



Bio-Rad Explorer Cloning and Sequencing Explorer Series

Curriculum Manual

Catalog #1665000EDU

Duplication of any part of this document is permitted for classroom use only. Please visit explorer.bio-rad.com to access our selection of language translations for Bio-Rad Explorer kit curricula.

This series ships in multiple shipments. Please see each individual module for storage conditions.

BIO-RAD

Dear Educator

Although the science of genetics goes back to the time of Gregor Mendel, an understanding of the chemical basis of heredity was not fully achieved until the latter half of the 20th century. The elucidation of DNA as the biochemical code occurred in the 1960s with individual genes being isolated (cloned) in the 1970s. The ability to determine the exact DNA sequence of genes emerged in the late 1970s, and a technique to synthesize large quantities of target regions of DNA using the polymerase chain reaction (PCR) technique was developed in the early 1980s. An electronic repository for the many genes being discovered was created in the late 1980s. This database, called GenBank, is operated by the National Center for Biotechnology Information (NCBI) and is funded by the U.S. National Institutes of Health. This database is accessible via the Internet by scientists, teachers, and students worldwide at no cost. Major efforts to completely sequence entire genomes were initiated in the 1990s and have now been completed for humans as well as for numerous model organisms studied by scientists, like the bacterium *Escherichia coli*, the common yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, the house mouse *mus musculus* and the brown rat, *rattus norvegicus*. The capacity for isolating and sequencing genes has grown so quickly that the number of submissions to GenBank has doubled every two years since 1993 (Benson et al. 2007). The challenge of analyzing all the DNA sequences deposited in GenBank spurred the development of numerous computer programs for interpreting DNA and protein sequence data. This computer-aided analytical approach is called bioinformatics.

The American Society for Biochemistry and Molecular Biology recommends that upon graduation, students have the skills to use computer databases as a research tool (Boyle 2004). Examples of research applications include the study of gene structure, gene regulation, protein structure and function, protein-protein interactions, and diseases caused by mutations (National Research Council 2003). Thus, the explosion and wealth of biological information in databases presents new challenges to incorporating data analysis as an educational tool (Campbell 2003, Honts 2003, Nichols et al. 2003).

Instructors of undergraduate laboratory courses involving molecular biology are faced with the challenge of staying abreast of the breathtaking advances occurring in the discipline. Laboratory exercises that were formerly assigned in upper-level biology courses (like PCR, digesting DNA with restriction enzymes, gel electrophoresis, and heat-shock transformation of bacteria) are now showing up in freshman-level and high school laboratory courses (Galewsky 2000, Lissemore et al. 2005, Kima and Rashe 2004, Schendel 1999). Educators are meeting this challenge by developing state-of-the-art laboratory exercises, but currently many of these are of limited didactic value as they are highly prescriptive, focus on only one or two techniques at a time, and are often not problem-based in their approach but simply require students to repeat exercises that are being carried out by thousands of other students on campuses worldwide (Streicher and Brodte 2002). Short, well-defined laboratory projects with known outcomes can encourage the notion that student experiments are just demonstrations, rather than encouraging the appreciation of research as a process (Caspers and Roberts-Kirchoff 2003, Gammie and Erdeniz 2004).

This laboratory project is less prescriptive, taking a multifaceted approach to incorporating both current molecular biology techniques as well as bioinformatics. Rather than being intimidated by the rapid pace at which molecular biology and bioinformatics has progressed this project embraces it. Due to technical advances that have been made in recent years (in PCR, DNA purification, gel electrophoresis, and documentation), molecular biology procedures have become more routine, safer, and relatively inexpensive with the widespread availability of basic cloning equipment. As another result of these advances, it is now possible to complete this project in a short length of time (6–8 weeks).

Cloning and Sequencing Explorer Series

In this Cloning and Sequencing Explorer Series, students can clone a housekeeping gene from a species that has not been studied before, then use bioinformatic approaches to study that gene. The DNA sequence obtained can be published in GenBank so that the sequence will be available to researchers throughout the world who are interested in gene structure, function, and evolution. Students involved in the project would also be able to see their names listed as coauthors of a permanent NCBI GenBank publication. This project allows students to become involved in a true investigatory research project, rather than a simple simulation of one. Our approach utilizes both active learning strategies and project-based laboratories as recommended by the National Research Council (2003). These strategies teach not only the process of science, but also basic laboratory methodologies (Handelsman et al. 2004). Evidence suggests that students who are engaged in active learning have better knowledge retention (Handelsman et al. 2004, National Science Foundation 1996). As a result, students develop critical thinking and problem-solving skills while at the same time learning to work cooperatively as a group.

This curriculum was developed in collaboration with Dr David Robinson and Dr Joann Lau of the Department of Biology, Bellarmine University, KY, Dr Kristi DeCoursey of the Fralin Biotechnology Center, Virginia Tech, VA, Dr. Sandra Porter of Digital World Biology and Christian Olsen of Biomatters, Inc. Bio-Rad thanks them for their invaluable guidance and contribution to this curriculum. We continually strive to evolve and improve our curricula and products. We welcome your suggestions and ideas!

Bio-Rad Explorer Team
Bio-Rad Laboratories
6000 James Watson Dr.
Hercules, CA 94547
Explorer@bio-rad.com

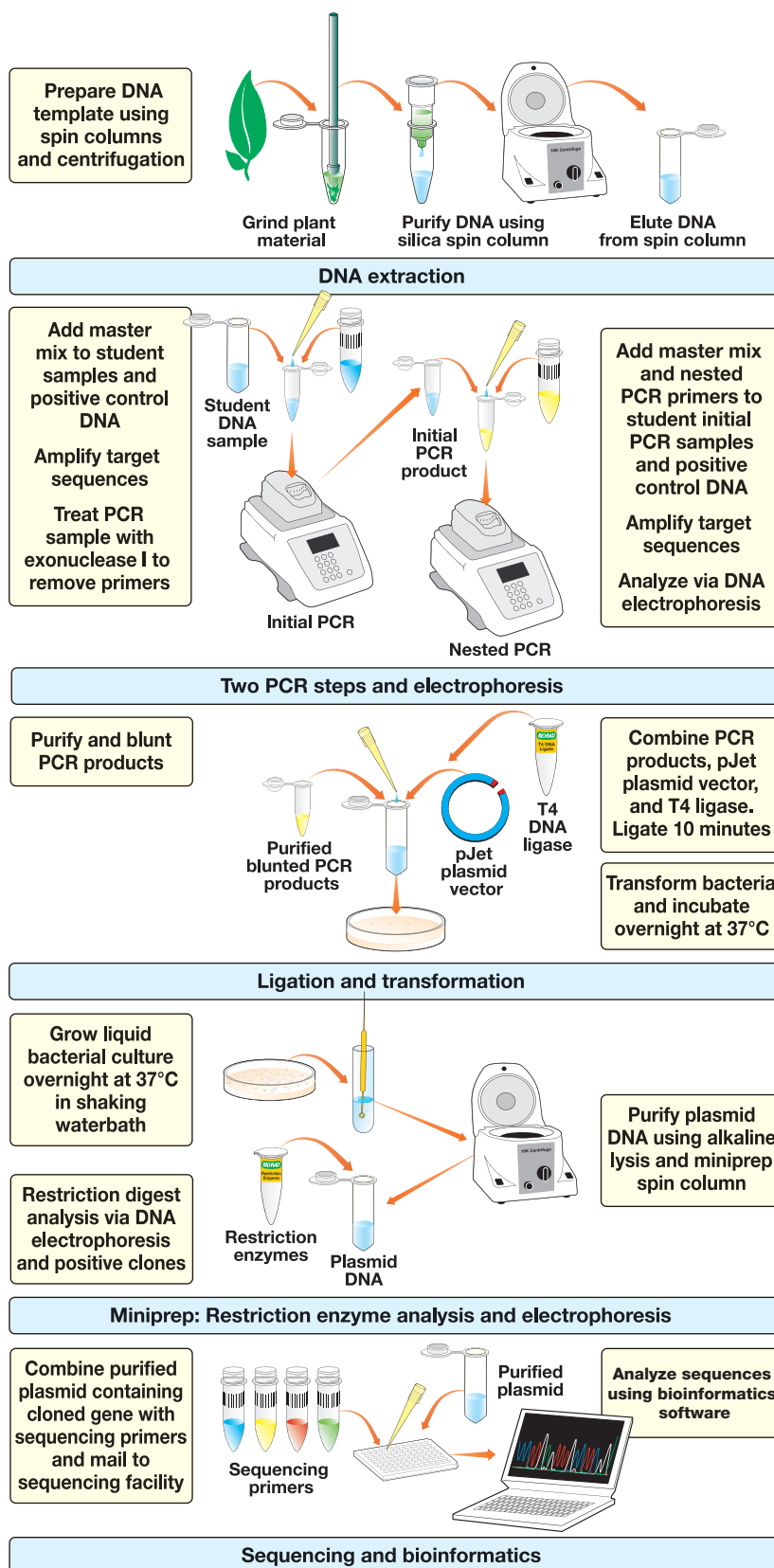


Table of Contents

	Page
Kit Summary	1
Storage Instructions	2
Checklists for Individual Modules	3
Course Objectives	9
Instructor's Advanced Preparation	10
Timeline for the Laboratory Course	11
Advice on Which Plant Species to Choose	13
General Background	15
Chapter 1: Nucleic Acid Extraction	
Background.....	21
Instructor's Advance Preparation	27
Protocol.....	29
Focus Questions	35
Quick Guide	36
Chapter 2: <i>GAPDH</i> PCR	
Background.....	39
Instructor's Advance Preparation	57
Protocol.....	60
Focus Questions	70
Quick Guide	72
Chapter 3: Electrophoresis	
Background.....	75
Instructor's Advance Preparation	76
Protocol.....	78
Focus Questions	83
Quick Guide	84
Chapter 4: PCR Purification	
Background.....	85
Instructor's Advance Preparation	86
Protocol.....	88
Focus Questions	91
Quick Guide	92
Chapter 5: Ligation	
Background.....	94
Instructor's Advance Preparation	100
Protocol.....	102
Focus Questions	106
Quick Guide	107

Chapter 6: Transformation	
Background.....	109
Instructor's Advance Preparation	112
Protocol.....	115
Focus Questions	120
Quick Guide	121
Chapter 7: Plasmid Purification	
Background.....	125
Instructor's Advance Preparation	130
Protocol.....	132
Focus Questions	139
Quick Guide	140
Chapter 8: DNA Sequencing	
Background.....	145
Instructor's Advance Preparation	149
Protocol.....	152
Focus Questions	157
Quick Guide	158
Chapter 9: Bioinformatics	
Background.....	159
Instructor's Advance Preparation	164
Protocol.....	174
Tour of the Geneious Prime platform	175
1. View Sequences and review the Quality of the Sequencing Data	177
Focus Questions	196
2. Assemble the Sequences and Correct Mistakes in the Basecalls.....	197
Focus Questions	210
3. Conduct BLAST Search on the Contig Sequence to Verify Identity of the Cloned Gene	212
Focus Questions	227
4. Determine Gene Structure (Exon/Intron/Exon Boundaries) Using BLAST Build a Gene Model	228
Focus Questions	243
5. Predicting an Amino Acid Sequence of the Cloned Gene (blastx).....	244
Focus Questions	250
Assemble Contig Sequences from the Entire Class	251

Appendices

Appendix A1:	Agarose Gel Electrophoresis Methods	252
Appendix A2:	Microbial Culturing Methods	255
Appendix A3:	Sterile Technique for PCR	261
Appendix B:	Additional Background on <i>GAPDH</i>	262
Appendix C:	Searching and Submitting Sequences to the GenBank	273
Appendix D:	Preparation of Research Papers and Presentations	279
Appendix E:	Custom BLAST Search Service Setup by Cut and Paste	284
Appendix F:	Tips and Tricks to Navigate the Geneious Prime Interface	287
Appendix G:	BLAST Searching on the NCBI Website	290
Appendix H:	Determining Sequencing Identities of Individual Sequences Using BLAST	300
Appendix I:	Glossary	313
Appendix J:	References	319

Kit Summary

The Cloning and Sequencing Explorer Series is a modular laboratory course designed to serve twelve student teams (two to four students per team). The aim of this course is to clone a portion of a glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene from plants, insert this gene fragment into a plasmid vector, and analyze the sequences of resulting clones using bioinformatics. Plants have multiple *GAPDH* genes; the specific ones that have been selected for cloning are the *GAPC* genes.

This project involves isolating, cloning, and analyzing a major portion of a plant *GAPC* gene, one of the family of *GAPDH* genes that code for the enzyme glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*). *GAPDH* genes are considered housekeeping genes because the encoded enzymes catalyze an important step of glycolysis, a stage of respiration that occurs in all living cells. Because of their central importance, *GAPDH* genes are located in the genome of all plants, as well as all other organisms.

This project offers an opportunity to perform novel research — to clone and sequence a gene that has not yet been analyzed and to add to the body of scientific knowledge around the world. The basic strategy is to extract genomic DNA (gDNA) and to synthesize large amounts of the DNA containing a portion of the *GAPC* gene. This will be done using polymerase chain reaction (PCR) on gDNA from a plant species you choose to study. Then this DNA will be inserted into a cloning vector and *E. coli* cells will be transformed with the resulting recombinant DNA. Cells containing this DNA will be screened, assessed, and multiplied so that large quantities of the DNA can be isolated and sequenced. The process of sequencing, typically done by a university or commercial laboratory, will reveal the exact sequence of DNA bases that makes up the *GAPC* gene cloned by the class. Once sequenced, DNA sequence files can be imported into the Geneious Prime software for analysis of DNA sequences. Sequences will be edited, assembled and annotated using bioinformatic tools. Once the DNA sequence is verified, this sequence can be published in NCBI's GenBank database and used by researchers throughout the world. The steps in this project to clone and analyze plant *GAPC* genes (and the corresponding chapters in this manual) are:

1. Identify the plant or plants to be studied and extract the gDNA (Nucleic Acid Extraction chapter).
2. Amplify a portion of the *GAPC* gene using PCR (*GAPDH* PCR chapter).
3. Assess the results of PCR (Electrophoresis chapter).
4. Purify the PCR product containing the *GAPC* gene fragment (PCR Purification chapter).
5. Ligate (insert) the *GAPC* gene fragment into a plasmid vector (Ligation chapter).
6. Transform bacteria with the plasmid (Transformation chapter).
7. Isolate the plasmid from the bacteria and confirm the presence of the insert by restriction enzyme digestion (Plasmid Purification chapter).
8. Prepare plasmid for DNA sequencing and send to a facility for sequencing (Sequencing chapter).
9. Obtain the sequence of the cloned *GAPC* gene fragment and analyze the cloned gene using bioinformatics (Bioinformatics chapter).

The first step in this exercise is to choose an interesting plant species. Some model species that plant biologists study, for example, *Arabidopsis thaliana*, the green alga *Chlamydomonas*, or crop plants like rice (*Oryza* species) and wheat (*Triticum* species), have already had their genomes sequenced, so the class may be reproducing and confirming this data. Alternatively, the class may be examining a species, variety, or cultivar that does not appear to have been studied well. There are over 250,000 known plant species, providing plenty of options with which to work.

It is possible that this laboratory will generate entirely novel sequence from plants that have not yet had a *GAPC* or closely related gene sequenced. It is also possible to clone *GAPDH* sequences that have been previously described and determine the validity of the data in GenBank. It is very beneficial to researchers to have data from multiple independently derived sequences from the same species – this greatly increases confidence in the data. If one objective is for students to participate in research that generates novel information, the instructor should perform an initial search of nucleotide databases (such as GenBank) to determine whether this gene has already been sequenced for a particular plant species (see Appendix C). In addition, if submission to GenBank is a goal of this activity, ensure the exact species of the plant is known, since this is required.

Because GAPDH is a crucial enzyme for all animals, protists, fungi, bacteria, and plants, there are lots of opportunities to draw connections between the molecular aspects of GAPDH and its biomedical and evolutionary significance. For instance, Altenburg and Greulich (2004) found that human GAPDH is highly expressed in 21 different types of cancer and that control of glycolysis may play an important role in future cancer therapies. GAPDH is a protein that appears to have roles in a wide range of cellular functions, including DNA replication and repair, cytoskeletal organization, and membrane fusion (Tatton et al. 2000), regulating transcription and programmed cell death (Kim and Dang 2005) as well as involvement in viral pathogenesis and aging-related neuronal diseases like Alzheimer's and Huntington's diseases (Sirover 1999).

About This Manual

The Cloning and Sequencing Explorer Series is accompanied by extensive curricula to enable students to gain a deep understanding about *GAPDH* genes and the theory behind the techniques they will use to perform the laboratory series.

Many of the individual modules that comprise the full Cloning and Sequencing Explorer Series contain an additional instruction manual- these smaller manuals are intended as guides when using the individual modules as stand-alone activities. When performing the entire Cloning and Sequencing Explorer Series, please use this comprehensive manual as your primary reference.

Storage Instructions

See Checklists for Individual Modules below for shipping and storage conditions for each module. Open each module as soon as it arrives and store components in recommended storage conditions.

Cloning and Sequencing Explorer Series Inventory Checklist (catalog #1665000EDU)

This section lists the components provided in the Cloning and Sequencing Explorer Series and then lists the components of each individual module comprising the series. It also lists required and optional accessories. Each Cloning and Sequencing Explorer Series contains sufficient materials to outfit 12 student workstations of one to four students per workstation. Use this checklist to inventory your supplies before beginning your advanced preparation.

Item	Quantity	(✓)
Nucleic Acid Extraction Module	1	<input type="checkbox"/>
GAPDH PCR Module	1	<input type="checkbox"/>
50x TAE buffer, 100 ml	2	<input type="checkbox"/>
Agarose powder, 25 g	1	<input type="checkbox"/>
UView 6x loading dye and stain, 1ml (found in GAPDH PCR module)	1	<input type="checkbox"/>
PCR Kleen Spin Purification Module	1	<input type="checkbox"/>
Ligation and Transformation Module	1	<input type="checkbox"/>
Microbial Culturing Module	1	<input type="checkbox"/>
Aurum Plasmid Mini Purification Module	1	<input type="checkbox"/>
1.5 ml EZ Micro Test Tubes, 500	1	<input type="checkbox"/>
Sequencing and Bioinformatics Module	1	<input type="checkbox"/>

Nucleic Acid Extraction Module (catalog #1665005EDU)

Item	Quantity	(✓)
Lysis buffer, 20 ml	1	<input type="checkbox"/>
DTT (Dithiothreitol), 0.3 g	1	<input type="checkbox"/>
Wash buffer, low stringency (5x concentrate), 20 ml	1	<input type="checkbox"/>
Sterile water, 2.5 ml	1	<input type="checkbox"/>
Micropestles	25	<input type="checkbox"/>
Aurum mini columns, purple*	25	<input type="checkbox"/>
Capless collection tubes, 2.0 ml	25	<input type="checkbox"/>
Microcentrifuge tubes, 1.5 ml	30	<input type="checkbox"/>
Microcentrifuge tubes, multicolor, 2.0 ml	60	<input type="checkbox"/>
Instruction manual**	1	<input type="checkbox"/>

Ships at room temperature. Immediately store DTT at 4°C.

* Both the mini columns in the Nucleic Acid Extraction module and the Aurum Plasmid Mini Purification Module are purple but they are functionally different. Be sure to keep them in their separate labeled bags so they do not get mixed or confused.

** Individual module instruction manuals are only included when the modules are purchased separately.

GAPDH PCR Module (catalog #1665010EDU)

Item	Quantity	(✓)
Initial <i>GAPDH</i> PCR primers, 50 µl	1	<input type="checkbox"/>
Nested <i>GAPDH</i> PCR primers, 50 µl	1	<input type="checkbox"/>
PCR master mix, 1.2 ml	3	<input type="checkbox"/>
pGAP control plasmid DNA for PCR	1	<input type="checkbox"/>
5x Control <i>Arabidopsis</i> genomic DNA, 20 µl	1	<input type="checkbox"/>
Exonuclease I, 50 µl	1	<input type="checkbox"/>
500 bp molecular weight ruler, 400 µl	1	<input type="checkbox"/>
UView 6x loading dye and stain, 1 ml	1	<input type="checkbox"/>
Sterile water, 2.5 ml	3	<input type="checkbox"/>
PCR tubes, 0.2 ml	150	<input type="checkbox"/>
Capless PCR tube adaptors, 1.5 ml	150	<input type="checkbox"/>
Microcentrifuge tubes, multicolor, 2.0 ml	120	<input type="checkbox"/>
Instruction manual*	1	<input type="checkbox"/>

Ships at 4°C. Immediately store temperature sensitive reagent bag at -20°C.

* Individual module instruction manuals are only included when the modules are purchased separately.

Electrophoresis Module (catalog #1660451EDU)

Item	Quantity	(✓)
Agarose powder, 25 g	1	<input type="checkbox"/>
Electrophoresis buffer, 50x TAE, 100 ml	2	<input type="checkbox"/>
UView 6x loading dye and stain, 1 ml**	1	<input type="checkbox"/>

Ships and stores at room temperature. Store UView loading dye at 4°C, and store TAE and agarose at room temperature.

** UView 6x loading dye and stain is included in the *GAPDH* PCR module.

PCR Kleen Spin Purification Module (catalog #7326300EDU)

Item	Quantity	(✓)
PCR Kleen spin columns, clear	25	<input type="checkbox"/>
Capless collection tubes, 2.0 ml	25	<input type="checkbox"/>
Microcentrifuge tubes, 1.5 ml	25	<input type="checkbox"/>
Instruction manual	1	<input type="checkbox"/>

Ships at room temperature, store at 4°C.

Ligation and Transformation Module (catalog #1665015EDU)

Item	Quantity	(✓)
T4 DNA ligase, 10 µl	1	<input type="checkbox"/>
2x Ligation reaction buffer, 100 µl	1	<input type="checkbox"/>
Proofreading polymerase, 10 µl	1	<input type="checkbox"/>
pJET1.2 blunted vector, 10 µl	1	<input type="checkbox"/>
BglII enzyme, 55 µl	1	<input type="checkbox"/>
10x Bg III reaction buffer, 1 ml	1	<input type="checkbox"/>
C-Growth medium, 30 ml	1	<input type="checkbox"/>
Transformation reagent A, 1.25 ml	4	<input type="checkbox"/>
Transformation reagent B, 1.25 ml	4	<input type="checkbox"/>
IPTG, 1 M, 0.1 ml	1	<input type="checkbox"/>
Sterile water, 1 ml	1	<input type="checkbox"/>
Microcentrifuge tubes, multicolor, 2.0 ml	120	<input type="checkbox"/>
Microcentrifuge tubes, 1.5 ml	30	<input type="checkbox"/>
Instruction manual*	1	<input type="checkbox"/>

Ships at 4°C. Immediately store reagents bags at -20°C.

Microbial Culturing Module (catalog #1665020EDU)

Item	Quantity	(✓)
Ampicillin, lyophilized	2	<input type="checkbox"/>
LB broth capsules	12	<input type="checkbox"/>
LB nutrient agar powder	1	<input type="checkbox"/>
Petri dishes, 60 mm, sterile	40	<input type="checkbox"/>
Cell culture tubes, 15 ml, sterile	75	<input type="checkbox"/>
Inoculation loops, sterile	80	<input type="checkbox"/>
<i>E. coli</i> strain HB101 K-12, lyophilized	1	<input type="checkbox"/>
Disposable plastic transfer pipets	10	<input type="checkbox"/>
Instruction manual*	1	<input type="checkbox"/>

Ships and stores at room temperature.

* Only included with modules that are purchased separately.

Aurum Plasmid Mini Purification Module (catalog #7326400EDU)

Item	Quantity	(✓)
Plasmid mini columns, purple**	100	<input type="checkbox"/>
Capless collection tubes	100	<input type="checkbox"/>
Resuspension solution, 25 ml	1	<input type="checkbox"/>
Lysis solution, 25 ml	1	<input type="checkbox"/>
Neutralization solution, 40 ml	1	<input type="checkbox"/>
Wash solution, 5x concentrate, 25 ml	1	<input type="checkbox"/>
Elution solution, 16 ml	1	<input type="checkbox"/>
Instruction manual	1	<input type="checkbox"/>

Ships and stores at room temperature.

** Both the mini columns in the Nucleic Acid Extraction module and the Aurum Plasmid Mini Purification Module are purple but they are functionally different. Be sure to keep them in their separate labeled bags so they do not get mixed or confused.

Sequencing and Bioinformatics Module (catalog #1665025EDU)

Item	Quantity	(✓)
pJET SEQ F primer, 50 µl	1	<input type="checkbox"/>
pJET SEQ R primer, 50 µl	1	<input type="checkbox"/>
GAP SEQ F primer, 50 µl	1	<input type="checkbox"/>
GAP SEQ R primer, 50 µl	1	<input type="checkbox"/>
pGAP control plasmid for sequencing, 100 µl	1	<input type="checkbox"/>
Barcoded 96-well plate	1	<input type="checkbox"/>
Sealing film	1	<input type="checkbox"/>
Microcentrifuge tubes, multicolor, 2 ml	120	<input type="checkbox"/>
Instruction manual*	1	<input type="checkbox"/>

Ships at 4°C. Immediately store reagents bag at -20°C.

* Only included with modules that are purchased separately.

Materials Required but Not Supplied

Reagents	Quantity
95–100% ethanol, molecular biology grade	500 ml
Distilled deionized water (ddH ₂ O)	4–5 L

Plastics and Consumables	Quantity
Weigh paper or weigh boats	1 box
Razor blades or scalpels	24
100–1,000 µl pipet tips, aerosol barrier (catalog #2112021EDU)	12 boxes
20–200 µl pipet tips, aerosol barrier (catalog #2112016EDU)	12 boxes
2–20 µl pipet tips, aerosol barrier (catalog #2112006EDU)	12 boxes
100–1,000 µl pipet tips, standard style (catalog #2239350EDU)	12 boxes
2–200 µl pipet tips, standard style (catalog #2239347EDU)	12 boxes
0.5–10 µl pipet tips, standard style (catalog #2239354EDU)	12 boxes
Parafilm sealing film	1

Temperature Control Equipment	Quantity
Microwave oven	1
−20°C freezer	1
Water bath (catalog #1660504EDU) or Digital dry bath (catalog #1660562EDU)	1
Incubation Oven (catalog #1660501EDU)	1
Shaking water bath or shaking incubator	1
Thermal cycler (catalog #1861096EDU)	1
Additional water bath, heating block, or incubator	1

Other Equipment	Quantity
Balance (capable of measuring to 5 mg)	1/class
100–1,000 µl adjustable micropipet (catalog #1660508EDU, 1660553EDU)	12
20–200 µl adjustable micropipet (catalog #1660507EDU, 1660552EDU)	12
2–20 µl adjustable micropipet (catalog #1660506EDU, 1660551EDU)	12
0.5–10 µl adjustable micropipet (catalog #1660505EDU, 1660550EDU)	12
Tube racks	12
Ice bath	12
Casting apparatus for agarose gels (catalog #1704422EDU)	12
Horizontal electrophoresis chambers (catalog #1664000EDU)	12
Power supply (catalog #1645050EDU)	3–12
Gel documentation system (catalog #1708270EDU)	1
Microcentrifuge with variable-speed setting $\geq 12K \times g$ (catalog #1660602EDU) (refrigeration feature recommended)	2
Desktop computers with Internet access	12

Miscellaneous	Quantity
Plants for DNA extraction	2
Marking pens	12

Optional materials	Quantity
Autoclave	1
Vortexer (catalog #1660610EDU)	1

Refills Available Separately

Each individual module is available to order as a stand alone kit. In addition, certain refill packs are also available:

Nucleic acid extraction module refill pack, catalog #1665006EDU includes DTT (0.3 g), lysis buffer (20 ml), low stringency wash buffer (5x concentrate) (20 ml), sterile water (2.5 ml)

GAPDH PCR module refill pack, catalog #1665011EDU includes Initial GAPDH PCR primers (50 µl), Nested GAPDH PCR primers (50 µl), PCR master mix (3.6 ml), pGAP control plasmid DNA for PCR (1 ml), 5x Control Arabidopsis gDNA (20 µl), Exonuclease I (50 µl), Molecular weight ruler (400 µl), UView 6x loading dye and stain (1 ml), and Sterile water (2.5 ml)

2x Master mix for PCR (1.2 ml), catalog #1665009EDU

Certified molecular biology agarose (25 g), catalog #1613100EDU

UView 6x Loading Dye and Stain, 6x, 1 ml, catalog #1665112EDU

50x TAE, 1 L, catalog #1610743EDU

Small Fast Blast DNA Electrophoresis Reagent Pack, catalog #1660450EDU includes agarose (25 g), 50x TAE (100 ml), 500x Fast Blast DNA stain (100 ml)

Ligation module refill pack, catalog #1665016EDU includes T4 DNA ligase (10 µl), 2x Ligation reaction buffer (100 µl), Proofreading polymerase (10 µl), pJET1.2 blunted vector (10 µl), and Sterile water (1 ml)

BglII reagent refill pack, catalog #1665018EDU includes BglII enzyme (55 µl) and BglII reaction buffer (1 ml)

Transformation module refill pack, catalog #1665017EDU includes Transformation reagent A (5 ml), Transformation reagent B (5 ml), IPTG 1M (0.1 ml) and C-Growth medium (30 ml)

Microbial Culturing module refill pack, catalog #1665021EDU includes Ampicillin (lyophilized, 2 vials), LB broth capsules (12), LB agar powder (1 pouch), and *E. coli* strain HB101 K-12 (lyophilized, 1 vial)

LB nutrient agar powder, catalog #1660600EDU

Ampicillin, catalog #1660407EDU

***E. coli* strain HB101 K-12**, catalog #1660408EDU

Course Objectives

The Cloning and Sequencing Explorer Series is appropriate for the laboratory portion of an undergraduate (or early graduate) course in Molecular Biology, Cell Biology, Genetics, Biotechnology, Recombinant DNA Techniques, or Advanced Plant Biology. It would also be suitable for students doing independent research. It would be excellent for inclusion in biotechnology degree programs offered by community or technical colleges. The exercise could also prove useful for employers in the biotechnology, pharmaceutical, or industrial sectors. This laboratory project is an effective way of demonstrating PCR, digestion with restriction enzymes, use of DNA vectors, ligation, transformation, recombinant bacterial screening, and bioinformatics for employees needing an introduction — or a refresher course — in biotechnology. Due to recent advances in the area of DNA technology, the actual laboratory procedures are routine, safe, and relatively inexpensive, provided that basic cloning equipment is available. In order to complete the laboratory project in 6–8 weeks, it is assumed that students meet at least once per week in a 3-hour laboratory session, and that there are other times during the week when students can meet to carry out a quick laboratory task or two.

Specific objectives met by this project

1. Students will experience a wide range of laboratory techniques. Some of the techniques implemented in this PCR-based project are: DNA extraction and purification, basic micropipetting, PCR optimization, gel electrophoresis, restriction enzyme digestion, working with cloning vectors, ligation, heat-shock transformation, subculturing, plasmid preparation, DNA sequencing, and bioinformatic analyses.
2. Students will see that these individual techniques are just steps in a longer investigatory process. Few researchers can complete an entire research project in one or two 3-hour laboratory sessions (the timeframe of most commercially available kits), so this 6–8 week project more accurately reflects what goes on in a contemporary molecular biology laboratory.
3. Students will be active participants in the process. There are numerous occasions during this project when students are asked to troubleshoot their results, or to make judgments about what to do next. This exercise does not take a simple “cookbook” approach, but rather involves more critical thinking.
4. By having publication of a DNA sequence be the long-term goal of the project, both the students and the instructors will be more engaged in the process. This can make the experience more fulfilling (personally and professionally) than doing simulations.
5. This project is not an exercise where students are carrying out redundant experiments or are competing with each other in the laboratory. This exercise is a cooperative effort and requires students to share their data to create a larger, more useful, final product (a gene sequence from a single organism). On a larger scale, it is also possible for different institutions to cooperate. If different institutions set about to isolate the same gene from different species, and publish their sequences in NCBI's GenBank, then over the years an enormous database could be developed. Future students would have access to this enlarged database to study the evolution of *GAPDH* genes using bioinformatic approaches.

Instructor's Advance Preparation

The amount of advance preparation performed for the students will vary greatly depending on the level of the students and goals of the instructor. A detailed description of the preparation required for each laboratory stage is provided in each chapter after the background section.

Standard techniques: Basic standard laboratory techniques for preparing, loading, and running agarose gels and preparing culture media and agar plates are included in Appendix A for students or instructors to perform prior to or as part of the laboratory.

Aliquoting reagents: Depending on the level and number of students, the instructor may prefer to prepare aliquots of reagents for student teams, or to have students take required reagents from a common stock. A list of requirements for each student workstation is provided at the start of each section. Additional requirements and information for the instructor at each stage are within individual modules later in this manual.

Sample tube colors: Color tubes can be used to distinguish closely related reagents. The instructor may also opt to assign tubes of a particular color to teams or sample types to help keep track of samples during and after preparation.

Use of aerosol barrier tips: PCR is very sensitive to contamination since it is designed to amplify minute quantities of DNA. DNA can aerosolize and gain entry into pipet barrels. Therefore it is highly recommended that aerosol barrier tips (also called filter tips) be used in all steps of sample preparation prior to the final PCR step. For this laboratory aerosol barrier tips should be used until the second nested PCR step has been completed, including electrophoresis of initial PCR products if this is performed prior to the nested PCR step. See Appendix A3 for further information on PCR amplification and sterile technique.

Reagents used multiple times: Some of the reagents supplied are used in multiple places in the course. For example, the 500 bp molecular weight ruler supplied in the *GAPDH* PCR module is used in both chapters three and seven for electrophoresis of PCR products and plasmid minipreps, respectively. Therefore, it is important that reagents not be discarded following use at a particular stage.

Skills Students Need to Perform This Laboratory

The Cloning and Sequencing Explorer Series assumes students and instructors have basic molecular biology and microbiology laboratory skills, such as loading and running agarose gels, micropipetting with care and accuracy, pouring and streaking agar plates, knowledge of molarity calculations, etc. Brief protocols for basic skills are included in the appendices as reminders rather than as thorough lessons. Also, this is not an ideal laboratory to introduce PCR to students. The Bio-Rad Explorer program has a full range of kits to help teach basic skills in individual laboratories prior to introducing students to how these separate skills connect into a single workflow, as is done in this laboratory course.

This laboratory also requires knowledge of basic computer skills, including familiarity with web browsers, navigation of software graphical user interfaces, search engines, and ways to find and navigate to files. For details, refer to Chapter 9.

Setting Class Expectations for Success

Before starting on the Cloning and Sequencing Explorer Series, it is important for the instructor to help the students understand what the curriculum involves and set expectations for the type of results they may experience along the way. This curriculum was not created to be a demonstration laboratory with guaranteed results of the sort to which some students may be accustomed. The Cloning and Sequencing Explorer Series takes learning to the next level since the curriculum is part of a real ongoing research

project involving novel and real experiments. Great attention has been given to the design of the experiments with respect to controls and the creation of robust protocols to try to ensure success. Specific controls have been added to the experiments to not only enable accurate interpretation of results, but to also provide a safety net so that students who do not achieve adequate results at any section can continue on with the known control and complete the series. Additionally, layers of repetition have been designed within the protocols so that multiple groups working on the same plant will not only be working to achieve the same results, but will also verify the data that each group produces. Despite this, successful results cannot be guaranteed, some failures may still occur; these may be due to technique, or real biological differences between plant species. If failures do occur, it is important for the instructor to determine whether to keep to the timeline of the course and use the results of the controls to continue or repeat steps depending on the objectives of the course. Some ways to increase successful outcomes are to have the whole class work with the same plant, choose plants from the “Plants Known to be Successful” list and have students practice molecular biology techniques for a couple of weeks prior to embarking on this project. Ultimately, this Series is designed to provide a safe starting place for students to delve into the real world of research and to see how exciting discovering new data can be while learning that real science also means learning about proper experimental design, interpretation of results and learning from failures.

Timeline for the Laboratory Course

The timeline will depend greatly on the level of the students, whether the ligation and transformation stages are combined or not, and whether other techniques and analyses are performed in addition to the basic protocol. A rough guide is provided here.

Laboratory Session	Chapter	Task	Estimated Duration	Module Containing Materials
1	1: Nucleic Acid Extraction	Extract DNA	2 hr	Nucleic Acid Extraction
	2: <i>GAPDH</i> PCR	Set up initial PCR	0.5–1 hr	<i>GAPDH</i> PCR
	2: <i>GAPDH</i> PCR	Run PCR reaction in thermal cycler	3–4 hr*	<i>GAPDH</i> PCR
2	2: <i>GAPDH</i> PCR	Treat initial PCR reactions with exonuclease I	1 hr	<i>GAPDH</i> PCR
	2: <i>GAPDH</i> PCR	Set up nested PCR reactions	0.5–1 hr	<i>GAPDH</i> PCR
	2: <i>GAPDH</i> PCR	Run PCR reaction in thermal cycler	3–4 hr*	<i>GAPDH</i> PCR
	Prep for activities in chapter 3	Pour agarose gels	0.5 hr	Electrophoresis
	Prep for activities in chapters 6 and 7	Pour LB and LB Amp IPTG agar plates	0.5 hr	Ligation and Transformation,
		Prepare LB and LB Amp broth	0.5 hr	Microbial Culturing
3**		Streak starter plate with bacteria	5 min	
		Grow starter plate at 37°C	16+ hr*	
	3: Electrophoresis	Electrophorese PCR products	0.5–1 hr	Electrophoresis
	3: Electrophoresis	Decide which PCR products to clone	0.5 hr	N/A
	4: PCR Purification	Purify PCR products	0.5 hr	PCR Purification
	Continue prep for activities in chapters 6 and 7	Complete prep from session 2	1 hr	Ligation and
		Inoculate single starter culture	5 min	Transformation,
		Grow starter culture at 37°C	8+ hr	Microbial Culturing

Timeline for the Laboratory Course (cont.)

4**	5: Ligation	Ligate PCR product	1 hr	Ligation and Transformation
5**	6: Transformation	Transform bacteria with ligated product and plate them on LB Amp IPTG agar plates	1 hr	Ligation and Transformation, Microbial Culturing
	6: Transformation	Grow bacteria at 37°C	16+ hr*	N/A
	Prep for activities in chapter 7	Pour agarose gels	0.5 hr	Electrophoresis
Between 5 and 6	7: Plasmid Purification	Inoculate transformed colonies into miniprep LB AMP broth for culturing	5 min	Microbial Culturing
	7: Plasmid Purification	Grow bacteria at 37°C	8+ hr*	N/A
6	7: Plasmid Purification	Purify plasmids from miniprep cultures	1 hr	Plasmid Purification
	7: Plasmid Purification	Perform restriction digestion of plasmids	1–1.5 hr	Ligation and Transformation
	7: Plasmid Purification	Electrophorese plasmid digests	0.5–1 hr	Electrophoresis
7	7: Plasmid Purification	Analyze results	1 hr	N/A
	8: Sequencing	Prepare sequencing reactions	0.5–1 hr	Sequencing and Bioinformatics
	8: Sequencing	Send sequencing reactions away to be processed	Up to 2 weeks*	Sequencing and Bioinformatics
8–10	9: Bioinformatics	Practice analyzing sample sequences	3 h	Sequencing and Bioinformatics
	9: Bioinformatics	Analyze sequences	6+ hr	

* Time indicated for these tasks is not hands-on time. It is the time needed for reactions to run, bacteria to grow or sequencing reactions to be processed.

** (Optional) Stages 3, 4, and 5 can be combined into a single laboratory session.

Advice on Which Plant Species to Choose

DNA can be extracted from plants by many methods, some of which are specific to a single type of plant. Although every attempt has been made to make this laboratory as universal as possible, the fact that it uses a single DNA extraction method means there will be some plants for which the extraction method does not succeed. Some plants have very tough cell walls or other characteristics that make it difficult to extract DNA. Moreover, different plants will yield different quantities of DNA, and the ability of that DNA to be amplified may vary. It is possible for plants to yield very little genomic DNA (gDNA) that is easily amplified, or to yield a lot of gDNA that is poorly amplified.

Likewise, primers have been designed to allow amplification of *GAPDH* DNA sequence from the majority of plants using PCR, the step used to generate sufficient DNA for cloning. However, although *GAPDH* is very highly conserved at the protein level, there is a good deal of variation among the DNA sequences of different plant species, so some plants may have DNA that is amplified poorly or not at all with the primers provided. Alternatively, a particular plant may contain a metabolite that interferes with PCR, preventing amplification.

To improve success rates and to demonstrate additional techniques, the PCR protocol consists of two rounds. In the first, DNA is amplified from gDNA using degenerate primers, which are designed based on the highly conserved amino acid sequence of *GAPDH*. Then, a further round of nested PCR is performed to increase both the specificity and quantity of PCR product. Depending on the plant used, the first round of PCR may produce no PCR product visible on an agarose gel or it may produce a well-amplified product. If it does, the instructor may decide to reduce the time spent on the amplification steps and to clone directly from the product of the first round of PCR. However, performing the protocol with two-rounds of PCR will have the best chance of student success.

The instructor is encouraged to choose the plants to use for the laboratory with both challenges of the techniques and teaching goals in mind. One option is to choose a plant known to be successful with this protocol (see table) to ensure student success and comfort with the techniques taught. Another is to choose a plant that is less well characterized to teach students the reality of research, with the risk that the experiment may be unsuccessful, but with great opportunities to troubleshoot and encourage independent research (such as by extracting DNA from different parts of the plants, using alternative extraction methods, optimizing PCR, and redesigning primers). To combine both the best chance of success and troubleshooting opportunities, an excellent approach is to choose two plants, one known to be successful and another that is less well-characterized. If the goal is to obtain solid sequence data to be uploaded into GenBank, the entire class should perform the experiment using the same plant, so that the multiple sequences generated will ensure more reliable accuracy of the data.

In addition, the laboratory contains controls (genomic DNA from *Arabidopsis thaliana* and an *A. thaliana GAPDH* gene fragment cloned into the plasmid vector) to assess the validity of the results. These controls ensure that at each step the products from the controls will allow continuation of the experiment in the event that the students' experiments are not successful, thus allowing the techniques to be learned and performed in a timely manner even if novel data are not acquired. However, student troubleshooting and repetition are extremely valuable and should be encouraged whenever time permits.

List of Plants Evaluated with This Protocol

Plants Known to Be Successful	Plants Known to Not Be Successful
Aluminum plant, leaves (<i>Pilea cadieri</i>)	Fern, leaves
Cabbage, leaves (<i>Brassica oleracea</i>)	Ivy, leaves (<i>Hedera helix</i>)
Carrot, root (<i>Daucus carota</i>)	Lily of the valley, leaves (<i>Convallaria major</i>)
Common sage, leaves (<i>Salvia officinalis</i>)	Pacific coast iris, petals and leaves (<i>Iris douglasiana</i>)
Croton, leaves (<i>Codiaeum variegatum</i>)	Pine, needles (<i>Pinus spp.</i>)
Green bean, seed case and bean (<i>Phaseolus vulgaris</i>)	Spider plant, leaves (<i>Chlorophytum comosum</i>)
Jade pothos, leaves (<i>Epipremnum aureum</i>)	
Lamb's ear, leaves (<i>Stachys byzantina</i>)	
Lawn grass, leaves	
Parsley, leaves (<i>Petroselinum crispum</i>)	
Petunia, leaves and petals (<i>Petunia x hybrida</i>)	
Pineapple sage, leaves (<i>Salvia elegans</i>)	
Purple heart, flowers and leaves (<i>Setcreasea pallida</i>)	
Snapdragon, leaves (<i>Antirrhinum majus</i>)	
Sow thistle, leaves (<i>Sonchus oleraceus</i>)	
Spinach, leaves (<i>Spinacia oleracea</i>)*	
Sugar cane, leaves (<i>Saccharum officinarum</i>)	
Sweet potato, tuber (<i>Ipomoea batatas</i>)	
Thyme, leaves (<i>Thymus vulgaris</i>)	
Umbrella plant, leaves (<i>Cyperus involucreatus</i>)	
Umbrella Schefflera, leaves (<i>Schefflera arboricola</i>)	
Wheatgrass, leaves (<i>Triticum aestivum</i>)	

* PCR fragment not visible after first round of PCR.

Plant Taxonomy

If the goal of this exercise is to submit gene sequences to GenBank, students could derive the taxonomic classification for the chosen plant species as an additional exercise.

The NCBI Taxonomy Browser at ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy can help. Use the browser to do a search using the plant's scientific name.

General Background

Glycolysis and the Role of Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH)

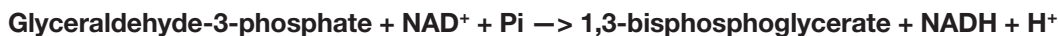
The gene that will be studied in this laboratory codes for the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). There are several reasons why this gene was selected. First, GAPDH is a crucial enzyme in glycolysis and is known as a housekeeping gene (a gene that is expressed constitutively and is necessary for cells to survive). Since GAPDH is abundant in cells and can be purified for study, much is known about this protein's structure and function. GAPDH consists of four subunits (hence, a tetramer) held together by noncovalent binding. All four subunits may be identical (designated as A4, a homotetramer) or they may be pairs of slightly different subunits (designated A2B2, a heterodimer). In both cases, each subunit has an active site (the part of the enzyme that carries out the enzymatic reaction) and can bind one molecule of the cofactor NAD^+ .

GAPDH protein has two major domains. The amino terminal region has an NAD^+ binding domain and the carboxy terminal region has glyceraldehyde-3'-phosphate dehydrogenase activity. In addition, recent research has found that GAPDH plays many other roles outside of glycolysis. For example, the human GAPDH gene is overexpressed (that is, expressed at levels much higher than normal) in 21 different classes of cancer (Altenburg and Greulich 2004). GAPDH plays roles in membrane fusion, endocytosis, microtubule bundling, and DNA repair. GAPDH is also involved in viral pathogenesis, regulation of programmed cell death, and human neuronal diseases including Alzheimer's and Huntington's diseases (Sirover 1999).

GAPDH catalyzes the sixth of the ten-step process of glycolysis, the pathway by which glucose is converted into pyruvate in a series of enzymatically catalyzed reactions. In mammals, most dietary polysaccharides are broken down to glucose in the bloodstream. In plants, glucose is synthesized from carbon dioxide in the Calvin cycle of photosynthesis. Glycolysis provides a number of usable products:

- ATP and NADH, produced during the process of glycolysis, providing energy for the cells
- Pyruvate, the end product of glycolysis, which feeds into the citric acid cycle, producing more energy for the cells
- Intermediate compounds in glycolysis, many of which are precursors for the formation of other biological molecules; for example, glucose-6-phosphate is a precursor for the synthesis of ADP, NAD^+ , and coenzyme Q, and phosphoenolpyruvate is a precursor for the synthesis of the amino acids tyrosine, phenylalanine, and tryptophan

The reaction catalyzed by GAPDH is:



Thus, GAPDH oxidizes glyceraldehyde-3-phosphate (GAP) by removing a hydrogen ion (H^+) and transferring it to the acceptor molecule, NAD^+ ($\text{NAD}^+ + \text{H}^+ \rightarrow \text{NADH}$). In addition, GAPDH adds a second phosphate group to GAP. This reaction is catalyzed by a cysteine in the active site of the GAPDH protein. When the source of carbohydrate for glycolysis is a sugar, glycolysis will occur in the cytosol, as it does in animal cells. When the carbohydrate source is starch, however, glycolysis can occur in plastids (a group of plant organelles that includes chloroplasts). For more information about the role of GAPDH in glycolysis, please refer to Appendix B.

Origin of *GAPDH* Genes

In plants, there are two metabolic pathways for carbohydrates: the Calvin cycle in chloroplasts and glycolysis in the cytosol. The pathways share some enzymatic reactions (including the reaction catalyzed by GAPDH), but the enzymes in the two pathways are not identical even though they catalyze the same chemical reactions in both pathways. The enzymes in the two pathways are isozymes or isoenzymes, homologous enzymes that catalyze the same reaction but differ in amino acid sequence. A separate gene encodes each isozyme, and all of the genes are encoded in the nucleus (Plaxton 1996). As one example of genes encoding isozymes, the enzyme hexokinase phosphorylates glucose both in the chloroplast and in the cytosol, but two separate genes in the plant cell nucleus encode cytosolic hexokinase and chloroplast hexokinase. Isozymes are very common in plants and animals, and typically appear to have resulted from gene duplication events that occurred millions of years ago. Sometimes the gene duplication event occurred within the nucleus itself, while other times, genes appear to have been translocated from nuclear DNA to mitochondrial or plastid DNA. One of the observations about mitochondria and plastids that led to the endosymbiotic theory of evolution (that these organelles exist due to an ancient symbiosis that resulted in prokaryotes being hosted within eukaryotes) is the observation that they contain DNA that is similar to bacterial DNA. Based on this theory it would have been more than one billion years ago that photosynthetic cyanobacteria were engulfed by eukaryotic cells, becoming the antecedents of modern plastids. The resulting subcellular organelles, plastids, have taken over many reactions for their host cells, including photosynthesis, carbohydrate metabolism, amino acid synthesis, lipid production, photorespiration, and nitrogen and sulfur reduction. At the same time, plastids still have their own DNA, as well as the machinery for replication, transcription, and translation. However, plastids retain only a fraction of the genome that their ancestors had. The plastid genome encodes 120–135 genes (López-Juez 2007), whereas the closest living relatives to the plastid ancestor, cyanobacteria of the genus *Nostoc* (Martin et al. 2002), have 3,000–7,000 genes. Most genes originally found in the symbiotic cyanobacteria are now found in the plant cell nucleus. Martin et al. (2002) report that about 18% of the protein coding genes in *Arabidopsis* are derived from cyanobacteria. However, the gene transfer was not unidirectional. Genes that preexisted in the nuclear genome have also been transferred to the plastid genome, but gene expression in the plastid is under nuclear control and most plastid proteins are encoded by nuclear genomic DNA (gDNA). All GAPDH isozymes found in eukaryotes are nuclear encoded and are believed to have originated in cyanobacteria (Martin et al. 2002). The genes that encode GAPDH are known as *GAPDH* genes (italicized abbreviations are a common convention to indicate a gene, whereas the non-italicized abbreviations refer to the protein). The duplication of *GAPDH* genes that gave rise to the chloroplast form is believed to have occurred during the period when land plants first emerged (Teich et al. 2007), and subsequent gene duplications resulted in the multiple forms now present in modern plants.

GAPDH Genes in Arabidopsis thaliana

As the “laboratory rat” of plant research, *A. thaliana* has been studied extensively. *Arabidopsis* is a small flowering plant of the mustard family. There are several reasons why it is used as a model system, including its rapid life cycle (6 weeks), small genome (125 Mb), and prolific seed production. Sequencing of the *Arabidopsis* genome was completed in the year 2000. (More information about *Arabidopsis* can be found on the web site known as The *Arabidopsis* Information Resource (TAIR) (arabidopsis.org)). Since so much basic research is done using *Arabidopsis* as the model system, much is known about the eight *Arabidopsis* GAPDH genes, including the chromosomes on which the genes are found. The nomenclature for the GAPDH genes is determined by the enzyme localization and function:

- GAPC — Cytosolic
- GAPCP — Plastid
- GAPA — Chloroplast
- GAPB — Chloroplast
- GAPN — Cytosolic, nonphosphorylating

Each gene represents a different GAPDH enzyme function, and several have multiple forms (isozymes). Remember that GAPDH is a tetramer, and for each GAPDH protein, the four subunits may be identical or of two different types.

GAPDH enzymes and the genes encoding them in *Arabidopsis*

Enzyme function	GAPDH protein subunits	EC* designation	Arabidopsis gene	Arabidopsis gene accession number **	Arabidopsis chromosome location	Position on chromosome
NAD ⁺ -dependent GAPDH in cytosol	GAPC GAPC-2	EC 1.2.1.12	GAPC GAPC-2	NM_111283 NM_101214	3 1	1,080,957–1,083,537 4,608,193–4,610,644
NAD ⁺ -dependent GAPDH in plastids	GAPCP GAPCP-2	EC 1.2.1.12	GAPCP GAPCP-2	NM_106601 NM_101496	1 1	29,920,795–29,924,127 5,574,304–5,574,616
NADP ⁺ -dependent GAPDH in chloroplast	GAPA GAPA-2 GAPB	EC 1.2.1.13	GAPA GAPA-2 GAPB	NM_113576 NM_101161 NM_103456	3 1 1	9,796,308–9,798,282 4,392,448–4,394,398 16,132,283–16,132,283
Non-phosphorylating NADP ⁺ dependent GAPDH in cytosol	NP-GAP or ALD11A3	EC 1.2.1.9	GAPN	NM_201797	1	—

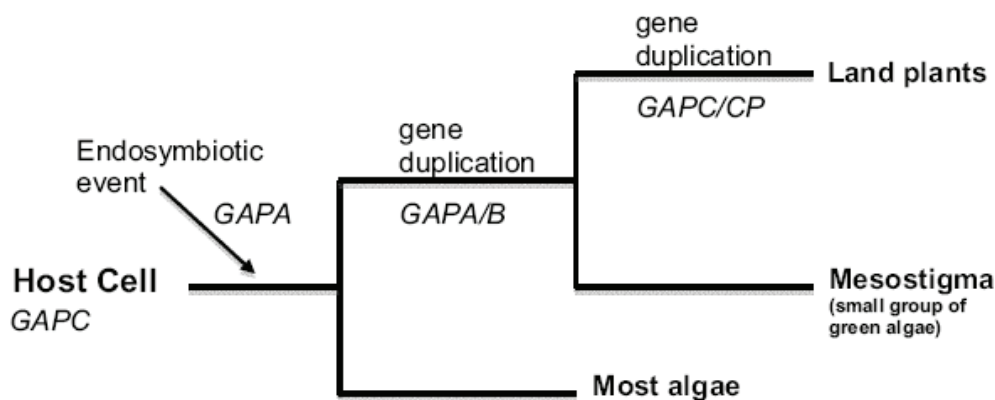
* EC stands for Enzyme Commission – see below.

** The accession number is the identifier assigned to the sequence record when it is submitted to the National Institutes of Health National Center for Biotechnology Information nucleotide sequence database, GenBank. The accession number for *Arabidopsis* chromosome 1 is NC_003070 and chromosome 3 is NC_003074. The sequence of bases of a chromosome are numbered from the 5' end to the 3' end. The position on the chromosome shows the specific location of the gene on its *Arabidopsis* chromosome.

The exact function of an enzyme is described by its EC designation, a numerical classification system from the Enzyme Commission (EC) that assigns numbers to all enzymes based on the reactions that they catalyze. Each number of the EC designation has a particular meaning. For example, in the case of the phosphorylating GAPDH found in the cytosol:

- EC 1._._._ designates enzymes that are oxidoreductases.
- EC 1.2._._ designates enzymes that act on the aldehyde or oxo group of donors
- EC 1.2.1._ designates enzymes specifically with NAD⁺ or NADP⁺ as acceptor
- EC 1.2.1.12 specifically designates a phosphorylating GAPDH enzyme

All of these *GAPDH* genes arose from a series of gene duplications over time. For example, *GAPB* originated by duplication of the *GAPA* gene, probably in early green algae (Brinkmann et al. 1989). Teich et al. (2007) developed a genetic tree to show phylogeny (inferred evolutionary relationships based on gene or protein sequences) using *GAPDH* sequences from *Arabidopsis* and several other plants. (The more closely related the sequences are, the closer together they appear on the phylogenetic tree.)



Phylogeny of *GAPDH* genes in plants. The original cells had the *GAPC* gene. When photosynthetic cyanobacteria were engulfed in the cells in an endosymbiotic event, the cells gained the *GAPA* gene. The next split was a gene duplication of *GAPA* to form *GAPB*, so all land plants (and *Mesostigma*, a small group of green algae) have the *GAPB* gene. The final gene duplication of *GAPC* to form *GAPCP* separated the *Mesostigma* from land plants (which include dicots, monocots, club mosses, mosses, and liverworts). Figure adapted from Teich et al. (2007).

Plant *GAPC* Genes and Nested Polymerase Chain Reaction (PCR)

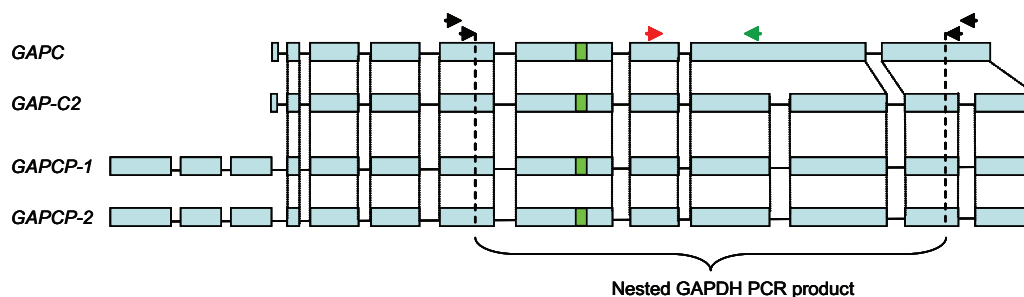
In this laboratory project, a portion of a *GAPC* gene, which is located in plant nuclear DNA and encodes an NAD⁺-dependent cytosolic GAPDH, will be amplified and sequenced. The section of the gene to be amplified encodes around two-thirds of the protein, including the active site of the enzyme. The technique of nested PCR will be used to amplify this portion of the gene.

All variants of PCR involve a repetitive series of reaction cycles, each of which consists of DNA denaturation to separate the strands of the double helix, annealing (binding) of pairs of primers (short oligonucleotides with a sequence complementary to the target gene of interest) to the two strands, and extension of the annealed primers using a heat-stable DNA polymerase. The primers are designed so that one binds on either end of the target gene sequence. After amplification, the PCR products of this extension process will be double-stranded copies of the gene sequence between the two primers. The primers for genes that have not yet been sequenced are typically designed to bind the regions of highest expected sequence homology to related genes. Genes that code for the same protein in different organisms are likely to have amino acid sequences that are conserved, meaning that they

are very similar or even identical in the different species. These conserved sequences usually code for parts of the protein that are essential for its function; for example, active sites typically have very high sequence homology.

Despite the high amino acid sequence homology within conserved sequences, it is not possible to predict the corresponding nucleic acid sequence that encodes it because the genetic code is redundant. More than one codon or set of three nucleic acids can code for most amino acids. To get around this problem, degenerate primers can be used. These are mixtures of primers that differ in exact nucleic acid sequence but code for the same conserved amino acid sequence. In theory, at least one pair of the primers in the mix will anneal to the target gene sequence, and others may anneal imperfectly, but sufficiently to allow extension of a PCR product. Of course, some of the primers may also anneal to other sequences in the DNA sample, resulting in unwanted PCR products as well. The technique of nested PCR is often used when degenerate primers are used because it allows more specificity in amplification of target sequences and increases the efficiency of the PCR. Nested PCR uses two sequential rounds of PCR with different primer pairs; the second-round set is nested within the first set. In this way, if the first primers amplify an unwanted DNA sequence, it is very unlikely that the second set of primers will also bind within the unwanted region. For more information on PCR, primer design, and nested PCR, refer to Chapter 2.

While the GAPDH protein sequence is highly homologous among family members and species, there is variability in the number, location, sequence, and length of introns (sequences that do not actually code for amino acids in the final protein) in the species examined to date. Such differences in gene structure are evident within the *Arabidopsis* GAPC gene family, where two introns present in the other family members are missing in this gene. This variability in gene structure results in PCR products of different lengths that can be identified by agarose gel electrophoresis. Studies of other plant species during the development of this laboratory series identified numerous other instances of absent introns.



Gene structure of the *Arabidopsis* GAPC family of genes. Blue bars indicate coding sequence (exons). *GAPC* differs from the rest of the family by the absence of two introns (noncoding sequence, indicated by lines), which shortens the gene. The *GAPCP* subfamily of genes has a signal peptide at the N-terminus that directs the protein to plastids. Arrows indicate annealing positions of first-round (outer arrows) and second-round nested (inner arrows) *GAPDH* PCR primers on structure of *GAPC* family genes. The green bar indicates the location encoding the enzyme active site. **Note:** Figure is not to scale.

Since the first-round primers used in this laboratory are degenerate and were designed based on a consensus sequence derived from a number of *GAPC* genes (including those encoding isozymes such as *GAPC* and *GAPCP*; see table above), they may anneal to the target DNA at several locations resulting in multiple bands of amplified DNA seen on an agarose gel after the initial round of PCR. In addition, the degenerate primers may anneal *GAPDH* genes other than *GAPC*, or unrelated sequences that have a high degree of complementarity to one or more of the degenerate primers.

The nested primers were designed to be more specific to *GAPC* and *GAPC-2* (rather than the *GAPCP* subfamily) and are not degenerate, so in the second round of nested PCR, only the *GAPC* or *GAPC-2* genes from the initial PCR should be amplified. For example, if the plant gDNA used in this laboratory is from *Arabidopsis*, then the nested PCR should amplify only *GAPC* and *GAPC-2*. The nested primers should not bind to DNA coding for GAPDH isozymes or to unrelated DNA sequences, so that DNA should not be amplified.

The sizes of PCR products expected from *Arabidopsis* *GAPC* family genes using the primers in this laboratory are shown below. Results will vary for other plant species, but in theory the nested PCR reaction should result in fewer bands (ideally one) that are also shorter than the ones obtained with the initial primers, and are visible on an agarose gel. The nested PCR products will then be cloned, sequenced, and analyzed in the remainder of the laboratory.

Length of *Arabidopsis* PCR product (bp)

<i>Arabidopsis</i> GAPC Gene	Length of Initial primers	PCR product (bp) Nested primers
<i>GAPC</i>	1,065	993
<i>GAPC-2</i>	1,216	1,145
<i>GAPCP-1</i>	1,303	1,231
<i>GAPCP-2</i>	1,205	1,133

Note: The pGAP control plasmid provided with the kit contains the sequence for the initial PCR product of the *GAPC* gene.

CHAPTER 1: NUCLEIC ACID EXTRACTION

Background

Why Extract DNA?

Cloning is the production of multiple exact copies of a piece of DNA, usually a gene, using the techniques of molecular biology. Cloning is frequently the first technique used in a research project, producing enough DNA for further study, and the first step in cloning is isolating the DNA that encodes the gene of interest from the organism that has it.

DNA is found in cells of both eukaryotes (animals, plants, and yeasts) and prokaryotes (bacteria). DNA is only one component of cells, which contain membranes, organelles, other nucleic acids, proteins, and many other chemical compounds. In cells, DNA is closely associated with proteins. The objective of DNA extraction is to separate the DNA from other cellular components and to remove contaminants from the DNA. The DNA must also be intact after extraction.

The barriers that need to be overcome to get to the DNA of interest depend on the cell type. For example, plant cells have a cell wall in addition to an outer cell membrane, and all eukaryotic cells also have membranes around the nucleus and other organelles. These must be disrupted to get to the DNA.

Eukaryotic cells contain most of their genes in the genomic DNA (gDNA) located in the chromosomes of the nucleus, with a few genes in two other types of organelles: mitochondria and (in plant cells) chloroplasts. Therefore, most protocols used to isolate DNA are designed to purify gDNA.

Bacteria can have, in addition to chromosomal DNA, small DNA elements called plasmids, which typically carry only one or a few genes, such as genes for antibiotic resistance. Plasmids are capable of being replicated independently of chromosomal DNA and transferred to other bacteria, characteristics that are useful in cloning. After isolating a gene from an organism of interest, it is generally introduced into a plasmid to allow cloning of the DNA.

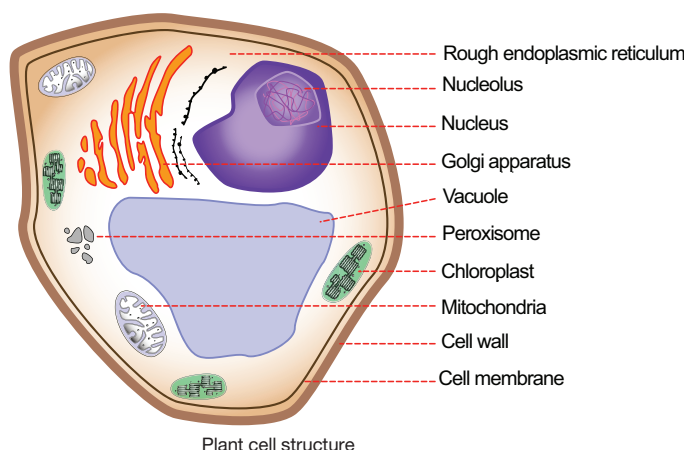
The basic steps in DNA extraction are:

- Grow (or collect) and concentrate cells
- Disrupt cell membranes and cell walls by chemical or physical means
- Remove cellular debris
- Digest remaining cellular proteins
- Purify DNA

All of these steps have variations depending on the cell type, how pure the DNA must be, and whether the gene to be cloned is encoded in the chromosomal DNA or is found in an organelle or on a plasmid.

Considerations in Purifying DNA from Plants

Plants are multicellular eukaryotic organisms distinguished by rigid cell walls, which maintain their shape and provide structural support, and chloroplasts, which carry out photosynthesis. Thus, extracting DNA from plant cells has the same challenges as extracting DNA from any other cell type, with a few additional complications. In addition to lysing the cells, the cell walls must be disrupted. Cell walls are composed of multiple layers of cellulose and other polysaccharides, and polysaccharides colocalize with DNA during the purification process, making it more difficult to separate the DNA from other components. In addition, the process of purifying the DNA could expose the DNA to damage. As in other types of cells, there are enzymes that can digest the DNA (nucleases) during extraction, and some organelles, such as plant vacuoles that are easily disruptable, can release acidic contents that can damage DNA, so the extraction protocol must minimize this damage. Some plant types may present additional complications. For example, many conifers and fruit trees contain a high concentration of compounds called polyphenols. When plant cells containing polyphenols are lysed without special precautions, the polyphenols bind irreversibly to the DNA, which makes it useless for experiments.



Most plant genes are encoded in the gDNA found in the nucleus, even if they code for proteins in organelles. However, if the gene of interest is encoded by mitochondrial or chloroplast DNA, the DNA purification protocol would need to be modified. The primary difference between gDNA and organelle DNA is the size of the DNA. The protocols for this lab are designed to isolate very long DNA molecules, because the *GAPC* gene is encoded in gDNA.

The steps in isolating gDNA from plants are:

1. Harvest cells	Harvest fresh plant tissue.
2. Grind cells	Physically disrupt plant cell tissue.
3. Lyse cells	Lyse cells in solution containing enzymes and chemicals that disrupt cell membranes.
4. Remove cellular debris	Remove insoluble cell components, such as the remains of the cell wall and membranes, by centrifugation. The DNA will be in the supernatant.
5. Purify DNA	Remove contaminating proteins and RNA, which will be in the supernatant with the DNA, using ion exchange chromatography, binding to silica, or enzyme treatment and extraction.
6. Concentrate DNA	Precipitate DNA using ice-cold ethanol in the presence of ions such as Na^+ and K^+ . If there is enough DNA present, it can be seen as a white precipitate. The DNA can be pelleted by centrifugation and then resuspended in a small volume of water or buffer.
7. Determine purity and concentration of DNA	Determine the concentration of DNA by reading the absorbance of the DNA solution at 260 nm in a spectrophotometer. Additionally, the ratio of the solution's absorbance at 230, 260, 280, and 330 nm can indicate the extent of contamination with proteins, phenol, or RNA.

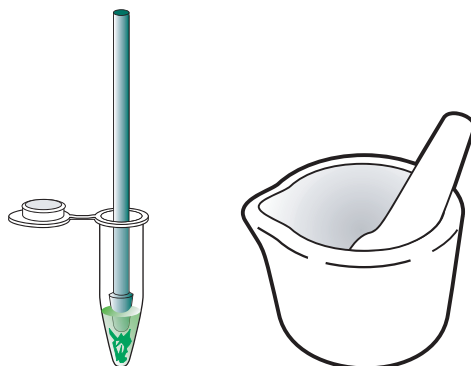
Choice of Plant Tissue

The plant species most suitable for this lab will depend on the teaching goals for the course. For general guidelines on choosing the type of plant to study, refer to the Advice on Which Plant Species to Choose section in the Introduction.

- Regardless of plant species, young leaves are the best source for gDNA. Young leaves are still growing and have a greater ratio of nuclear volume to cytoplasmic volume, and younger leaves have fewer chemicals that interfere with DNA isolation.
- Fresh samples will have higher yields of DNA than preserved or dried leaves.
- While fresh tissue is best, if necessary, plant tissue can be stored for 2–3 days at 4°C in a sealed bag or tube. Do not put tissue samples into the –20°C or –80°C freezer for storage — slow freezing of tissue will disrupt cells, release DNases, and may degrade the genomic DNA.
- If the plant can be moved, placing it in the dark for 1–2 days before harvesting tissue reduces the amount of polysaccharide.

Grinding and Lysing Plant Tissue

Because of the rigid cell walls, plant tissue must be physically ground or crushed prior to DNA isolation. The most common way to grind plant tissue is with a mortar and pestle, although blenders and mechanical tissue grinders are also used with some plant species. There are different sizes of grinders available for crushing different amounts of plant tissue, from large porcelain mortars and pestles to small microcentrifuge tube grinders such as micropestles.



Different types of grinders. Micropestles (left) are intended for use with small samples and porcelain mortars and pestles (right) for use with large samples.

Grinding of plant tissue is frequently done in the presence of lysis buffer, as in this protocol. Alternatively tissue can be flash frozen and ground in liquid nitrogen. If the tissue is frozen, the ground plant material is added to lysis buffer before it can thaw. This prevents damage to the DNA by the cell contents. Plant lysis buffer typically contains ethylenediamine tetraacetate (EDTA), which serves a dual function. EDTA removes magnesium ions by chelation (binding them), destabilizing the cell wall and the cell membrane, and it also inhibits nucleases, enzymes that could digest the DNA.

Lysis buffer must have buffering capacity. When cells are lysed, acidic compounds are released from vacuoles, so the lysis buffer must be formulated to ensure that the lysate does not change its pH dramatically. Most lysis solutions are prepared at pH 8.0 with Tris (tris[hydroxymethyl] aminomethane), a commonly used buffer in molecular biology.

The plant cell outer membrane is selectively permeable, allowing some molecules to pass through while blocking others. This membrane, like other cell membranes, is composed of phospholipids (a type of fat) and proteins. The cell membranes need to be disrupted, and this can be done with detergents or chaotropic agents. Detergents break up cell membranes by removing lipid molecules from the membranes. The choice of detergent depends on the application. Ionic detergents, such as sodium dodecyl sulfate (SDS) or Sarkosyl (N-laurylsarcosine), are commonly used, but nonionic detergents such as Triton X-100 may also be used. Nonionic detergents are milder than ionic detergents and usually leave proteins intact and functional. Chaotropic agents disrupt proteins and membranes by destabilizing their three-dimensional structure. Frequently, the lysis buffer also contains a reducing agent, such as dithiothreitol (DTT) or β -mercaptoethanol, to minimize protein oxidation. These reagents are added shortly before extraction, since they break down quickly.

The classic method of DNA extraction from plants uses a nonionic detergent, cetyltrimethylammonium bromide (CTAB), to lyse the plant cells and precipitate the DNA, leaving most polysaccharides in solution.

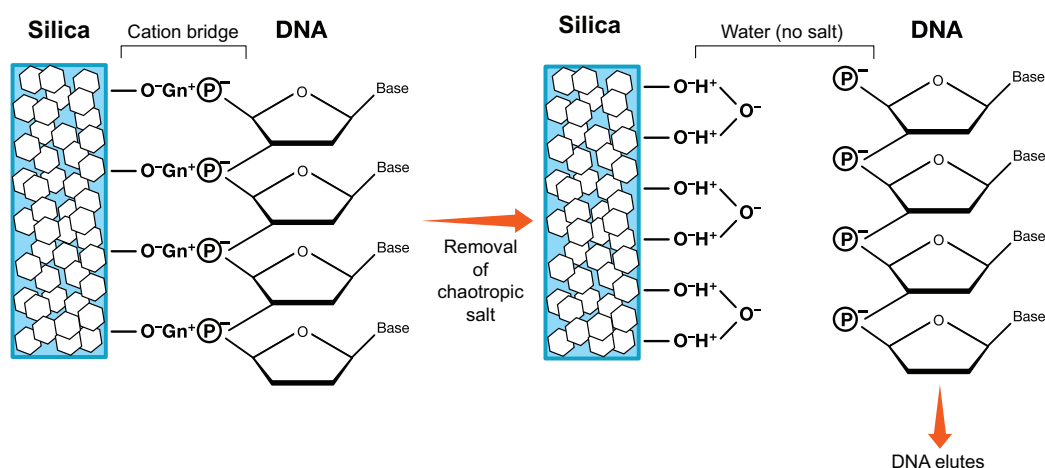
Because of special problems found in some plants, specialized techniques have been developed for DNA extraction. For example, for plants with high levels of polyphenols, including grape species, many fruit trees, and conifers, lysis buffer containing polyvinylpyrrolidone (PVP) and a high concentration of salt works best for DNA isolation. PVP binds the polyphenols, preventing them from complexing with the DNA, and the high salt reduces the coprecipitation of polysaccharides with the DNA.

Purifying DNA

It is important to remember that during the purification of DNA the lysate should never be vortexed or mixed vigorously, as that would result in breaking the large molecules of gDNA into smaller pieces (known as shearing of DNA). Intact, full-length DNA molecules are required for subsequent steps in the experiment. If DNA becomes sheared, your gene of interest may be broken into multiple pieces and therefore may not be amplifiable.

After cell lysis and centrifugation to remove cellular debris, proteins and RNA will be in the supernatant with the DNA. Several gentle methods can be used to remove these contaminants.

- Silica-based purification relies on the strong binding of DNA to silica in the presence of high concentrations of chaotropic salts, salts like guanidine that disrupt hydrophobic interactions. The mechanism of binding is not fully understood. One theory is that the binding is due to the exposure of anions on the DNA and silica as a result of dehydration by the salts. The phosphates on the DNA bind to the silica through the formation of a cation bridge formed by the salt. The DNA can be released from the silica by reducing the salt concentration. This chemistry is the basis for many of the commercially available kits for DNA purification, including the kit used in this laboratory activity.
- Purification of DNA by ion exchange chromatography uses a positively charged resin or other matrix, usually in a chromatography column. DNA and RNA are both negatively charged and will bind to the positively charged matrix through ionic interactions. Other components of the cell lysate that are not negatively charged will not bind to the matrix and will be discarded with the liquid that flows through the column when the sample is applied, or during subsequent washing steps.



Proposed mechanism of DNA binding to silica column. DNA in high salt conditions binds to silica through formation of a cation bridge. In low salt conditions there is no cation bridge and DNA elutes from the column. Gn⁺ stands for guanidine.

Molecules bound to the matrix can be removed (eluted) by increasing the salt concentration of the solution on the column. Molecules elute based on the strength of their binding to the column, with more weakly bound molecules eluting at lower salt concentration than more strongly bound molecules. Using this differential elution, DNA can be separated from proteins and RNA.

- Purification of DNA can rely on a two-step process of enzyme treatment and organic extraction. Proteins in the cell lysate are degraded by treatment with proteases, enzymes that specifically break down proteins. Besides getting rid of protein contaminants, this step also degrades enzymes called nucleases that might destroy the target DNA. Proteinase K is one of the most commonly used proteases, degrading most proteins and inactivating enzymes under a broad range of conditions.

After the proteins have been degraded, organic extraction is used to precipitate proteins in the lysate. Phenol or phenol mixed with chloroform will cause proteins to coagulate at the interface of the organic and the aqueous solutions. The proteins are removed by centrifugation, leaving RNA and DNA in the aqueous solution. The RNA is removed by treating the solution with enzymes called ribonucleases (RNases) that degrade RNA but do not damage DNA. Ribonuclease A, a nuclease that cleaves only single-stranded RNA molecules, is commonly used.

Concentrating DNA

Many DNA preparations result in a solution that is too dilute for experimental purposes, so the DNA must be concentrated. The most common methods use alcohol. In the presence of high concentrations of salt (for example, NaCl), ethanol or isopropanol will precipitate DNA. The cations neutralize the charge on the phosphate backbone of the DNA and allow the DNA molecules to be closer together. (In aqueous solutions, the strong negative charges of the DNA molecules normally repel each other.) The ions do not bind strongly to DNA in aqueous solution, but when an organic solvent like alcohol is added, the binding becomes much stronger and the DNA-cation complexes precipitate out of solution. If there is enough DNA present, a white precipitate should be visible, but DNA may be present even if a precipitate is not observed.

The precipitated DNA is pelleted by centrifugation, after which the pellet should be washed at least once with 70–80% ethanol to remove any remaining salt. The ethanol concentration used for the wash cannot be 100% ethanol, as salt will not dissolve in pure ethanol, nor can the ethanol concentration be below 70%, as the DNA might be resuspended with the salt and be lost. After washing, the DNA pellet should be dried to evaporate any remaining ethanol and resuspended in water or the desired buffer. The concentration and purity of the resuspended DNA can be determined by spectrophotometry or by fluorometry.

Instructor's Advance Preparation

In this Nucleic Acid Extraction Chapter, students will extract all nucleic acids from the plant cells using tissue grinding, followed by DNA purification using silica-based column chromatography. The final sample will contain gDNA and RNA from the cells.

Nucleic Acid Extraction Checklist

Components from Cloning and Sequencing Explorer Series

	Where Provided	(✓)
Microcentrifuge tubes, 1.5 ml	Nucleic Acid Extraction Module	<input type="checkbox"/>
Microcentrifuge tubes*, multicolor, 2 ml	Nucleic Acid Extraction Module	<input type="checkbox"/>
Capless collection tubes, 2 ml	Nucleic Acid Extraction Module	<input type="checkbox"/>
Micropestles	Nucleic Acid Extraction Module	<input type="checkbox"/>
Sterile water	Nucleic Acid Extraction Module	<input type="checkbox"/>
Lysis buffer	Nucleic Acid Extraction Module	<input type="checkbox"/>
DTT (dithiothreitol)	Nucleic Acid Extraction Module	<input type="checkbox"/>
Wash buffer, low stringency (5x concentrate)	Nucleic Acid Extraction Module	<input type="checkbox"/>
Aurum mini columns, purple**	Nucleic Acid Extraction Module	<input type="checkbox"/>

* Optional: tube colors may be assigned to teams or samples to help keep track of samples.

** The mini columns in both the Nucleic Acid Extraction module and the Aurum Plasmid Mini Purification Module are purple but they are functionally different. Be sure to keep them in their separate labeled bags so they do not get mixed or confused.

Required Accessories (Not Provided)

	Quantity	(✓)
Ethanol, 95–100%, molecular biology grade	300 ml	<input type="checkbox"/>
Distilled deionized water (ddH ₂ O)	6 ml	<input type="checkbox"/>
Water bath at 70°C	1	<input type="checkbox"/>
Balance with weigh paper or weigh boats	1	<input type="checkbox"/>
Microcentrifuge with variable speed setting $\geq 12K \times g$	2	<input type="checkbox"/>
200 μ l adjustable-volume micropipets and aerosol barrier filter tips	12	<input type="checkbox"/>
1,000 μ l adjustable-volume micropipets and aerosol barrier filter tips	12	<input type="checkbox"/>
Tube racks	12	<input type="checkbox"/>
Marking pens	12	<input type="checkbox"/>
Tubes for aliquoting (optional)	36	<input type="checkbox"/>
Razor blades or scalpels	24	<input type="checkbox"/>
–20°C freezer	1	<input type="checkbox"/>
Plants for DNA extraction	2	<input type="checkbox"/>

Optional Steps

Optional analysis of the sample may be performed using standard techniques. Once extracted, DNA can be viewed using agarose gel electrophoresis. Loading ~10 µl of gDNA on a 0.8–1% agarose gel and electrophoresing at 75 V for 1 hour is recommended. gDNA will be seen as a faint band at the top of the gel. RNA will be seen as two major bands much further down the gel.

Note: Some plant extractions may have yielded too little DNA to be visible on a gel. However, even if the DNA is not visible on a gel, it will likely be amplified in the next PCR step. Fluorometry, which specifically quantifies double-stranded nucleic acids, can be used to quantify the DNA. Reading the absorbance of the DNA at 260 nm using spectrophotometry has lower accuracy due to the presence of RNA in the sample preparation. RNA can be removed by treating the sample with RNase I and then precipitating the DNA using standard techniques to remove nucleotides.

Tasks to Perform Prior to the Lab

1. Set water bath to 70°C.
2. Place tubes of sterile water for elution of DNA into water bath at 70°C at start of lab.
3. Prepare wash buffer.

Add 95–100% ethanol to the 5x low stringency wash buffer prior to use in the extraction. The amount of ethanol to add is indicated on the wash buffer label. Prepared wash buffer can be stored at room temperature for up to 1 year.

4. Prepare lysis buffer.

Add 0.3 g of DTT to the lysis buffer (20 ml) for a final concentration of 100 mM DTT. Once DTT has been added, excess lysis buffer should be stored at –20°C to remain usable for up to 2 months. If longer storage is required, prepare only enough lysis buffer for the current lab. Store excess lysis solution (without DTT) at room temperature and excess DTT at 4°C.

5. Prepare 70% ethanol.

Add 14 ml of 95–100% ethanol to 6 ml of distilled deionized water (ddH₂O).

6. Obtain plant material — see Advice on Which Plant Species to Choose in the Introduction. The best results will be obtained from young, fresh leaves. However, other parts of plants have been used with success. It is important **not** to put the fresh plant tissue in the freezer for storage. If tissue needs to be stored, keep at 4°C for up to 3 days in a sealed container and ensure it is kept moist.

Protocol

Overview

This project is an opportunity to perform novel research — to clone and sequence a gene that has not yet been analyzed and to add to the body of scientific knowledge around the world. The first step in this exercise is to choose an interesting plant species. Some model species that plant biologists study, for example, *Arabidopsis thaliana*, the green alga *Chlamydomonas*, or crop plants like rice (*Oryza* species) and wheat (*Triticum* species), have already had their genomes sequenced, so you may be reproducing and confirming this data.

Alternatively, you may be examining a

species, variety, or cultivar that does not appear to have been studied well. There are over 250,000 known plant species, providing plenty of options with which to work.

In order to clone a gene from an organism, DNA must first be isolated from that organism. In this module, you will isolate genomic DNA (gDNA) from one or two plants using column chromatography. Plant material is first weighed, then ground in lysis buffer with high salt and enzyme inhibitors using a micropestle. The solid plant material is removed by centrifugation, then ethanol is added to the lysate and the lysate is applied to the column. In the presence of ethanol and high salt solution, the DNA will bind to the silica in the chromatography column. The column is then washed three times to remove contaminating molecules before the DNA is eluted using sterile water warmed to 70°C.

Safety Issues

Eating, drinking, smoking, and applying cosmetics are not permitted in the work area. Wearing protective personal equipment such as eyewear, gloves, and labcoats should be standard laboratory practice. The lysis buffer in this product contains guanidine thiocyanate (CAS# 593-84-0) in solution. Basic laboratory practices should be followed to avoid contact with eyes, skin, and clothing. Wash your hands with soap and water before and after this exercise. If any of the solution gets into eyes, flush with water for 15 minutes. If these buffers are spilled, clean with a suitable laboratory detergent and water. Avoid contact with acids or bleach as these will liberate toxic gas. Please refer to material safety data sheet for complete safety information.

Cloning the GAPC gene

• Identify and extract gDNA from plants

- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items for DNA extraction from two plants.

Material Needed for Each Workstation	Quantity
Razor blades or scalpels	2
Microcentrifuge tubes, 1.5 ml	2
Microcentrifuge tubes, multicolor, 2 ml	4
Capless collection tubes, 2 ml	2
Micropestles	2
Sterile water (at 70°C)	200 µl
Lysis buffer (prepared with DTT)	1.5 ml
1x wash buffer (with 95–100% ethanol added)	5 ml
Aurum mini columns, purple	2
1,000 µl adjustable micropipet and aerosol barrier tips	1
200 µl adjustable micropipet and aerosol barrier tips	1
Marking pen	1
70% ethanol	2 ml
Plant tissue samples	2
Tube racks	1

Common Workstation

Material Required	Quantity
Balance with weigh paper or weigh boats	1
Water bath at 70°C (place tubes of sterile water into the water bath for elution of DNA)	1
Microcentrifuges with maximum speed $\geq 12,000$ g	1

Experimental Procedure for DNA Extraction

Ensure that appropriate personal safety equipment such as gloves, goggles, and lab coat is worn during this laboratory activity.

Select one or two plants (very little material is required for the DNA extraction). The younger the plant tissue, the better the DNA yield. If necessary, clean the plant material to remove soil or debris.

Ensure that sterile water for elution of DNA is at 70°C.

1. Label two 1.5 ml microcentrifuge tubes with your initials and plant name.
2. Pipet 200 µl of lysis buffer into both 1.5 ml tubes.

Lysis buffer contains protective agents (inhibitors) and a buffer to maintain pH.

3. For each plant, weigh 50–100 mg of plant tissue. Record the weight of the tissue.

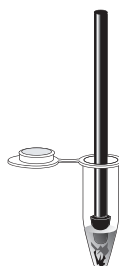
Note: For most leaves 50–75 mg of tissue is sufficient. For plant tissue that is high in water content, such as cabbage leaves, use 75–100 mg.

Plant Name	Part of Plant (Leaf, Root, etc.)	Weight (mg)

4. For each plant, use a razor blade or scalpel to chop the plant tissue into small pieces. Try to make the pieces less than 1–2 mm in diameter. (**Note:** Use a new razor blade or scalpel for each plant type.) Add the chopped plant material to the lysis buffer in the microcentrifuge tube.
5. For each plant, use a clean micropestle to grind the plant tissue for at least 3 min. Be careful not to let the lysis buffer spill over the side of the tube, which would result in loss of sample. Move the pestle up and down as well as twisting it to ensure good grinding. If material compacts at the bottom of the tube, a clean pipet tip can be used to dislodge it and permit further grinding. Prior to moving to the next step, check that the plant tissue has been ground to very fine particles (that is, particles difficult to see by eye, rather than visible chunks), even if this requires further grinding.



The mechanical action lyses the cellular components of the cell and releases the DNA.



6. Once a homogeneous lysate is generated, add an additional 500 μ l of lysis buffer and grind further using the micropestle until the lysate is homogeneous. Inefficient tissue lysis may result in lower DNA yield.
7. Cap (close) the microcentrifuge tube. Spin for 5 min at full speed in a microcentrifuge at room temperature. Make sure that the tubes are placed in the rotor so that the microcentrifuge is balanced; accommodate classmates' tubes to ensure economic use of the microcentrifuge.

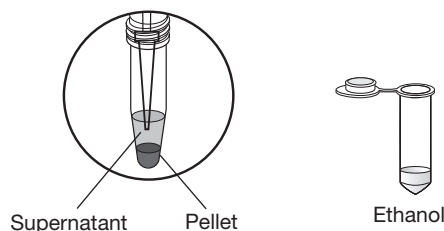
This step pellets the cell debris.



8. While tubes are centrifuging, label one microcentrifuge tube of a different color for each extract and add 500 μ l of 70% ethanol to each tube.

The ethanol will help the DNA bind to the column.

9. Retrieve samples from microcentrifuge. For each sample, carefully remove 400 μ l of supernatant (taking care not to disturb the pellet) and add it to the 500 μ l of 70% ethanol in the appropriately labeled tube. Avoid transferring any solid plant material to the ethanol; if necessary, recentrifuge the lysate. Pipet up and down to thoroughly mix the lysate and ethanol into a homogeneous solution. Cap tubes.

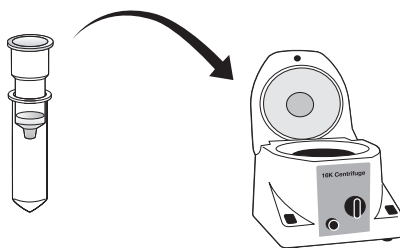


Note: If a precipitate is visible (common in starchy sources such as potato), spin the microcentrifuge tube again for 5 min in a microcentrifuge at full speed to pellet the precipitated starch and use the supernatant for the next stages.

10. Label the top outside edge of a mini column for each sample with your initials and plant name. Place columns in 2 ml capless collection tubes.
11. For each sample, pipet 800 μ l of cleared plant lysate and apply to column.

Note: Be sure that there is no plant tissue in the lysate that is added to the column. If necessary, centrifuge the lysate again to pellet any plant material prior to loading the lysate on the column.

In this step, the DNA binds to the silica membrane filter while protein, polysaccharides, and lipids flow through.



12. Place the capless collection tubes containing the columns into the microcentrifuge. Ensure that the centrifuge is balanced. Spin for 1 min at full speed at room temperature. Discard flowthrough from the collection tubes.

Note: If all of the supernatant does not pass through column, centrifuge again for 1 min. If there is still supernatant remaining in the column, carefully remove the excess supernatant with a pipet and discard, taking care not to disrupt the bed of the column, then continue to the next step. Further centrifugation will not help. Some plant samples can block the column with either carbohydrate or pigments. Even when blockage occurs, it is likely that some DNA will have bound to the column matrix and, although yield will be lowered, there will probably be sufficient DNA for the rest of the experiment.

13. Add 700 μ l of wash buffer to each column. Spin at full speed at room temperature for 1 min. Discard flowthrough. Repeat wash step 2 more times for a total of 3 washes. Check the box for each wash.

These wash steps help to remove contaminants such as polysaccharides and proteins.

- Wash 1 ☐
- Wash 2 ☐
- Wash 3 ☐



14. After the final wash step, discard flowthrough and replace the columns in the capless collection tubes. Dry the columns by centrifuging for 2 min at full speed in the microcentrifuge at room temperature.

This step is vital to remove the residual ethanol because it will interfere with PCR.

Final spin ☐



15. Transfer each column to a fresh, appropriately labeled color microcentrifuge tube.

16. Obtain the sterile water from the 70°C water bath. Immediately add 80 µl of 70°C sterile water to the bed of each column, making certain that the water wets the column bed. Let sit for 1 min.

This step will solubilize the DNA.



17. Place the column, still in the microcentrifuge tube, into the microcentrifuge. Orient the loose cap of the microcentrifuge tube downward, toward the center of the rotor, to minimize friction and damage to the cap during centrifugation. Spin at full speed in the microcentrifuge at room temperature for 2 min. Remove the spin column from the microcentrifuge tube. Collect the eluate that contains your extracted genomic DNA and store at -20°C. Be sure that your tubes are labeled as purified genomic DNA with your initials, plant name, and the date.

The purified sample contains both gDNA and RNA from your plant samples. RNA will not interfere with subsequent PCR steps.

Optional: If time permits, you may proceed directly to the *GAPDH* PCR Module and set up PCR reactions using freshly extracted gDNA.

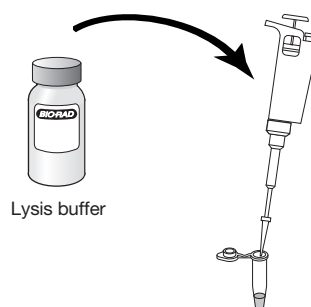
Optional: Depending on time constraints, your instructor may help you analyze your sample prior to the next steps. This analysis may include agarose gel electrophoresis, fluorometry, or spectrophotometry. More information on these options is provided in the Nucleic Acid Extraction Module manual that comes in the kit box.

Focus Questions for Nucleic Acid Extraction

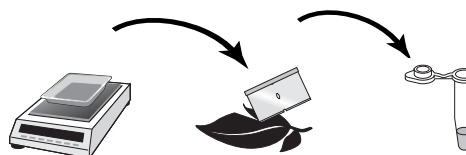
1. Where is DNA found in eukaryotic cells? (Hint: Think of different kinds of cells.)
2. What parts of the cell must be broken down to extract DNA? (Hint: Think about cell structure.)
3. Why is it more difficult to extract DNA from plants?
4. Why are young plants the best source for DNA?
5. Briefly explain how you will achieve the basic steps in the DNA extraction.

Nucleic Acid Extraction — Quick Guide

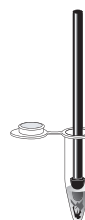
1. Label one 1.5 ml microcentrifuge tube with your initials and plant name for each plant sample.
2. Pipet 200 μ l of lysis buffer into each tube.
3. Weigh 50–100 mg of tissue for each plant and record the mass.



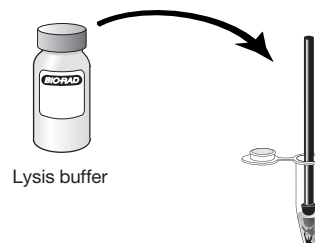
4. For each plant, use a razor blade or scalpel to chop the plant material into 1–2 mm pieces. Use a new razor blade for each sample. Add the chopped material to the lysis solution.



5. For each plant, use a micropestle to grind the plant material for at least 3 min. Use a new micropestle for each sample.



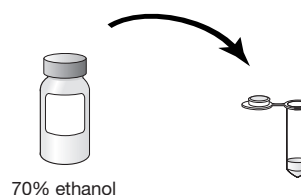
6. Once a homogenous lysate has been generated (That is, chunks of plant material are no longer visible), add 500 μ l of lysis buffer to the lysate. Grind further if homogeneity has not yet been achieved.



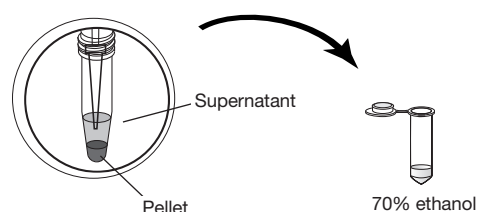
7. Cap microcentrifuge tube and centrifuge for 5 min at top speed.



8. For each plant sample, label a new microcentrifuge tube with your initials and plant name. Add 500 μ l of 70% ethanol into each tube.



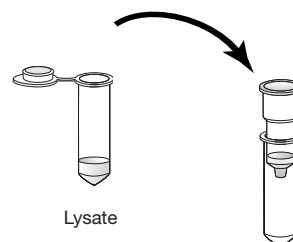
9. Retrieve samples from microcentrifuge and add 400 μ l of the supernatant to the 70% ethanol in the appropriately labeled tube, taking care not to disturb the pellet. Pipet up and down to mix lysate and ethanol. Change pipet tips for each sample.



10. Label the top edge of a purple mini DNA extraction column for each plant and place columns in 2 ml capless collection tubes.



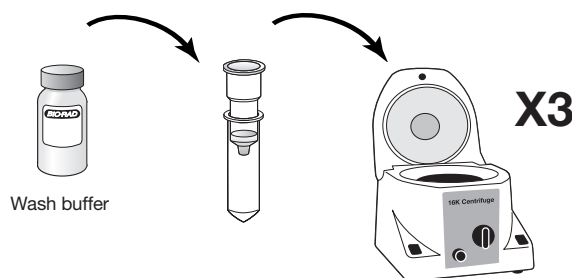
11. For each sample, pipet 800 μ l of cleared plant lysate into the appropriate column. Change pipet tips for each sample.



12. Centrifuge columns for 1 min. Discard the flowthrough from the collection tube and replace the column in the appropriate collection tube.



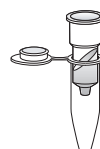
13. Add 700 μ l of wash buffer to each column. Spin at full speed for 1 min. Discard the flowthrough. Repeat 2 more times for a total of 3 washes.



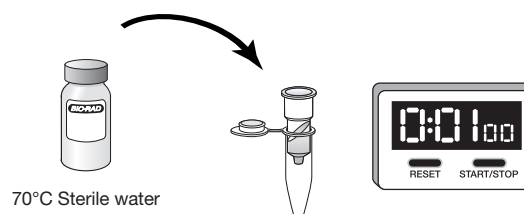
14. After the final wash step, discard the flowthrough and replace the columns in the capless collection tubes. Dry the columns by spinning for 2 min at full speed in microcentrifuge.



15. Transfer each column to a fresh, appropriately labeled, capped microcentrifuge tube.



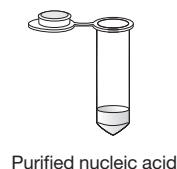
16. Add 80 μ l of 70°C sterile water to the bed of each column. Let sit for 1 min. Change pipet tips for each sample.



17. Spin the columns in the capped microcentrifuge tubes for 2 min at full speed in the microcentrifuge.



18. Discard column and cap the labeled microcentrifuge tube containing purified nucleic acid. Store at -20°C .



CHAPTER 2: *GAPDH* PCR

Background

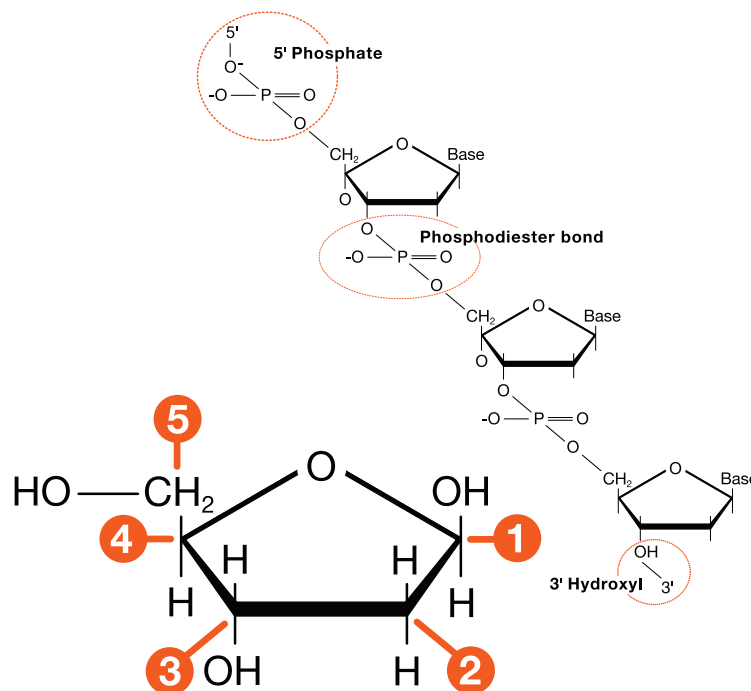
PCR

Polymerase chain reaction (PCR) is a technique for rapidly generating multiple copies of a segment of DNA utilizing repeated cycles of DNA synthesis. PCR has revolutionized molecular biology and forensics, allowing amplification of small quantities of DNA into amounts that can be used for experimentation or for forensic testing. Kary Mullis, who later won a Nobel Prize for his work, developed PCR in 1983. The subsequent discovery of a DNA polymerase that is stable at high temperatures and the introduction of thermal cyclers, instruments that automate the PCR process, brought the procedure into widespread use in the late 1980s.

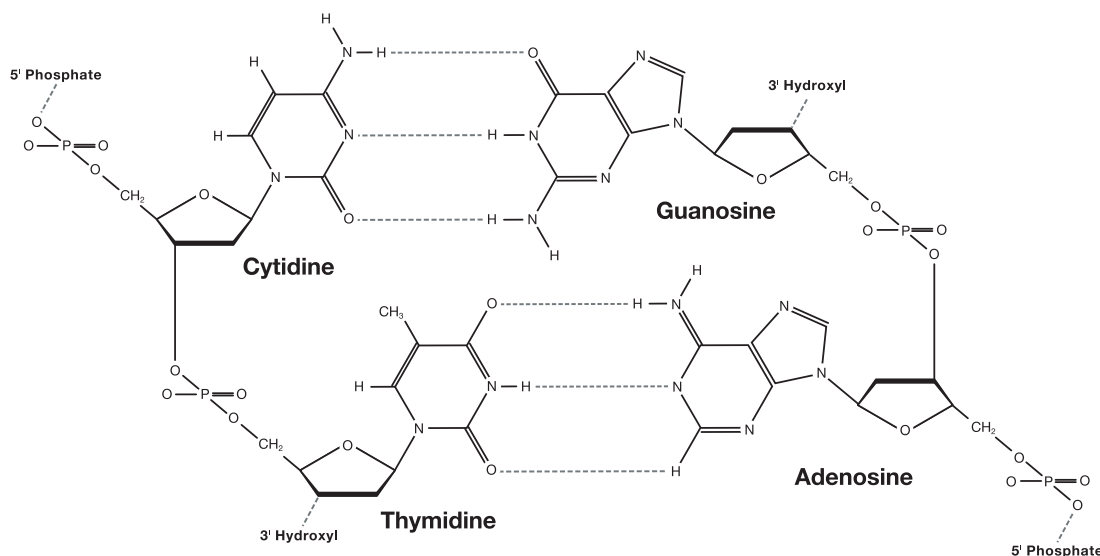
From trace amounts of the DNA used as starting material (template), PCR produces exponentially larger amounts of a specific piece of DNA. The template can be any form of DNA, and only a single molecule of DNA is needed to generate millions of copies. PCR makes use of two normal cellular activities: 1) binding of complementary strands of DNA, and 2) replication of DNA molecules by DNA polymerases.

DNA Structure

DNA strands are polymers of nucleotides, molecules comprising a sugar, a phosphate group, and one of four bases: adenine, thymine, guanine, or cytosine (A, T, G, or C). The sugars and phosphates form the backbone of the DNA polymers. Each sugar has five carbons, making it a pentose. Each sugar is actually a deoxyribose because it has a hydrogen instead of a hydroxyl group at carbon number 2 (in RNA, the sugar is ribose, as it has the hydroxyl group). Each carbon in the sugar is numbered (see figure), and the numbering is the source of the 3' and 5' nomenclature used for DNA. For example, the 5'-phosphate is the phosphate to which the next nucleotide will be attached to the DNA molecule, and it is called 5' because the phosphate group is attached to carbon number 5 of the sugar.



Each base forms hydrogen bonds with its complementary base, A with T (two hydrogen bonds) and G with C (three hydrogen bonds). These pairings of A-T and G-C are called base pairs. Double-stranded DNA consists of two complementary strands of DNA held together by hydrogen bonding between the base pairs. The two strands are antiparallel, meaning that the strands are oriented in opposite directions. One strand, the sense or coding strand, has bases running 5' to 3', and the second strand, the antisense strand, has the complementary bases running 3' to 5'. When DNA is transcribed, the antisense strand serves as the template for synthesis of messenger RNA (mRNA). The mRNA will have the same sequence as the sense or coding strand of DNA (with uracil instead of thymine).



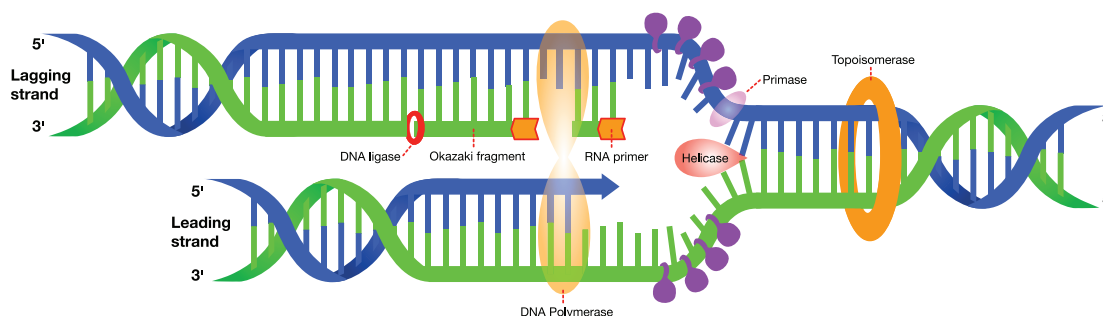
Base pairs of DNA. Cytosine base pairs with guanine by three hydrogen bonds. Likewise thymine base pairs with adenine by two hydrogen bonds.

DNA Replication

DNA replication is an essential part of life. As cells divide, DNA must be duplicated, and the new DNA molecules must be exact copies of the original DNA. DNA polymerases are enzymes that synthesize the new DNA strands, and they are found in all cells. DNA polymerases link together free nucleotides in the order determined by the template DNA that the polymerase follows. The new strand will be complementary to the template strand. In other words, each base of the new strand is the complement of the base in the template strand. For each A in the template, the new strand will have a T. For each G in the template, the new strand will have a C, etc. Since a DNA polymerase can use only single-stranded DNA as a template (and since it can synthesize DNA in only one direction, 5' to 3'), double-stranded DNA must be uncoiled and the strands separated before the DNA can be replicated.

DNA polymerase also needs a signal to determine where to start synthesis. This primer is a short strand of nucleotides that binds to the template DNA at the starting point and becomes the 5' end of the new DNA strand. In DNA replication in cells, the primers are small RNA molecules, but for PCR in the lab, the primers are DNA molecules.

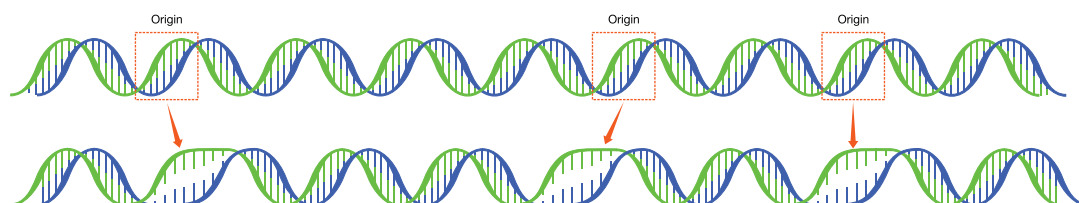
In its essentials, DNA replication sounds simple: unwind the double-stranded template, bind a primer to each strand to give the DNA polymerase a starting point, and the enzyme will produce replicated DNA strands. In reality, the process is much more complicated. There are as many as 40 proteins involved in DNA replication in eukaryotes. Without detailing all of the proteins involved, the basic steps of DNA replication are:



DNA replication fork.

1. Template DNA strands begin to separate at the origin of replication. An enzyme called DNA helicase breaks the hydrogen bonds between the base pairs to separate the strands. The point where the two strands separate is called the replication fork.
2. As the strands unwind and separate, the DNA ahead of the replication fork starts to form supercoils. An enzyme named topoisomerase moves ahead of the replication fork, nicking single strands of the double-stranded DNA and relaxing the supercoiled structure.
3. To keep the two strands from reannealing (binding to each other again), single-stranded DNA-binding proteins bind to each of the separated strands.
4. Since DNA polymerase can add nucleotides only to the 3' end of an existing nucleotide, an enzyme named RNA primase binds to each of the template DNA strands and assembles a short primer of RNA. (The RNA primer will later be removed and replaced by DNA in the new strands.)
5. DNA polymerase begins to synthesize DNA by adding new nucleotides to the RNA primers.

6. Since DNA polymerase can synthesize DNA in only one direction, from 5' to 3', synthesis actually proceeds differently on the two template strands. On the 3' to 5' template strand, called the leading strand, DNA synthesis proceeds continuously, moving toward the replication fork. On the second strand, called the lagging strand, synthesis moves away from the replication fork and is discontinuous. DNA on the lagging strand is synthesized in short pieces (100 to 2,000 bases) called Okazaki fragments. After the Okazaki fragments are synthesized, they are joined together by DNA ligase.
7. Although DNA polymerase is a high-fidelity enzyme, meaning that it makes few mistakes in replicating the bases, it does make some mistakes. In eukaryotic replication, the error rate is one mistake in every 10,000 to 100,000 base pairs. Many DNA polymerases also have proofreading activity, which means that they can find mistakes and correct them as the enzyme moves along the template.
8. DNA replication in eukaryotes does not begin at a single origin of replication, but at numerous locations along a DNA molecule. Origins of replication are found about every 100 kilobases in eukaryotic cells. Mammalian cells are estimated to have ~30,000 origins of replication. In addition, at each origin of replication, actually two replication forks form that head in opposite directions, and, although the description above refers to replication at only one fork, replication occurs simultaneously (and in the opposite direction) at the other fork.



DNA replication uses multiple origins of replication.

Replication moves along the template DNA molecule until the replication fork meets a fork coming from the opposite direction.

Each replication of DNA produces two strands of DNA, each identical to the original strand. Eukaryotic DNA replication is called semiconservative because each double-stranded product consists of one original strand and one newly synthesized strand.

PCR Step by Step

The strength of PCR lies in its ability to make many copies of (amplify) a single region (target) of a longer DNA molecule. For example, a researcher wanting to study a single human gene needs to amplify only that portion from the enormous human genome of approximately 3.3×10^9 base pairs! The first step is to identify and sequence areas upstream and downstream from the DNA of interest. Once this is done, short strands of DNA that are complementary to the upstream and downstream DNA are synthesized. As in cellular DNA replication, these oligonucleotide primers are used as the starting point for copying the DNA of interest, but the primers used in PCR are DNA oligonucleotides, not RNA.

Taq DNA polymerase. Originally, the DNA synthesis step of PCR was performed at 37°C using DNA polymerase from the bacterium *E. Coli*, but the enzyme was inactivated during the high-temperature denaturation step in each cycle. So, the enzyme had to be added anew during each cycle. The 1988 discovery of a thermally stable DNA polymerase brought PCR into the mainstream. *Taq* DNA polymerase was isolated from *Thermus aquaticus*, thermophilic bacteria that live in hot springs in Yellowstone National Park. Since the hot springs frequently approach boiling temperatures, *T. aquaticus* and other bacterial species that live in these waters must have enzymes that are functional at high temperatures, so the DNA polymerase from *T. aquaticus* is not inactivated by the denaturation step in PCR.



Since the discovery of *Taq*, several other heat-stable DNA polymerases have been isolated. *Taq* has a drawback for DNA synthesis in PCR, which is that it lacks a proofreading mechanism to catch and correct errors in the new DNA strand. Therefore, *Taq* is said to have low replication fidelity. In 1991, scientists discovered and characterized *Pfu* DNA polymerase from *Pyrococcus furiosus*, a thermophilic type of Archaeobacteria. *Pfu* DNA polymerase has the proofreading capacity that *Taq* lacks, so *Pfu* generates fewer errors in the new DNA strands. Subsequently, a number of companies have developed modified versions of DNA polymerases. For example, Bio-Rad's iProof polymerase, which is a DNA polymerase with proofreading capability similar to *Pfu*, is fused to a protein that binds double-stranded DNA.

PCR involves a repetitive series of cycles, each of which consists of template denaturation, primer annealing (binding to the template DNA strand), and extension of the annealed primer by a heat-stable DNA polymerase.

Cloning and Sequencing Explorer Series

All of the components needed for PCR are mixed in a microcentrifuge tube. They are:

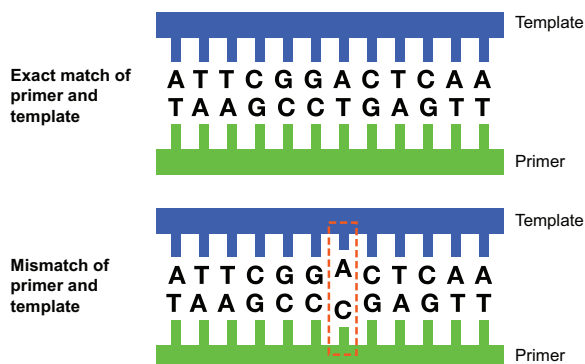
- Template DNA
- *Taq* DNA polymerase (or another thermally stable DNA polymerase)
- Primers — synthesized to complement a specific region on the template DNA. The two primers in a pair are designed to anneal to opposite ends of the region of interest. The primers are added in excess (that is, there are many more primer molecules than template molecules in the reaction tube)
- Nucleotides — the four individual bases in the form of deoxynucleoside triphosphates (dNTPs), which allows them to be added to a DNA polymer. The dNTP mixture includes the same amounts of dATP, dTTP, dGTP, and dCTP
- Reaction buffer — prepared with the correct ionic strength of monovalent and divalent cations needed for the reaction and buffered to maintain the pH needed for enzyme activity

The microcentrifuge tubes are specialized tubes used only for PCR. PCR tubes are plastic with very thin walls, allowing rapid transfer of heat through the plastic, and the tubes usually hold only 0.2 or 0.5 ml. The PCR reaction tubes are placed in a thermal cycler, an instrument developed in 1987 that automates the heating and cooling cycles needed during PCR. Thermal cyclers contain a metal block with holes for the PCR tubes. The metal block can be heated or cooled very rapidly. Thermal cyclers are programmable, so they can store the PCR reaction parameters (temperatures, time at each temperature, and number of cycles). This means that the user can just load the samples and push a button to run the reactions. Contrast this to early researchers who had to sit by a series of water baths with a timer, manually switching the tubes from one temperature to another for hours!

The first step of the PCR reaction is the denaturation step. Since DNA polymerase can use only single-stranded DNA as a template, the first step of PCR is uncoiling and separating the two strands of the template DNA. In cells, enzymes such as helicase and topoisomerase do this work, but in PCR, heat is used to separate the strands. When double-stranded DNA is heated to 95°C, the strands separate, or denature. Since complete denaturation of the template DNA is essential for successful PCR, the first step is frequently an extended denaturation period of 2–5 minutes. The initial denaturation is longer than subsequent denaturation steps because the template DNA molecules are longer than the PCR product molecules that must be denatured in subsequent cycles. Denaturation steps in subsequent PCR cycles are normally 30–60 seconds.

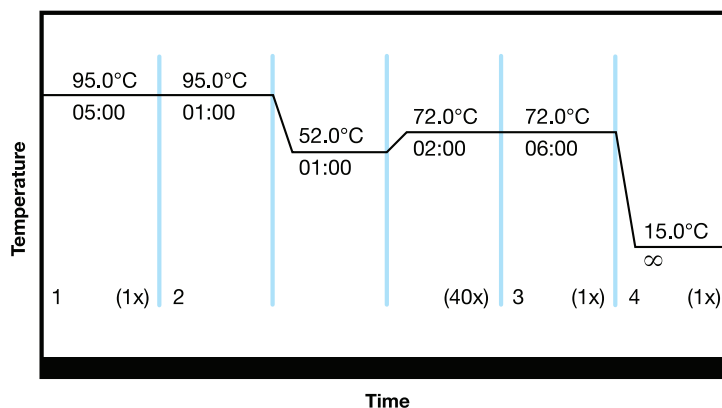
The thermal cycler then rapidly cools the reactions to 40–60°C to allow the primers to anneal to the separated template strands. The temperature at which the primers anneal to the template DNA depends on several factors, including primer length, the G–C content of the primer, and the specificity of the primer for the template DNA. If the primer sequences match the template sequences exactly, the primers will anneal to the template DNA at a higher temperature. As the annealing temperature is lowered, primers will bind to the template DNA at sites where the two strands are not exactly complementary. In many cases, these mismatches will cause the strands to dissociate as the temperature rises after the annealing step, but they can also result in amplification of DNA other than the target.

In the annealing step, the two original strands may reanneal to each other, but the primers are in such excess that they outcompete the original DNA strands for the binding sites.



Mismatched base pairs affect DNA annealing.

The final step is extension, in which the reaction is heated to 72°C, the optimal temperature for *Taq* DNA polymerase to extend the primers and make complete copies of each template DNA strand.

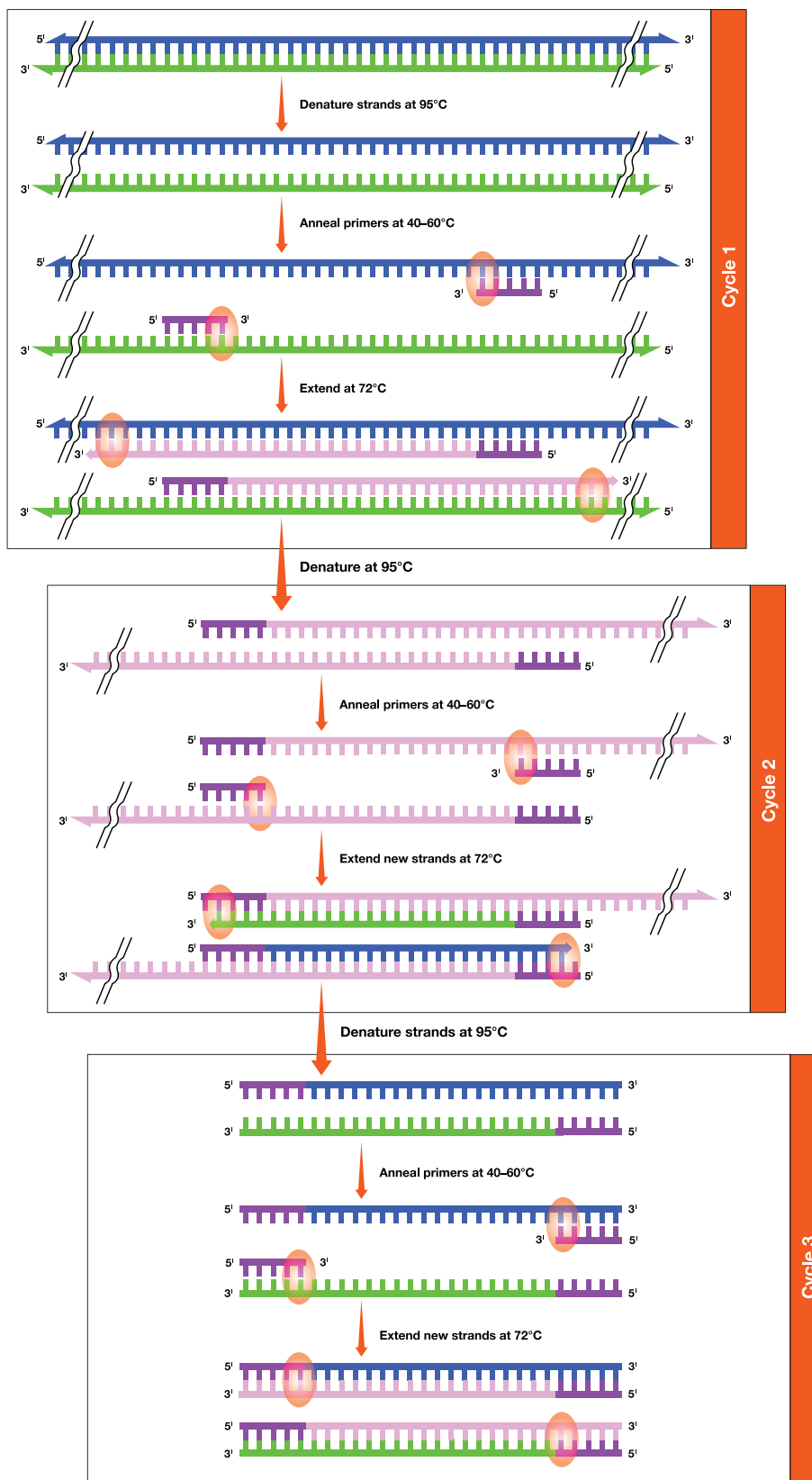


Example of thermal cycling profile. In this profile an initial denaturation step of 95°C for 5 min is followed by 40 cycles of 1-min denaturation, 1-min annealing, and 2-min extension. A final 6-min extension time is added to ensure completion of DNA synthesis. The final hold ensures samples are kept stable until retrieved.

At the end of the first PCR cycle (one round of denaturation, annealing, and extension steps define one cycle), there are two new strands for each original double-stranded template, which means there is twice as much template DNA for the second cycle of PCR. As the cycle is repeated, the number of strands doubles with each reaction. After 35 cycles, there will be over 30 billion times more copies of the target sequence than at the beginning. The number of cycles needed for amplification depends on the amount of template DNA and the efficiency of the reaction, but reactions are frequently run for 30–40 cycles.

PCR generates DNA of a precise length and sequence. During the first cycle, primers anneal to the original template DNA strands at opposite ends and on opposite strands. After the first cycle, two new strands are generated that are shorter than the original template strands but still longer than the target DNA, because the original template sequence continues past the location where the other primer binds. It isn't until the third PCR cycle that fragments of the precise target length are generated.

Cloning and Sequencing Explorer Series



First three cycles of PCR. PCR takes three cycles before a product of the correct length is generated.

Primer Design

Probably the most important variable in PCR is the design of the primers, which will determine whether or not the correct piece of DNA is amplified. Primers are short, single-stranded oligonucleotides, synthesized in the laboratory and designed to bind to the DNA template strands at the ends of the sequence of interest. (Actually, very few laboratories prepare their own primers, as there are many companies that make primers to order — quickly, cheaply, and accurately.) Primer design is usually the responsibility of the researcher, although there are computer programs and websites that assist in this process. Normally, two different primers are needed, one for each of the complementary strands of template DNA.

Factors to be considered in designing primers include:

- Length — primers of 18–30 nucleotides are likely to be specific for their target sequence. In other words, primers of that length are less likely to bind to sites on the template DNA other than the sites for which they were designed
- Melting temperature of primers (T_m) — the T_m is the temperature at which half the primers dissociate from the target DNA. It is important that the two primers used in each PCR reaction have similar melting temperatures (within 5°C). If the T_m values are very different, the primers will not bind equally during the annealing stage. T_m is a function of length and GC content, because more energy is required to dissociate the three hydrogen bonds between G and C compared to the energy to dissociate the two hydrogen bonds between A and T. The T_m for primers around 18–24 bases in length can be estimated from their nucleotide content using the formula:

$$T_m = 2^\circ\text{C} (A+T) + 4^\circ\text{C} (G+C)$$

Another formula for T_m determination is:

$$T_m = 81.5 + 16.6 (\log_{10}[I]) + 0.41 (\%G+C) - 600/n$$

where I is the molar concentration of monovalent cations and n is the number of bases in the primer. This formula gives accurate T_m in °C for primers from 20 to 100 bases long.

There are many website tools that will calculate the exact T_m for primers*.

- Annealing temperature — the temperature for the annealing step of PCR should be about 5°C below the T_m of the primers. Primers should generally have an annealing temperature of ~50–60°C
- GC content — the primer should be composed of 40–60% Gs and Cs. Primers with higher GC content require more energy and thus a high T_m for PCR. (If the T_m is too high, the annealing temperature can exceed the optimal temperature for *Taq* polymerase extension of the DNA strand.) In addition, long stretches of any single base may lead to gaps, hairpin structures, or mismatches, and should be avoided. An ideal primer would have a random mix of bases with ~50% GC content

* Free calculators include OligoCalc (hosted at Northwestern University, University of Pittsburgh, JustBio.com and others) and PrimerFox (primerfox.com). There are also calculators available on many biotech company websites and some fee-based sites.

- Intra- or inter-primer complementarity — primers should not have any regions of complementarity longer than three bases. Otherwise, they can form hairpins by internal annealing or generate double-stranded structures that will interfere with PCR. Also, it is very important that there not be complementarity at the 3' ends of the two primers. If primers hybridize at their 3' ends, the hybrid molecule can act as a template for DNA polymerase, resulting in an unwanted PCR product called a primer-dimer. Primer-dimers are more likely to be produced when the primers do not bind efficiently to the template DNA
- GC-clamp — the sequence of the primers at the 3' end is important to ensure correct and strong binding of the primer to the template. If the primers contain GC clamps, which are 1–3 G or C bases at the 3' end of the primer, they will form a more stable complex with the template DNA

Testing Primers in PCR

Although one should spend time and energy designing PCR primers, there are no guarantees that a well-designed primer will work. The only way to find out if a primer will actually amplify template DNA is to test it in a PCR reaction. Frequently primers that appear less efficient on paper work better than seemingly perfect primers. Researchers will frequently design multiple primers and spend a lot of time optimizing the PCR to find the best primer pair.

Designing Degenerate Primers from Consensus DNA Sequences

Normally PCR primers are designed based on the known sequence of the target DNA, and therefore consist of a single, unique sequence. When the sequence of the template DNA is not known, there are several alternative approaches for primer design. One approach is to take advantage of genetic homology among closely related organisms. For example, the target DNA may not have been sequenced in the species of interest, but the gene may have been sequenced in several other species. Genes that code for the same protein in different organisms are likely to have sequences that are conserved, therefore very similar or even identical in the different species. These conserved sequences usually code for parts of the protein that are essential for function; in other words, mutations in these areas are likely to be detrimental to the organism, so evolution discourages any changes.

If genomic DNA (gDNA) or messenger RNA (mRNA) sequences from similar species are aligned, a consensus sequence can be derived. The consensus sequence may be exactly the same in all species, or it may have one or more bases that vary among the species. For example, a consensus sequence could be represented by A-C-T-G-G-N-T-T-A-C-C-G, where A, C, G, and T represent the bases that are the same in all of the species compared, and N represents a base that varies in different species. In other words, the base at the N position might be G, C, A or T.

Since the goal of PCR is to amplify the DNA region of interest, primers are designed to bracket that region. Once the two primers have been designed based on the consensus sequences derived from other organisms, it is possible that they will have enough complementarity with the target DNA to bind during the annealing step. However, to increase the probability that the primers will bind to the target DNA, one or more bases within the primers is substituted with the other three bases, introducing degeneracy, or wobble, to the primer sequences. In a simplified example, if the consensus sequence is NATC, the set of degenerate primers would be AATC, TATC, GATC, and CATC.

However, in many cases, not all of the bases are used to substitute for the variable base. To increase the probability that the primer will anneal to the target DNA, the variable base is substituted with a similar base. For example, if the variable base is T, it might be replaced only with C (the other pyrimidine). There is a code from the International Union of Biochemistry (IUB) used to tell the company synthesizing the primers which bases to substitute at each variable position shown in the table below:

IUB codes.

IUB Code	Bases	Derivation of IUB Code
N	A/G/C/T	Any
K	G/T	Keto
S	G/C	Strong
Y	T/C	Pyrimidine
M	A/C	Amino
W	A/T	Weak
R	G/A	Purine
B	G/T/C	—
D	G/A/T	—
H	A/C/T	—
V	G/C/A	—

The following table shows how alignment of *GAPC* genes from different plant species can be used to derive a consensus sequence that can then be used for primer design. The plant species are listed on the left and the *GAPC* genes are aligned on the right. The vertical highlighting in the sequences shows bases conserved across all the species. Deriving the consensus sequence for the gene begins with the conserved bases. For example, all of the sequences begin with GA, so the consensus sequence will also begin with GA. Twelve out of the 23 bases do not vary between plant species and are highlighted.

Although the other bases are not conserved, the differences between the species are not random. For example, the bases in positions 4 and 5 are always either A or T. (A and T are considered weak bases, as the base pairs they form are not as strong as those formed by G and C). The base that is more commonly found in that position will be the one used in the consensus sequence; that is, position 5 will be A, as A is found in 15 of the 19 sequences.

After most of the consensus sequence has been determined, degeneracy can be introduced at one or more positions, normally at the positions that show the most variability among species, but this can also depend on experimental optimization. Degeneracy cannot be introduced at each base where there is variation; this would introduce too much nonspecific binding and also decrease the concentration of a matching sequence such that amplification would be too inefficient.

Degeneracy is achieved by having multiple bases introduced at specific base positions during the manufacture of the oligonucleotides (oligos). Oligos are short individual DNA or RNA sequences that are usually synthetically manufactured and have a wide range of applications in molecular biology. A primer is a type of oligo that is used as a starting point for DNA synthesis. A primer typically refers to a single oligo sequence; however, degenerate primers are composed of multiple oligo sequences. In the case, however, of the initial forward PCR primer, position 3 is an A in two genes, G in eight genes, C in four genes and T in five genes (see table: Design of initial forward primer). Since A is less frequent, only G, C, or T were chosen to be represented in the degenerate primer. During manufacture of the primer, G, C, and T will be randomly incorporated into each individual oligo at position 3, resulting in a pool of oligos with three different sequences, each differing at position 3.

Design of initial forward primer.

Plant	Gene	GenBank Accession Number	Sequence
<i>Arabidopsis</i>	<i>GAPC1</i>	AT3G04120	GACTACGTTGTTGAGTCTACTGG
<i>Arabidopsis</i>	<i>GAPC2</i>	AT1G13440	GAC TTTGTTGTTGAGTCTACTGG
<i>Arabidopsis</i>	<i>GAPCP1</i>	AT1G79530	GATTATGTTGTTGAGTCTTCCGG
<i>Arabidopsis</i>	<i>GAPCP2</i>	AT1G16300	GAGTATGTTGTTGAGTCTTCAGG
Pepper	<i>GAPCP</i>	CAN272042	GATTATGTTGTTGAATCTTCTGG
Liverwort	<i>GAPC</i>	AJ246023	GAGTACGTCGTCGAGTCTACCGG
Corn	<i>GAPC1</i>	ZMGPC1	GAGTACGTCGTGGAGTCCACCGG
Corn	<i>GAPC2</i>	U45855	GAGTATGTCGTGGAGTCCACCGG
Corn	<i>GAPC3</i>	U45856	GAATATGTTGTTGAGTCTACTGG
Corn	<i>GAPC4</i>	X73152	GAATATGTTGTTGAGTCTACTGG
Pea	<i>GAPC1</i>	L07500	GATATCATTGTTGAGTCTACTGG
Wheat	<i>GAPC</i>	EF592180	GAGTACGTTGTTGAGTCCACCGG
Rye grass	<i>GAPC3</i>	EF463063	GACTACGTTGTTGAGTCCACTGG
Tobacco	<i>GAPC</i>	AJ133422	GATTACATTGTGGAGTCGACTGG
Tobacco	<i>GAPDH*</i>	DQ682459	GATTTTCGTTGTGGAATCCACTGG
Carrot	<i>GAPDH*</i>	AY491512	GAGTACATTGTGGAGTCCACTGG
Blue gem	<i>GAPDH*</i>	X78307	GAGTACGTCGTTGAGTCGACTGG
Tomato	<i>GAPDH*</i>	AB110609	GACTTCGTTGTTGAATCAACCGG
Snapdragon	<i>GAPDH*</i>	X59517	GAGTATATTGTGGAGTCCACTGG
Initial forward primer			GABTATGTTGTTGARTCTTCWGG
Positions of base			123456789

* Specific *GAPDH* gene not listed in GenBank

The IUB code designates what degenerate bases are incorporated (see table: IUB codes) and the code for incorporation of G or C or T is represented by the letter B. In the initial forward primer there are two more degenerate bases: position 15 is R, as all the bases at that position are purines (G or A) and position 21 is W (A or T). By adding additional degenerate bases, the number of DNA sequences in the pool of oligos also increases, thus the initial forward primer is actually composed of 12 different oligo sequences (3 x 2 x 2). The concentration of the one oligo that is the best match is therefore reduced by one 12th, which can reduce PCR efficiency.

GAGTATGTTGTTGA (GA) TCTTC (AT) GG
 GATTATGTTGTTGA (GA) TCTTC (AT) GG
 GACTATGTTGTTGA (GA) TCTTC (AT) GG

3 bases for position 3

GA (GTC) TATGTTGTTGAGTCTTC (AT) GG
 GA (GTC) TATGTTGTTGAATCTTC (AT) GG

2 bases for position 15

GA (GTC) TATGTTGTTGA (GA) TCTTCAGG
 GA (GTC) TATGTTGTTGA (GA) TCTTCTGG

2 bases for position 21

$$3 \times 2 \times 2 = 12 \text{ oligos}$$

Degenerate primers are beneficial because they increase binding to the target region during PCR, but they have drawbacks. First, they decrease the concentration of specific oligos available for amplification, which reduces the efficiency of the PCR. Second, they increase the chance of amplification of nonspecific/unwanted PCR products. This nonspecific binding can be partially ameliorated by increasing the annealing temperature for the primers, which will discourage nonspecific annealing. However, the annealing temperature cannot be too high because it is unlikely that any of the oligos match the target sequence 100%. (Consider that although degenerate base pairs have been introduced in three highly variable locations, there are still nine bases where the consensus sequence may not match the target sequence of the specific plant gene being studied.) In the example below, the snapdragon sequence matches a degenerate oligo in all three bases (red, italics), but still has four mismatched bases (green, bold).

Initial forward primer: GA***G***TATGTTGTTGA***G***TCTTC***T***GG
 Snapdragon *GAPDH* sequence: GA***G***TAT***A***TTGT***G***GA***G***TC***C***A***C******T***GG

Designing Degenerate Primers from a Protein Sequence

In some cases, a researcher may purify a protein of interest and obtain some amino acid sequence data from the protein but not have any of the DNA sequence for the protein. When that happens, there is another approach for designing primers. Since organisms use more than one codon of three nucleotides to specify some amino acids (see table: Amino acids and the DNA codons for each), primer mixtures can be synthesized that include all possible codons for each amino acid. Although it seems as though there would be huge numbers of oligonucleotides needed, the task can be simplified in several ways, such as choosing an area of protein sequence that is heavy in amino acids that are encoded by only one or two codons.

Amino acids and the DNA codons for each.

Ala (A)	Arg (R)	Asp (D)	Asn (N)	Cys (C)	Gln (Q)	Glu (E)	Gly (G)	His (H)	Ile (I)
GCA	CGA	GAC	AAC	TGC	CAA	GAA	GGA	CAC	ATA
GCC	CGC	GAT	AAT	TGT	CAG	GAG	GGC	CAT	ATC
GCG	CGG						GGG		ATT
GCT	CGT						GGT		
	AGA								
	AGG								
Leu (L)	Lys (K)	Met (M)	Phe (F)	Pro (P)	Ser (S)	Thr (T)	Trp (W)	Tyr (Y)	Val (V)
CTA	AAA	ATG	TTC	CCA	TCA	ACA	TGG	TAC	GTA
CTC	AAG		TTT	CCC	TCC	ACC		TAT	GTC
CTG				CCG	TCG	ACG			
GTG									
CTT				CCT	TCT	ACT			GTT
TTA					AGC				
TTG					AGT				

Note: When there are multiple codons for an amino acid, the codons are very similar. For all amino acids with up to four codons, only the third base differs between codons (for example, the four codons for valine, which all begin with GT). There are three codons that code for a stop signal: TAG, TGA, TAA.

Cloning and Sequencing Explorer Series

By choosing amino acids with fewer codons, the number of degenerate primers can be minimized. For example, to make degenerate primers to the DNA that codes for the amino acid sequence Gly-Leu-Ser-Val, the mixture would include 576 different oligonucleotides:

Amino Acid	Number of Codons	Number of Degenerate Primers
Glycine (Gly)	4	$4 * 6 * 6 * 4 = 576$
Leucine (Leu)	6	
Serine (Ser)	6	
Valine (Val)	4	

In comparison, degenerate primers to the amino acid sequence Asp-Trp-Cys-Glu would include only eight different oligonucleotides:

Amino Acid	Number of Codons	Number of Degenerate Primers
Aspartic acid (Asp)	2	$2 * 1 * 2 * 2 = 8$
Tryptophan (Trp)	1	
Cysteine (Cys)	2	
Glutamic acid (Glu)	2	

These sequence examples are only four amino acids in length, making the primers only twelve oligonucleotides long. Degenerate primers are usually longer, meaning more oligonucleotide combinations will be needed. To keep the number needed to a minimum, choose a target amino acid sequence containing amino acids coded by only one or two codons and try to avoid amino acids that have six codons.

Nested PCR

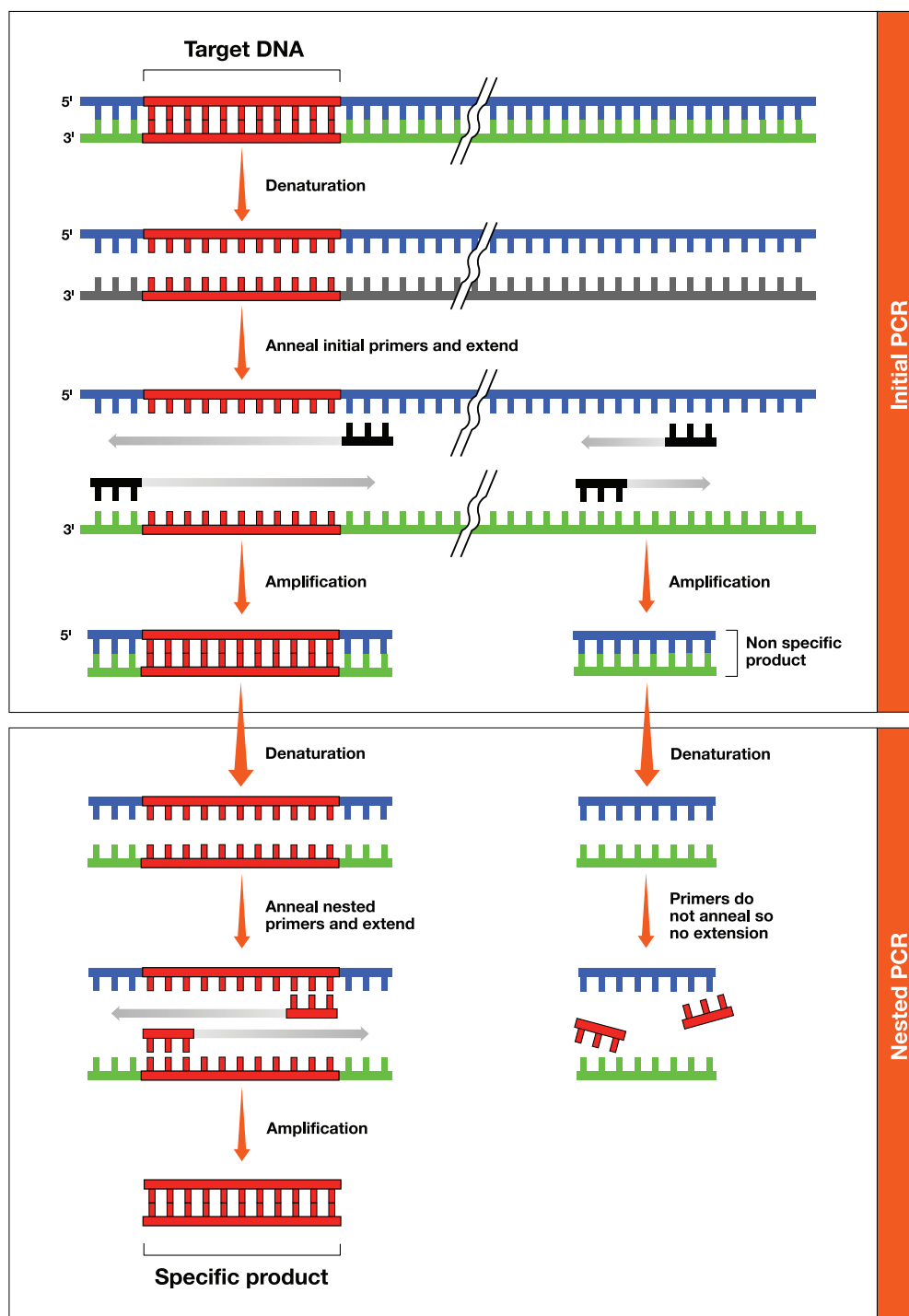
A number of variations of PCR have been developed in the last 20 years to address specific research questions. Some of these variations include inverse PCR, in situ PCR, long PCR, real-time PCR, and nested PCR. When there is the potential for primers to bind to sequences of the template DNA other than at the target area (for example, when using degenerate primers), nested PCR can increase the yield and specificity of amplification of the target DNA. Nested PCR uses two sequential sets of primers. The first primer set binds to sequences outside the target DNA, as expected in standard PCR, but it may also bind to other areas of the template. The second primer set binds to sequences in the target DNA that are within the portion amplified by the first set (that is, the primers are nested). Thus, the second set of primers will bind and amplify target DNA within the products of the first reaction. One advantage of nested PCR is that if the first primers bind to and amplify an unwanted DNA sequence, it is very unlikely that the second set of primers will also bind within the unwanted region.

A second advantage of nested PCR is that the initial PCR step enriches the pool of potential targets for the second set of nested primers. The initial PCR using degenerate primers, which is less efficient than a PCR using homologous primers, results in a lower concentration of PCR product than is desirable for ligation and also amplifies undesirable nonspecific PCR products. The low efficiency is due to the following reasons (see Designing Degenerate Primers from Consensus DNA Sequences, above, for more explanation):

- The concentration of primers that bind efficiently in the reaction is lower than normal due to the presence of multiple oligo sequences for each primer
- The annealing temperature used for the PCR is quite high (52°C) to discourage primers from binding nonspecifically
- Even though degenerate primers are used, the oligo sequences are still not 100% homologous to the target sequence, thus reducing annealing efficiency

However, by performing the initial PCR reaction, the pool of targets for a second, nested round of PCR is greatly enriched. During the initial PCR, there are only a few target regions within the millions of base pairs of DNA in a genomic DNA sample, while after the initial PCR has completed, the number of target regions has increased by a millionfold or more. This enrichment greatly increases the efficiency of the nested PCR reaction.

The second round of nested PCR does not use degenerate primers, which increases its specificity for *GAPC* and *GAPC-2* genes over other *GAPDH* family members. However, to gain this specificity, PCR efficiency is sacrificed, since the consensus sequences of the invariable nested primers are even less complementary to the plant target sequence than the initial degenerate primers. If the nested primers are used directly on genomic DNA, they usually amplify poorly because they do not bind well to the template DNA. However, because of the enrichment of target DNA by the initial round of PCR, there are many more targets for the primers to anneal, increasing the chances of binding, and boosting the PCR efficiency. In addition, to encourage the noncomplementary primers to anneal, the annealing temperature for the nested PCR is also reduced to 46°C. In contrast to the initial round of PCR where nonspecific binding was discouraged through use of a higher annealing temperature, here as much binding as possible is encouraged, since most nonspecific binding sites were screened out during the initial PCR.



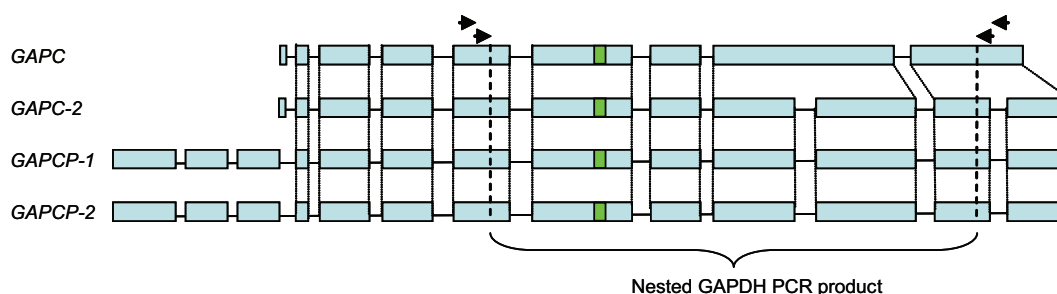
Nested PCR. Nested PCR involves two rounds of PCR with the product of the first round acting as template for the second round.

The PCR variation of nested PCR will be used to amplify a portion of *GAPC*, the plant gene for NAD⁺-dependent cytosolic GAPDH. The section of the gene to be amplified encodes around two-thirds of the protein, including the active site of the enzyme. The primer annealing sites within the *Arabidopsis GAPC* gene are shown below.

GACTACGTTGTTGAGT**CTACTGGTGTCTTCACTGACAA**AGACAAGGCTGCAGCTCACTTGAAGGTTTGTCTTATTTGAATTGGTTATTTTTGTCTTGTAAATGATATAAAATAGTTTATGTGCTAGAATTTGCTTAGTATCATTCAACTAAATTTGTGACTTGTGTATTTTCAGGGTGGTGCCAAGAAGGTTGTTATCTCTGCCCCCAGCAAAGACGCTCCAATGTTTGTGTGGTGTCAACGAGCAGCAATACAAGTCCGACCTTGACATTGTCTCCAACGCTAGCTGCACCACTAACTGCCTTGCTCCCCTTGCCAAGGTAAAATATCTGATATTCTATATGATCAAATTTGACTTTGTATTTCAAGTTGAAGTGACTAATTTTCATTTAACGTTCTTTGATTTCAATTGTGTAGGTTATCAATGACAGATTTGGAATTGTTGAGGGTCTTATGACTACAGTCCACTCAATCACTGGTAAATTTATCAATCAGTTAGAAAGTTATTACAACTTGCTTGCCTATAGGTGGAAAATTTGTGATTTAATGGGGTTTGCTTTATGATTTCAGCTACTCAGAAGACTGTTGATGGGCCTTCAATGAAGGACTGGAGAGGTGGAAGAGCTGCTTCATTCAACATATATCCCAGCAGCACTGGAGCTGCCAAGGCTGTCGGAAGGTGCTTCCAGCTCTTAACGGAAGTTGACTGGAATGTCTTTCCGTGTCCCAACCGTTGATGTCTCAGTTGTTGACCTTACTGTCAGACTCGAGAAAGCTGCTACCTACGATGAAATCAAAAAGGCTATCAAGTAAGCTTTTGAGCAATGACAGATTAAGTTTACTTATATTCCAGTAGTGATCAAATTACTCACCAAGTGTTTTTACCACCAATACATAGGGAGGAATCCGAAGGCAAACTCAAGGGAATCCTTGGATACACCGAGGATGATGTTGTCTCAACTGACTTCGTTGGCGACAACAGGTCG**AGCATTT****TTGACGCCAAGGCTG**GAATTGCATTGAGCGACAAGTTTGTGAAATTGGTGTC**TGGTACGACAACGAATGG**

GAPDH PCR primers. Positions of first-round initial *GAPDH* PCR primers (blue, bold) and second-round nested PCR primers (yellow, underlined italics) on *Arabidopsis GAPC* gDNA are indicated. **Note:** Reverse primers are complementary to the antisense DNA strand, the one that does not encode the gene.

While the GAPDH protein sequence is highly homologous among family members and species, the gene structure (the sequence that does not actually code for amino acids in the final protein), including the number, locations, sequence, and length of introns, is more variable. This can be observed in the differences in gene structure within the *Arabidopsis GAPC* gene family, where *GAPC* is missing two introns present in the other family members, resulting in a shorter PCR product. This variability in gene structure results in PCR products of different lengths that can be identified by agarose gel electrophoresis. Studies of other plant species during the development of this lab series identified numerous other instances of absent introns.



Gene structure of the *Arabidopsis GAPC* family of genes. Blue bars indicate coding sequence (exons). *GAPC* differs from the rest of the family by the absence of two introns (noncoding sequence, indicated by lines), which shortens the gene. The *GAPCP* subfamily of genes has a signal peptide at the N-terminus that directs the protein to plastids. Arrows indicate annealing positions of first-round (outer arrows) and second-round nested (inner arrows) *GAPDH* PCR primers on the structure of *GAPC* family genes. Green bar indicates location encoding the enzyme active site. **Note:** Figure is not to scale.

Since the first-round primers used in this lab are degenerate and were designed based on a consensus sequence derived from a number of *GAPC* genes (including those encoding isozymes such as *GAPC* and *GAPCP*), they may anneal to the target DNA at several locations. These locations may be sequences of *GAPDH* genes other than *GAPC*, or they may be unrelated sequences that have a high degree of complementarity to one or more of the degenerate primers. So it is likely that multiple bands of amplified DNA may be seen on an agarose gel after the initial round of PCR. The nested primers were designed to be more specific to *GAPC* (rather than the *GAPCP* subfamily) and are not degenerate, so in theory only the *GAPC* genes from the pool of *GAPDH* genes amplified during the initial PCR will be amplified in the second round of nested PCR. For example, if the plant gDNA used in this lab is from *Arabidopsis*, then the nested PCR should amplify only *GAPC* and *GAPC-2*. The nested primers should not bind to DNA coding for *GAPDH* isozymes or to unrelated DNA sequences, so that DNA should not be amplified.

Analyzing Results

PCR products can be visualized by agarose gel electrophoresis, with the agarose concentration determined by the expected size of the products. In addition to the experimental samples, the negative control reaction and a size marker should be included on the gel. The size markers help determine if the size of the PCR product is as expected. The negative control reaction should not yield any amplified DNA. If it does, then the reactions may have been contaminated and the experimental results are suspect. If the PCR products are around 50–100 base pairs, the reactions may have formed primer-dimers.

However, occasionally *GAPCP* genes are cloned using these primers. The sizes of PCR products expected from *Arabidopsis* *GAPC* family genes using the primers in this lab are shown in the table below:

Expected length of *Arabidopsis* *GAPC* gene family PCR products.

Enzyme Function	<i>GAPDH</i> Protein Subunit	<i>Arabidopsis</i> Gene	<i>Arabidopsis</i> Chromosome Location	Length of PCR Product (bp)	
				Initial Primers	Nested Primers
NAD ⁺ -dependent GAPDH in cytosol	<i>GAPC</i>	<i>GAPC</i>	3	1,065	993
	<i>GAPC-2</i>	<i>GAPC-2</i>	1	1,216	1,145
NAD ⁺ -dependent GAPDH in plastids	<i>GAPCP</i>	<i>GAPCP</i>	1	1,303	1,231
	<i>GAPCP-2</i>	<i>GAPCP-2</i>	1	1,205	1,133

Note: The pGAP plasmid, which is used as a PCR control, contains the sequence for the first-round PCR product of the *Arabidopsis* *GAPC* gene.

What does it mean if no DNA is visible on an agarose gel after the nested PCR? If the experimental controls worked (meaning that the problem was not with the reagents or the thermal cycler), then it is likely that no *GAPC* was amplified from the gDNA sample. The most probable reason is that the initial primers did not bind to any target DNA because there was too little complementarity between the primers and the target. Alternatively, there could have been too little gDNA or PCR inhibitors present in the gDNA preparation.

Instructor's Advance Preparation

In this *GAPDH* PCR chapter, students will perform two rounds of PCR to amplify a portion of the *GAPC* gene from their gDNA. The initial round of PCR uses degenerate primers to amplify a pool of DNA fragments. This means that one or more positions in the primer nucleotide sequence have had more than one nucleotide synthesized at that position. Whereas primers are usually synthesized with one nucleotide at each position, degenerate primers have had different nucleotides synthesized for one or more positions in the sequence so that these primers are actually a mixture of oligonucleotides of closely related sequence. This idea can be demonstrated in the example below, where an N stands for any deoxynucleotide.

Degenerate primer sequence: 5'-AGCTTTGC**N**TGTGAAC-3'

The primer reagent is actually composed of four different oligonucleotides, each with a different base at the N position

5'-AGCTTTGC**A**TGTGAAC-3'

5'-AGCTTTGC**T**TGTGAAC-3'

5'-AGCTTTGC**G**TGTGAAC-3'

5'-AGCTTTGC**C**TGTGAAC-3'

For this lab, the initial primer nucleotide sequence is degenerate at three positions (3, 15, and 21). The initial forward primer is:

GAB**T**ATGTTGTTGAB**R**TCTTC**W**GG

Degeneracy in initial primer sequence.

Oligonucleotide Position	IUB Code	Bases
3	B	G/T/C
15	R (purine)	G/A
21	W (weak)	A/T

Thus the initial primer reagent contains 12 different oligonucleotides, increasing the probability that one of the oligonucleotides will be complementary to the template DNA. The initial reverse primer is prepared using the same approach.

The second round of nested PCR uses more specific primers to amplify a specific *GAPC* gene from the pool of amplified PCR products. The nested primers are not degenerate. They are intended to bind internally within the target sequence amplified by the initial primers. The primers will not bind equally well to DNA sequences from all plant species, so different plants will result in different reaction efficiencies in PCR. Students may well obtain a clonable fragment after the first round of PCR. In this case, clonable means that a PCR product of sufficient quantity (generally, visible on an agarose gel is sufficient), quality (meaning few or single bands), and the expected size.

GAPDH PCR Checklist

Components from Cloning and Sequencing

Explorer Series	Where Provided	(✓)
PCR master mix (2x)	GAPDH PCR Module	<input type="checkbox"/>
Initial GAPDH PCR primers	GAPDH PCR Module	<input type="checkbox"/>
Nested GAPDH PCR primers	GAPDH PCR Module	<input type="checkbox"/>
5x Control <i>Arabidopsis</i> gDNA, 25 ng/μl	GAPDH PCR Module	<input type="checkbox"/>
pGAP control plasmid DNA	GAPDH PCR Module	<input type="checkbox"/>
Exonuclease	GAPDH PCR Module	<input type="checkbox"/>
UView 6x loading dye and stain, 1 ml	GAPDH PCR Module	<input type="checkbox"/>
Sterile water	GAPDH PCR Module	<input type="checkbox"/>
PCR tubes, 0.2 ml	GAPDH PCR Module	<input type="checkbox"/>
Capless PCR tube adaptors, 1.5 ml	GAPDH PCR Module	<input type="checkbox"/>
Microcentrifuge tubes, multicolor, 2 ml	GAPDH PCR Module	<input type="checkbox"/>

Required Accessories (Not Provided)	Quantity	(✓)
Thermal cycler	1	<input type="checkbox"/>
Water bath, heating block, or incubator set to 37°C*	1	<input type="checkbox"/>
Water bath, heating block, or incubator set to 80°C*	1	<input type="checkbox"/>
20 μl adjustable-volume micropipets and aerosol barrier filter tips	12	<input type="checkbox"/>
200 μl adjustable-volume micropipet and aerosol barrier filter tips	12	<input type="checkbox"/>
Tube racks	12	<input type="checkbox"/>
Tubes for aliquoting (optional)	60	<input type="checkbox"/>
Microcentrifuge (optional)	1	<input type="checkbox"/>
Marking pen	12	<input type="checkbox"/>

* **Note:** A thermal cycler can also be used.

Note: PCR is exceptionally sensitive to contamination by DNA from many sources. All reagents to be used for PCR should be handled with care to minimize contamination. It is recommended either to set up PCR in an area of the lab or classroom that is separate from the DNA extraction area or to thoroughly swab down the lab benches used with a commercial cleaner or 10% bleach (ethanol does not destroy DNA). In addition, filter pipet tips should always be used to set up PCR. The micropipets should be carefully cleaned with a 10% solution of bleach before performing PCR. In research labs, PCR hoods are frequently used to prevent contamination. Refer to Appendix A3: Sterile techniques for PCR for further information.

Tasks to Perform Prior to Lab

1. Dilute the required amount of 5x *Arabidopsis* gDNA (25 ng/μl) in sterile water to 5 ng/μl. For 12 student teams, combine 20 μl of gDNA with 80 μl sterile water and mix well. Each student team requires 6 μl. Store at –20°C when not in use.
2. Optional: Prepare PCR master mix with primers. Depending on the level of your students, you may wish to prepare this mix ahead of time. The student protocol includes an option with molarity calculations for students to prepare their own master mix with primers. The preparation method is the same for each different primer set. However, add primers to master mix a maximum of 30 minutes prior to use. Each student team requires 120 μl of 2x master mix with primers. For 12 student teams, 30 minutes prior to use, add 30 μl of specific primers (blue initial primers for the first round of PCR and yellow nested primers for the second round) to 1,500 μl of master mix for PCR (2x) and mix thoroughly. Store on ice.
3. For the second round of PCR, place the exonuclease I enzyme on ice.
4. Program thermal cycler.

For the first round of PCR, program the thermal cycler with the Initial *GAPDH* PCR program:

Initial denaturation 95°C for 5 min

Then 40 cycles of:

 Denaturation 95°C for 1 min

 Annealing 52°C for 1 min

 Extension 72°C for 2 min

Final extension 72°C for 6 min

Hold 15°C hold (∞)

For the second round of PCR, program the thermal cycler with the Nested *GAPDH* PCR program:

Initial denaturation 95°C for 5 min

Then 40 cycles of:

 Denaturation 95°C for 1 min

 Annealing 46°C for 1 min

 Extension 72°C for 2 min

Final extension 72°C for 6 min

Hold 15°C hold (∞)

Protocol

Overview

The strategy for this experiment uses nested PCR to amplify portions of the *GAPC* gene from gDNA of the plant of interest. Three control reactions will also be set up. A negative control will be run with sterile water instead of gDNA to test for contamination, and two positive controls will be run with a plasmid and *Arabidopsis* gDNA. The plasmid control ensures the PCR reagents and thermal cycler are functioning. The pGAP control plasmid contains the *GAPC* target region from *Arabidopsis* and should yield an intense single band around 1 kb. The second positive control with *Arabidopsis* gDNA more

closely matches your extracted plant gDNA. The gDNA control provides a more representative PCR result since amplification of genomic DNA is much less efficient than plasmid DNA. The band intensity for this reaction will be lower than for the plasmid and may also yield multiple bands representative of the four *GAPC* genes in *Arabidopsis*. The gDNA control also provides a control template for the nested PCR, and products from this control nested PCR may be used to continue the lab in the event that the plant samples did not amplify. In the initial round of PCR, a set of blue primers using degenerate (less specific) sequences will amplify the *GAPC* gene from the gDNA. Then, in the second round of PCR (the nested PCR), a more specific set of yellow primers will amplify *GAPC* from the initial PCR products. It is very important not to reverse the order in which the primers are used or to mix the two primer sets together in the PCR reactions.

Before performing the nested PCR, the primers that were not incorporated into PCR product in the first round must be removed so that they do not amplify target DNA in this round of PCR. To remove the primers, exonuclease I, an enzyme that specifically digests single-stranded DNA such as primers but not double-stranded DNA such as the template, will be added to the PCR products from the first round. After exonuclease I digests the primers, the enzyme must be inactivated to prevent the exonuclease from digesting the nested primers that will be added for the next round of PCR.

Following exonuclease I treatment and inactivation of the enzyme, the PCR products from the gDNA templates generated in the first round of PCR need to be diluted. The PCR products from the initial round contain a high proportion of *GAPC*-like sequences relative to the total amount of gDNA. By diluting the gDNA, it is even less likely that the gDNA will be a template for contaminating PCR products when using the nested primers in the next round of PCR. Because the complexity of the pool of available DNA templates has been greatly reduced, nested PCR is very efficient.

As each PCR reaction takes approximately 3–4 hours to run, it is most practical to run the PCR reactions on separate days. Since the reagents used in these experiments function optimally when prepared fresh, it is highly recommended that the reagents be prepared just prior to setting up the reactions.

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- **Amplify region of *GAPC* gene using PCR**
- Assess the results of PCR
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items to set up 5 of the **initial** PCR reactions:

Material Needed for Each Workstation	Quantity
PCR master mix (2x)	120 µl
Initial <i>GAPDH</i> PCR primers, blue	4 µl
gDNA — previously extracted	from 2 plants
<i>Arabidopsis</i> gDNA (diluted to 5 ng/µl)	6 µl
pGAP control plasmid DNA	6 µl
Sterile water	100 µl
20 µl adjustable-volume micropipet and filter tips	1
0.2 ml PCR tubes	5
Capless PCR tube adaptors	5
Microcentrifuge tube	1
Marking pen	1
Ice bath	1
Tube rack	1

Each student team will require the following items to set up 5 of the **nested** PCR reactions:

Material Needed for Each Workstation	Quantity
Exonuclease I on ice	3 µl
PCR master mix (2x)	120 µl
Nested <i>GAPDH</i> PCR primers, yellow	4 µl
PCR reactions from initial PCR	3
pGAP control plasmid DNA	25 µl
Sterile water	350 µl
20 µl adjustable-volume micropipet and filter tips	1
200 µl adjustable-volume micropipet and filter tips	1
0.2 ml PCR tubes	5
Capless PCR tube adaptors	5
Microcentrifuge tubes	4
Marking pen	1
Ice bath	1
Tube rack	1

Common Workstation

Material Required	Quantity
Water bath, incubator, or heating block set to 37°C	1
Water bath, incubator, or heating block set to 80°C	1

Note: A thermal cycler can also be used.

Preparation for Initial PCR (First-Round PCR)

1. Plan the first round of PCR. You will perform one initial PCR for each of the two plant gDNA samples you have extracted, two positive controls, one using control gDNA and the other using pGAP plasmid DNA, and one negative control with sterile water instead of DNA, for a total of 5 PCR reactions. Use the table below to record the label on each PCR tube, the DNA template, and the primers used to amplify the DNA.
2. Prepare a master mix. (**Note:** This mix should be prepared no more than 30 minutes prior to performing PCR. Your instructor may have prepared it just prior to the lab.) A master mix is a mixture of all the reagents required for PCR except the template DNA. Making a single mixture reduces potential pipetting errors, and increases the consistency between PCR samples.

Tube Label	Template	Primers

Bio-Rad's 2x master mix is provided as a 2x (double strength) colorless reagent; when mixed with an equal amount of DNA template, all components are at optimal concentrations in the final reaction. Bio-Rad 2x master mix contains *Taq* DNA polymerase, dNTPs, buffer, and salt, but does not contain primers. It is necessary to add primers to Bio-Rad 2x master mix to form a complete 2x master mix with initial primers — "2x MMIP".

Consider why commercial master mixes are not sold with primers added to them. Give one reason. _____

Each PCR reaction volume is 40 μ l. Thus, each reaction requires 20 μ l of 2x MMIP (Bio-Rad 2x master mix with blue initial primers) plus 20 μ l of DNA template. The amount of 2x MMIP to prepare is calculated by multiplying the number of reactions plus one (to allow for pipetting errors) by the volume of 2x MMIP required for each reaction (20 μ l).

Referring to your table, calculate how much 2x MMIP is required.

Volume of 2x MMIP required = (# PCR reactions + 1) x 20 μ l = _____ μ l

For the initial round of PCR, blue initial primers will be used. These primers are designed for a section of DNA bracketing the target sequence. The initial primers are supplied at a 100 μ M concentration. For this PCR, the concentration of primers in the 2x MMIP should be 2 μ M. Calculate the volume of initial primers required in the 2x MMIP. Remember the formula $M_1V_1 = M_2V_2$, where:

M_1 = Required concentration of primers (μ M)

V_1 = Required volume of 2x MMIP (μ l)

M_2 = Given concentration of primers (μ M))

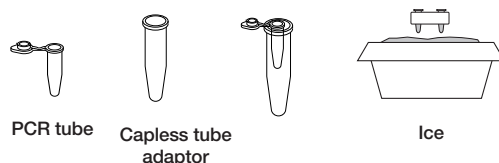
V_2 = Required volume of primers (μ l)

Required volume of initial primers (V_2): _____ μ l

Label a microcentrifuge tube 2x MMIP. No more than 30 min before use, add the calculated volume of initial primers to the required volume of Bio-Rad 2x master mix in a labeled tube. Mix well by pipetting up and down several times or vortexing. If a microcentrifuge is available, spin tube briefly to collect the contents at the bottom of the tube. Keep on ice.

Experimental Procedure for Initial PCR

1. Referring to your table, label your PCR tubes with your initials and the tube label.
2. Place each PCR tube into a tube adaptor and cap each tube. Place the adaptor tube with PCR tube on ice.

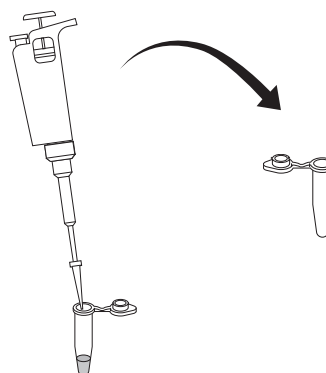


3. Ensure all the reagents are thoroughly mixed, especially the gDNA. Mix tubes containing reagents thoroughly by vortexing or flicking to ensure the gDNA is homogeneously distributed. Before opening the tubes, spin in a microcentrifuge for 5–10 seconds to force contents to the bottom of the tube (to prevent contamination).

Each PCR needs to be set up with the following reagents:

4. Pipet 20 μ l of 2x MMIP into each PCR tube.
5. Add 15 μ l of sterile water to each tube.
6. Referring to your initial PCR plan, use a fresh pipet tip to add 5 μ l of the appropriate DNA template to each tube and gently pipet up and down to mix reagents. Use a fresh filter tip each time. Recap tubes tightly to prevent any evaporation during PCR.

Reagent	Amount
Blue master mix (2x MMIP)	20 μ l
Sterile water	15 μ l
DNA template or negative control	5 μ l
Total	40 μl



- When your instructor tells you to do so, place your PCR tubes into the thermal cycler.

The PCR reaction will run for the next several hours using the following Initial *GAPDH* PCR program:

Initial denaturation: 95°C for 5 min

Then 40 cycles of:

Denaturation: 95°C for 1 min

Annealing: 52°C for 1 min

Extension: 72°C for 2 min

Final extension: 72°C for 6 min

Hold: 15°C hold (∞)

- Store PCR products at 4°C for up to 2 weeks and at -20°C long term.

Note: Store any remaining gDNA at -20°C.

(Optional) Analyze PCR products by agarose gel electrophoresis. This can be performed on the same gel used to analyze the nested PCR products (Chapter 3, Electrophoresis).

Prepare agarose gels to analyze the results of your experiment. You will need 1% agarose gels with the sufficient number of wells for each of your samples plus an additional well for your molecular size marker. See Appendix A for detailed instructions on preparing agarose gels.

Note: Do not add loading dye directly to initial PCR reactions — the loading dye can inhibit the next round of PCR. Refer to Chapter 3, Electrophoresis for instructions on using electrophoresis to analyze PCR products.

The plasmid control should yield a visible band of around 1 kb. It is relatively common for plant gDNA to not yield a visible band during the initial round of PCR and yet be amplified after the nested round of PCR.

Results Analysis for Initial PCR

Following electrophoresis, consider the following questions regarding your controls and samples:

- Did the negative control generate a PCR product? If yes, what size were the DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?
- Did the pGAP plasmid generate a PCR product? If yes, what size is the DNA band and what does this mean for the experiment? If no, what does this mean for the experiment?

- Did the control gDNA generate a PCR product? If yes, what size DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?

- Did your plant gDNA generate PCR products? If yes, what size DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?

Plant 1:

Plant 2:

Preparation for Nested PCR (Second-Round PCR)

- Plan the nested PCR experiment. Three nested PCRs will be set up, using as a template the PCR products of each gDNA sample amplified in the initial round of PCR: the PCR product from the control *Arabidopsis* gDNA and the two PCR products from the newly extracted plant gDNAs. In addition, two control reactions will be set up with the nested PCR primers: pGAP plasmid as a positive control and sterile water as a negative control. Complete the table below to record the label for each nested PCR tube, the DNA template, the dilution factor, and the primers used.

Tube Label	Template	Dilution Factor	Primers

2. Prepare a master mix. (**Note:** Your instructor may have done this just prior to the lab.) Each nested PCR reaction volume is 40 μl . Thus, each nested PCR requires 20 μl of 2x MMNP composed of 2x master mix and yellow nested primers. Referring to your table, and calculations from the initial round of PCR if necessary, calculate how much 2x MMNP is required.

Volume of 2x MMNP required = (# nested PCR reactions + 1) \times 20 μl = _____ μl

The nested PCR uses nested primers that specifically target DNA interior to the locations where the primers for the initial PCR bound. The yellow nested primers are supplied at a 25 μM concentration. For this experiment, the concentration of primers in the 2x MMNP should be 0.5 μM . Calculate the volume of nested primers required in the 2x MMNP. Remember the formula $M_1V_1 = M_2V_2$ where:

M_1 = Required concentration of primers (μM)

V_1 = Required volume of 2x MMNP (μl)

M_2 = Given concentration of primers (μM)

V_2 = Required volume of primers (μl)

Required volume of nested primers (V_2): _____ μl

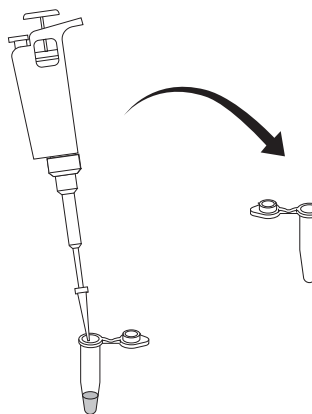
Label a microcentrifuge tube 2x MMNP. Not more than 30 minutes before use, add the calculated volume of initial primers to the required volume of 2x master mix in a labeled tube. Mix well by pipetting up and down several times or vortexing. If a microcentrifuge is available, spin tube briefly to collect the contents at the bottom of the tube. Keep on ice.

3. Obtain tubes from initial round of PCR. Use only the PCR tubes that contained the gDNA, including the positive control gDNA but not the plasmid DNA or the negative control reactions — refer to your table for the first round of PCR.

Experimental Procedure for Nested PCR

1. Prepare template DNA. Treat with exonuclease I to remove unincorporated primers from initial PCR tubes. Using a fresh tip each time, pipet 1 μl of exonuclease I into each initial PCR sample of amplified gDNA. Mix well by pipetting up and down.
2. Incubate at 37°C for 15 min.
3. Incubate at 80°C for 15 min to heat-inactivate the exonuclease I enzyme.

Why is it necessary to inactivate the exonuclease?

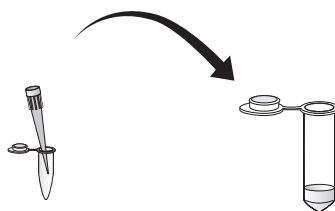


4. Label a microcentrifuge tube for each exonuclease-treated initial PCR tube.

- The initial PCR sample will be diluted 100 times in the second PCR. Rather than pipetting 0.4 μl of the initial PCR sample, the sample is first diluted to allow a larger volume to be pipetted.

To dilute each initial PCR sample to 1/50 the original concentration, pipet 98 μl of sterile water into each of the labeled microcentrifuge tubes. Remember the initial PCR sample will be further diluted when added to 2x master mix.

- Using a fresh tip each time, pipet 2 μl of the appropriate initial PCR into the appropriate microcentrifuge tube. Close the cap.



- Vortex or flick the tube with your finger to mix. Spin briefly in a microcentrifuge to collect the liquid at the bottom of the tube.
- Label PCR tubes according to your plan and place in PCR tube adaptors on ice.

Each PCR needs to be set up with the following reagents:

Reagent	Amount
Yellow master mix (2x MMNP)	20 μl
DNA template or negative control	20 μl
Total	40 μl

- Pipet 20 μl of 2x MMNP into each PCR tube.
- Referring to your nested PCR plan, use a fresh pipet tip to add 20 μl of the appropriate DNA template (diluted gDNA, plasmid DNA, or sterile water for the negative control) to each PCR tube. Gently pipet up and down to mix reagents. Recap tubes.
- When your instructor tells you to do so, place your PCR tubes into the thermal cycler.

The PCR will run for the next several hours using the following Nested *GAPDH* PCR program:

Initial denaturation: 95°C for 5 min

Then 40 cycles of:

Denaturation: 95°C for 1 min

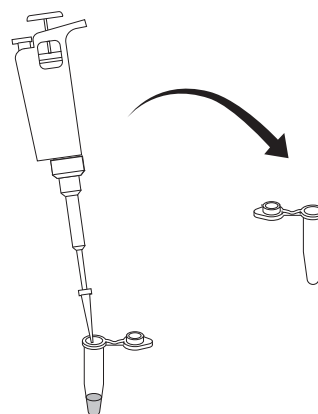
Annealing: 46°C for 1 min

Extension: 72°C for 2 min

Final extension: 72°C for 6 min

Hold: 15°C hold (∞)

- Store PCR products at 4°C for up to 2 weeks and -20°C for long term.



Results Analysis for Nested PCR

Following electrophoresis, consider the following questions regarding your controls and samples:

1. Did the negative control generate a PCR product? If yes, what size DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?
2. Did the pGAP plasmid generate a PCR product? If yes, what size is the DNA band and what does this mean for the experiment? If no, what does this mean for the experiment?
3. Did the control gDNA generate a PCR product? If yes, what size DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?
4. Did your plant gDNA generate PCR products? If yes, what size DNA band(s) and what does this mean for the experiment? If no, what does this mean for the experiment?

Plant 1:

Plant 2:

Next Steps

1. After both rounds of PCR are completed, the products will be analyzed by agarose gel electrophoresis (Chapter 3). If you have not already done so, prepare agarose gels and make electrophoresis running buffer according to standard protocols with the sufficient number of wells for all PCR samples, including one for your molecular size marker (see Appendix A). Refer to Chapter 3, Electrophoresis, to prepare a plan for your electrophoresis experiment to determine the required number of wells.
2. Plan ahead for the Ligation and Transformation chapters (5 and 6).
3. Prepare LB agar and LB amp IPTG agar plates: At least **3 days** prior to the transformation, prepare one LB agar plate per team to be used to inoculate starter cultures. Prepare this according to standard protocols (see Appendix A). Use remaining LB agar to make 2 LB amp IPTG plates per team.
4. Prepare the *E. coli* starter plate: At least **2 days** prior to the transformation, streak an LB agar starter plate with *E. coli* bacteria. First, rehydrate the bacteria in the *E. coli* stock vial with 250 μ l of sterile water. Streak 10 μ l of the rehydrated bacteria on the LB plate using standard microbiology techniques to allow formation of single colonies (see Appendix A). Incubate the plate at 37°C overnight. Once colonies have grown, wrap the plate in Parafilm and store at 4°C for up to 2 weeks.

Note: Any strain of *E. coli* bacteria normally used for transformation, such as DH5a or DH10b may be used in place of HB101.

5. Prepare LB broth: At least **2 days** prior to the transformation, prepare at least 25 ml of LB broth per team according to standard procedures (see Appendix A).

Focus Questions for *GAPDH* PCR

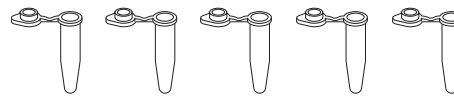
1. What are the steps of DNA replication in the cell?
2. How do researchers mimic these steps in a test tube during PCR?
3. How do researchers target the portion of DNA to be amplified (copied)?
4. What are degenerate primers and why would you use them?
5. Why is it necessary to conduct two rounds of PCR in this lab?

6. What is the purpose of the exonuclease I enzyme?
7. Why was the discovery of *Taq* DNA polymerase important for the development of PCR?
8. The protein GAPDH is necessary for cellular function and is highly conserved between organisms. Why is it probable that proteins needed for cell survival will be very similar (highly conserved) in many different organisms?
9. What is a consensus sequence?
10. What would be the consensus sequence for the following aligned sequences?
ATTGCTTC
AATGCTAC
ATTCCTAC
ATTGCTAC

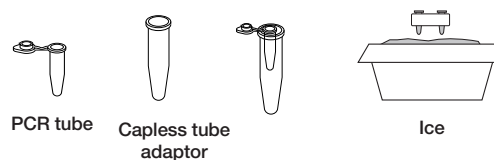
GAPDH PCR — Quick Guide

Part I Initial PCR

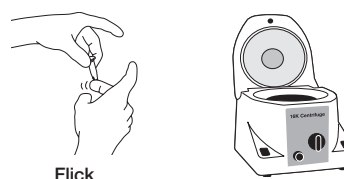
1. Label five PCR tubes with your initials and the following labels:
 1 – Negative control (sterile water)
 2 – *Arabidopsis* gDNA
 3 – Positive control pGAP plasmid
 4 – Plant 1 gDNA
 5 – Plant 2 gDNA



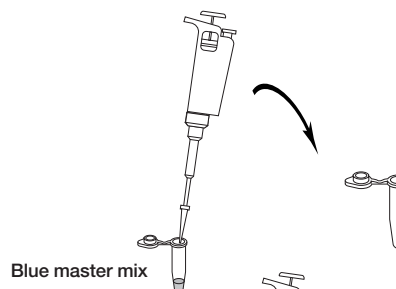
2. Place each PCR tube into a capless tube adaptor and cap each tube. Place the adaptor tube with PCR tube on ice.



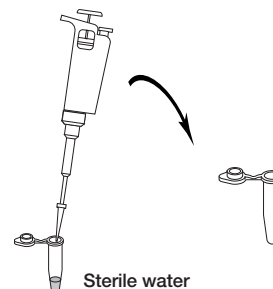
3. Thoroughly mix tubes of 2x blue master mix and DNA samples and then centrifuge 10 sec to force contents to bottom of tubes.



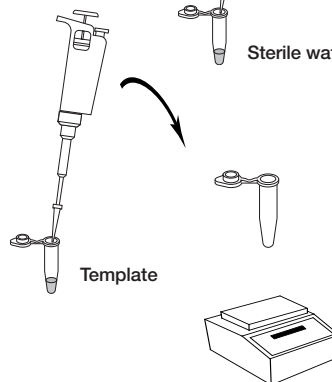
4. Pipet 20 μ l of 2x blue master mix containing initial primers (2x MMIP) into each PCR tube. Change pipet tip for each sample.



5. Using a fresh tip, pipet 15 μ l of sterile water to each tube.



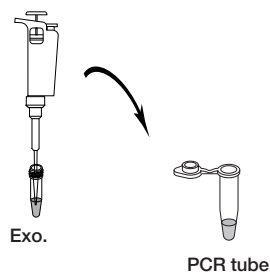
6. Using a fresh tip every time, add 5 μ l of the appropriate DNA template to each tube. Gently pipet up and down to mix reagents. Recap tubes.



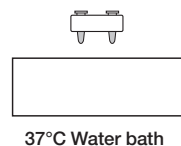
7. When instructed, place PCR tubes in thermal cycler.
8. After PCR is complete, store PCR products at 4°C for up to 2 weeks or at -20°C for long term.

Part II Nested PCR

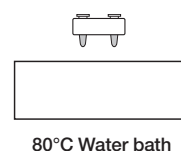
1. Using a fresh tip each time, pipet 1 μ l of exonuclease I into each blue initial PCR sample containing amplified genomic DNA; i.e., tubes 2, 4, and 5 from step 1 of the previous page. Mix well by pipetting up and down.



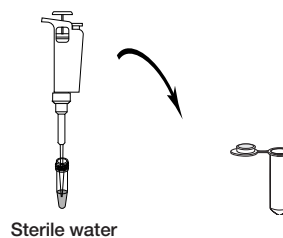
2. Incubate exonuclease reactions at 37°C for 15 min.



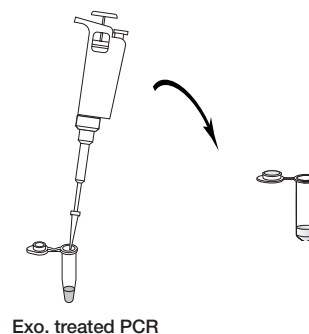
3. Heat inactivate exonuclease by incubating at 80°C for 15 min.



4. Label 3 microcentrifuge tubes for the exonuclease-treated PCR reactions.
 - Exo *Arabidopsis* initial PCR
 - Exo plant 1 initial PCR
 - Exo plant 2 initial PCR



5. Add 98 μ l of sterile water to each microcentrifuge tube.
6. Using a fresh tip each time, add 2 μ l of each exonuclease-treated PCR reaction into the appropriate microcentrifuge tube.



7. Vortex or flick each tube to mix. Spin briefly in a microcentrifuge to force the contents to the bottom of the tube.

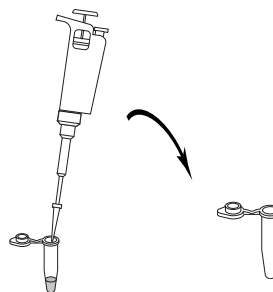


Flick



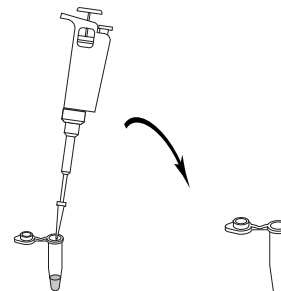
8. Label five PCR tubes with your initials and the following labels.
 6 – Negative control (sterile water)
 7 – Exo *Arabidopsis* gDNA
 8 – Positive control pGAP plasmid
 9 – Exo plant 1
 10 – Exo plant 2

9. Pipet 20 μ l of 2x yellow master mix with nested primers (2x MMNP) into each PCR tube.



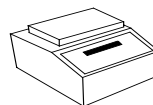
Yellow master mix

10. Using a fresh tip each time, pipet 20 μ l of template (according to your tube labels) into the appropriate PCR tube. Gently pipet up and down to mix reagents. Recap tubes.



Template

11. When instructed, place PCR tubes in thermal cycler.



12. After PCR is complete, store PCR products at 4°C for up to 2 weeks or at –20°C long term.

CHAPTER 3: ELECTROPHORESIS

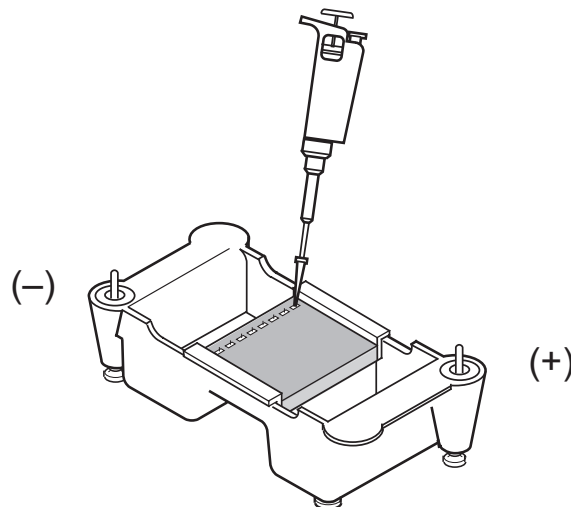
Background

Agarose Gel Electrophoresis

Agarose gel electrophoresis separates DNA fragments by size. PCR products or other DNA fragments are loaded into an agarose gel slab, which is in a chamber filled with a conductive buffer solution. A direct current is passed between wire electrodes at each end of the chamber. Since DNA fragments are negatively charged, they will be drawn toward the positive pole (anode) when placed in an electric field. The matrix of the agarose gel acts as a molecular sieve through which smaller DNA fragments can move more easily than larger ones. Therefore, the rate at which a DNA fragment migrates through the gel is inversely proportional to its size in base pairs (bp). Over a period of time, smaller DNA fragments will travel farther than larger ones. Fragments of the same size stay together and migrate as single bands of DNA.

The sizes of DNA fragments are determined through comparison with a molecular weight standard — in the case of this lab, a molecular weight ruler that has a series of bands differing in size by 500 bp increments from 500 bp up to 5,000 bp. The expected sizes of the target *GAPDH* PCR products range from 500 bp to 2,500 bp, depending on the plant species chosen.

Agarose gel electrophoresis is a useful tool since it allows determination of the number of PCR products (useful since multiple *GAPDH* genes may be amplified using this protocol), the size of these PCR products, the success of the PCR (the presence and intensity of the DNA bands), whether there has been contamination of the sample based on examination of the negative control, and whether primer-dimers have been amplified.



Instructor's Advance Preparation

In this Electrophoresis chapter, it is assumed that students are experienced with agarose gel electrophoresis and will require minimal guidance. Time constraints may determine when students analyze their PCR reactions.

The products to be analyzed prior to the ligation stage are:

- Initial PCR reactions (5 samples)
- Nested PCR reactions (5 samples)
- Purified PCR products (2 samples; optional from Chapter 4)

To save time it is acceptable for students to analyze their initial PCR and nested PCR products on a single gel after the nested round of PCR has been completed. This will also allow direct comparison of PCR products. However, be sure to load 4x less nested PCR product (5 μ l rather than 20 μ l), since the nested PCR bands are often very intense compared to the initial PCR products. You may also wish to compare initial PCR reactions before and after exonuclease treatment. However, it is rare to actually see a difference because the primers are at a very low concentration after the PCR reaction. If this is done, ensure that sufficient PCR product is available for the nested PCR reaction.

UView 6x Loading Dye and Stain, a safe, nontoxic combination loading dye and nucleic acid stain, is included with this series. The loading dye and stain is added directly to samples, the gel is run, and gels can be directly imaged on a UV imaging system. An alternative means of visualizing DNA is Bio-Rad's Fast Blast DNA staining solution (catalog #1660420EDU), which is a biologically safe DNA stain that does not require any documentation system to visualize the DNA. Fast Blast stain is around 5x less sensitive than ethidium bromide, which may mean some faint DNA bands that might be visible with ethidium bromide may not be visible with Fast Blast stain. Another alternative is SYBR[®] Green I, a DNA stain that also requires a visualization and documentation system.

Note: To save time on electrophoresis, refer to Appendix A1, step 3 for an alternative protocol for running buffer that allows gels to be electrophoresed faster than usual.

Electrophoresis Checklist

Components from Cloning and Sequencing Explorer Series

	Where Provided	(✓)
Agarose powder	Electrophoresis Module	<input type="checkbox"/>
Electrophoresis buffer, 50x TAE	Electrophoresis Module	<input type="checkbox"/>
500 bp molecular weight ruler	<i>GAPDH</i> PCR Module	<input type="checkbox"/>
UVView 6x loading dye and stain	<i>GAPDH</i> PCR Module	<input type="checkbox"/>
Microcentrifuge tubes for preparing and aliquoting samples	Cloning and Sequencing Series	<input type="checkbox"/>

Required Accessories (Not Provided)

	Quantity
Horizontal electrophoresis chambers	12
Power supplies	3–12
20 µl adjustable-volume micropipets and aerosol barrier filter tips	12
Gel visualization and documentation system	1
Equipment to cast agarose gels	12

Tasks to Perform Prior to the Lab

1. (Optional) If desired, instructors may prepare 1% agarose gels and electrophoresis running buffer for students according to standard protocols (see Appendix A1).
2. Prepare molecular weight ruler: Add 80 µl of 6x loading dye and stain to 400 µl of 500 bp molecular weight ruler. Use 10 µl of molecular weight ruler per gel. Store at 4°C.

Protocol

Overview

The products of both the initial and nested PCR reactions will be analyzed by agarose gel electrophoresis to assess PCR success; the number of amplified bands and their sizes will be examined to evaluate the success of each round of PCR.

Safety Note: UV light is used to visualize UView-stained DNA. UV light can cause eye damage and burns. Wear UV-protective eye goggles, do not look directly at the light, and avoid skin exposure.

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- **Assess the results of PCR**
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items to electrophorese their PCR samples:

Material Needed for Each Workstation	Quantity
Horizontal electrophoresis chamber	1
Power supply	1
20 µl adjustable-volume micropipets and filter tips	1
Microcentrifuge tubes for preparing samples and aliquoting	10
500 bp molecular weight ruler	12 µl
UView 6x loading dye and stain	25 µl

Common Workstation

Material Required

Materials to cast agarose gels
Gel visualization and documentation system

Preparation for Electrophoresis

1. Plan the gel electrophoresis and complete the table with the samples that will be run on the gel.

Lane Number	Sample	Sample Volume (μl)	Resulting Band Sizes (bp)
1	500 bp molecular weight ruler	10 μl	500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000

2. Refer to Appendix A1 for detailed instructions on casting agarose gels, preparing electrophoresis running buffer, and selecting electrophoresis conditions.

Cast a 1% agarose gel with the appropriate number of wells for analysis.

Prepare sufficient electrophoresis running buffer to run your samples.

Note: Running buffer can be reused 2–3 times.

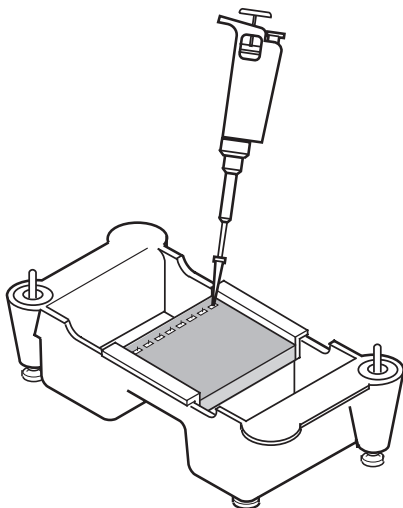
Experimental Procedure for Electrophoresis

1. Label a microcentrifuge tube for each PCR sample.
2. Transfer 20 μl of each initial PCR reaction into a clean labeled microcentrifuge tube. Add 4 μl of loading dye and stain into the tube. Mix up and down to mix.

Note: Do not add loading dye and stain directly to the blue initial PCR tubes — the loading dye can inhibit the next round of PCR. Loading dye and stain is supplied as a 6x concentrate.

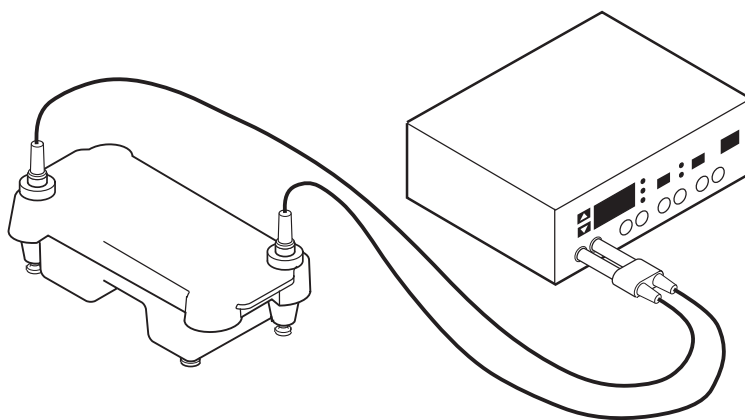
3. To assess the success of the nested PCR round, pipet 5 μl of each yellow nested PCR into a microcentrifuge tube and mix it with 1 μl of 6x loading dye and stain. A smaller volume of the nested PCR is loaded because nested PCR is usually more efficient and produces very intense bands that can obscure bands in adjacent wells if samples are overloaded.
4. Place a 1% agarose gel in the electrophoresis chamber. Pour electrophoresis running buffer into the chamber until it just covers the gel by 1–2 mm.
5. Load 10 μl of the 500 bp molecular weight ruler into the first well. The bands of the ruler are 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500 and 5,000 bp. See example gel in next section.

6. Load 20 μ l from each microcentrifuge tube containing an initial PCR with added loading dye. Load 5 μ l from each tube containing a nested PCR with added loading dye into the wells of the gel according to your plan.



7. Connect your electrophoresis chamber to the power supply and turn on the power. If you are using 1x TAE as the electrophoresis running buffer, run the gel at 100 V for 30 min. If you are using the fast gel protocol (Appendix A1, step 3) with 0.25x TAE buffer, run the gel at 200 V for 20 min.
8. On completion of electrophoresis, visualize the DNA bands (using appropriate safety equipment) and acquire an image of your gel according to your instructor's directions.

Paste image below and label it.



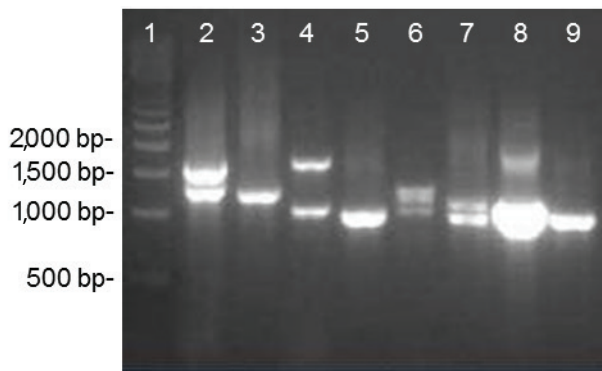
Results Analysis from the PCR

Use the table below to tabulate your results for each sample (including controls) that includes: the sample name, whether a band or multiple bands are present, the molecular weight of the bands, and the intensity (darkness) of the bands relative to a particular band on the 500 bp molecular weight ruler.

Sample Name	# of Bands	Molecular Weight of Band(s)	Intensity of Band(s)

The fragment of *GAPC* that has been targeted in this laboratory series varies in size between plant species. The expected molecular weights of fragments from the initial round of PCR are 0.5 to 2.5 kb. The expected molecular weights of fragments from the second round of nested PCR are around 80 bp smaller than those of the PCR product from the initial round. It is likely that multiple DNA fragments of similar molecular weights will be amplified from the genomic DNA (gDNA) of some plants, most likely due to amplification of multiple *GAPC* genes that are highly homologous.

For example, three bands are frequently amplified from the initial PCR of *Arabidopsis* gDNA and two of these are amplified after nested PCR.



An example of electrophoresis results of the initial and nested PCR. A 1% TAE agarose gel was loaded with initial (20 μ l) or nested (5 μ l) PCR samples generated from specified genomic DNA extracted using the Nucleic Acid Extraction module or control DNA.

Lane 1 — 500 bp molecular weight ruler (10 μ l)

Lane 2 — PCR of grass gDNA with initial primers (20 μ l)

Lane 3 — PCR of product from lane 2 with nested primers (5 μ l)

Lane 4 — PCR of thistle gDNA with initial primers (20 μ l)

Lane 5 — PCR of product from lane 4 with nested primers (5 μ l)

Lane 6 — PCR of control *Arabidopsis* gDNA with initial primers (20 μ l)

Lane 7 — PCR of product from lane 6 with nested primers (5 μ l)

Lane 8 — PCR of pGAP plasmid control with initial primers (20 μ l)

Lane 9 — PCR of pGAP plasmid control with nested primers (5 μ l)

What conclusions can you draw from your PCR experiment? Did you obtain PCR products of the expected molecular weights from your experimental plants?

Deciding Which Plant *GAPC* to Clone

Once all students in the class have their results, it is time to decide which plant will be used for cloning. Although two plants were chosen for investigation, only a single plant's *GAPC* gene should be cloned. It is recommended that the plant chosen be the one that generated the cleanest PCR product (fewest background bands), with good band intensity of an appropriately sized fragment. It is acceptable to clone doublets since each linearized plasmid is expected to ligate to a single DNA fragment. Be aware that cloning doublets may result in colonies of transformed *E. coli* containing different gene sequences.

It is highly recommended that the entire class clone *GAPC* from the same plant, so that the data obtained will be more reliable. Cloning the same gene multiple times will provide this reliability, called depth of coverage, which will help to resolve any ambiguous base pairs when the gene is sequenced. Remember that the sequencing data obtained from this lab may provide new data for the scientific community at large; thus, it is vital the data provided be as accurate as possible.

It is recommended that one or two student groups perform an additional PCR purification, ligation, and transformation of the *Arabidopsis* *GAPC* PCR product, both as a positive control for the class and as a backup in case ligations using PCR products from an experimental plant are unsuccessful.

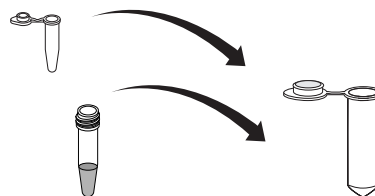
State which plant gene the class has decided to clone:

Focus Questions for Electrophoresis

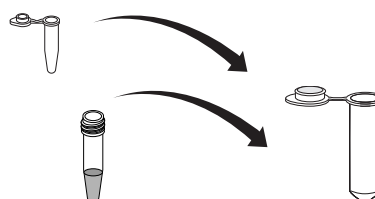
1. What is the purpose of the agarose gel?
2. What purpose does electrophoresis running buffer serve?
3. What purpose does the loading dye serve?
4. What is the purpose of the molecular weight ruler?

Electrophoresis – Quick Guide

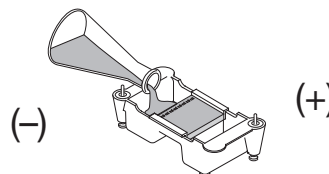
1. Label a microcentrifuge tube for each PCR.
2. DO NOT add loading dye directly to blue initial PCR reactions. In a separate tube add 4 μ l of 6x loading dye and stain and add 20 μ l of initial PCR reaction and pipet up and down to mix.



3. Add 1 μ l of 6x loading dye and 5 μ l of yellow nested PCR reaction to a separate tube and pipet up and down to mix.

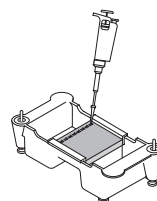


4. Place a 1% agarose gel in the electrophoresis chamber. Pour electrophoresis buffer into the chamber until it just covers the gel by 1–2 mm.

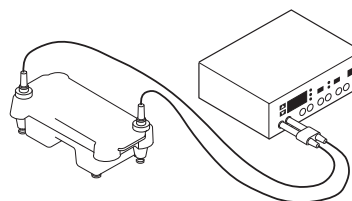


5. Load 10 μ l of the 500 bp molecular weight ruler into lane one of the gel.

6. Referring to your electrophoresis plan, load 20 μ l of blue initial PCR reactions with loading dye and 5 μ l of the corresponding yellow nested PCR reactions with loading dye into the wells of a 1% gel.



7. Connect the electrophoresis chamber to the power supply and turn on the power. Run the gel at 100 V for 30 min unless otherwise instructed.



8. Upon completion of electrophoresis, visualize your DNA fragments according to your instructor's directions.

CHAPTER 4: PCR PURIFICATION

Background

Purifying PCR Products for Further Analysis

Before PCR products can be used for cloning, they must be purified. In addition to the PCR product, the PCR reaction mix contains unincorporated dNTPs, primer-dimers, unused primers, and DNA polymerase that may interfere with subsequent experiments. Unfortunately, the commonly used methods for purification of DNA, such as alcohol precipitation and phenol-chloroform extraction, do not remove primers. Chromatography, a separation method based on characteristics of the sample components, such as their size and charge, is the most common method used to purify PCR products from other components of the reaction mix. The two common types of chromatography used are ion exchange and size exclusion, which separate molecules based on charge and size, respectively. In this laboratory activity, size exclusion is used to purify PCR products.

Principles of Size Exclusion Chromatography

In size exclusion (also called gel filtration) chromatography, molecules in solution are separated by size as they pass through a column of cross-linked beads that form a three-dimensional network. These polymer beads (frequently made of dextran, agarose, or acrylamide) have pores of a specific size. There are many types of size exclusion materials. The choice of material depends on the range of size of the molecules to be separated and the goal of the separation, for example clean separation of molecules above a certain size vs. fractionation of molecules having a range of sizes. Different bead types have pores of different sizes, which allow you to achieve your goal.

To visualize how separation occurs using these beads, imagine that the beads are like tiny perforated wiffle balls, and the molecules in the solution are different sized particles falling through a vertical column of the wiffle balls — small particles would enter the wiffle balls and be slowed in their progress through the column, while larger particles would not enter the wiffle balls, and will go through the column much faster than the small particles.

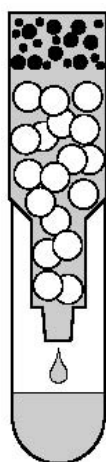
As the components of a sample pass through the column, there are two routes that molecules can take through the column. The path they take will depend on their size and on the size of the pores in the beads. Molecules that are larger than the pores will not enter the beads, staying in the solution surrounding the beads. Hence they elute first from the column. Smaller molecules will enter the pores in the beads and so move more slowly through the column (see figure on next page). Molecules of intermediate size will enter the beads to some extent, but will not spend as much time there as do the smaller molecules. To summarize, larger molecules will elute from the column first, and the smallest molecules will elute last, with intermediate-size molecules strung out in between.

Size exclusion chromatography is used to purify large PCR products (the longer molecules) from the smaller primers, dNTPs and enzymes. A small spin column is filled with size exclusion beads, which are packed to form a column bed. The PCR reaction mix is loaded on top of the beads, and the column is placed in a microcentrifuge tube and centrifuged. The PCR product will pass through the beads and be collected in the bottom of the tube, while smaller nucleic acids, such as primer-dimers, primers, and dNTPs, as well as enzymes, remain in the column, which is discarded. Recovery of PCR products is dependent on product size, with larger fragments yielding higher recovery than smaller fragments.

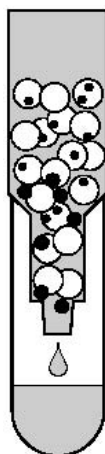
Instructor's Advance Preparation

In this PCR Purification chapter, students will purify their PCR reactions using the PCR Kleen spin columns to remove excess primers, nucleotides, and enzymes. Columns are prepared for chromatography by resuspending the beads in the column and then centrifuging to create the column bed. Samples are then applied to the column bed and centrifuged to elute the purified PCR product. It is important that the speed of the centrifuge for this stage be adjusted according to the protocol. Details are in the protocol, but the speed required (735 x g) is much slower than most benchtop microcentrifuges.

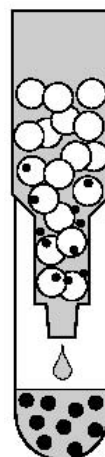
CHAPTER 4 ADVANCE PREPARATION



A mixture of large and small molecules is applied to a column of porous beads.



As the buffer flows through the column, the small molecules penetrate into the beads and are slowed.



The larger molecules emerge from the column first.

Size exclusion chromatography. Open circles represent the size exclusion beads, large filled circles are large molecules, and small filled circles are small molecules.

Note: As an alternative to the calculation for determining rpm in the main protocol, many conversion tables and calculators are available on the Internet.

Following purification of the PCR, you have the option to analyze the PCR products before and after purification. This will depend on the time constraints of the course.

PCR Purification Checklist

Components from Cloning and Sequencing Explorer Series

	Where Provided	(✓)
PCR Kleen spin columns, clear	PCR Purification Module	<input type="checkbox"/>
Microcentrifuge tubes, 1.5 ml	PCR Purification Module	<input type="checkbox"/>
Capless collection tubes, 2.0 ml	PCR Purification Module	<input type="checkbox"/>
(Optional) electrophoresis reagents	Electrophoresis Module	<input type="checkbox"/>

Required Accessories (Not Provided)

	Quantity
Microcentrifuge with variable-speed setting	1
Vortex mixer (if available)	1
200 µl adjustable-volume micropipet and aerosol barrier filter tips	12
(Optional) electrophoresis equipment	12

Tasks to Perform Prior to the Lab

1. (Optional) Prepare agarose gels and electrophoresis running buffer for analysis of products.
2. Ensure that all preparation for the Transformation Stage has been completed. Have a fresh starter plate of bacteria, LB broth, and LB amp IPTG plates ready for the transformation activity.

Protocol

Overview

The next step after using PCR to generate DNA fragments is to find a way to maintain and sequence these products. Ligating (inserting) the fragments into a plasmid vector that can be propagated in bacteria accomplishes this. To optimize the success of ligation, unincorporated primers, nucleotides, and enzymes must be removed from the PCR. Using size exclusion chromatography, small molecules, like proteins, primers, and nucleotides, get trapped inside the chromatography beads, while large molecules, like the PCR products, are too large to enter the beads and pass through the column into the collection tube. The columns will also remove loading dye.

Bio-Rad PCR Kleen spin columns are designed to be used in variable-speed benchtop microcentrifuges capable of generating a force of 735 x g. The top speed on most benchtop microcentrifuges is 12,000–14,000 rpm, which equates to 14,000–16,000 x g, much more force than is recommended. Thus, the columns should be centrifuged on a slow setting. The speed setting will depend on the radius of the rotor in the microcentrifuge. The settings on most microcentrifuges are in revolutions per minute (rpm) rather than in relative centrifugal force (rcf), also known as gravitational force (g-force); therefore, a conversion must be performed. The g-force created by a particular number of rpm is a function of the radius of the microcentrifuge rotor. Consult the microcentrifuge instruction manual for conversion information from rpm to g-force. Alternatively, to calculate the speed (in rpm) required to reach a g-force of 735 x g, use the following equation:

$$rcf (g) = (1.12 \times 10^{-5})(s)^2r \quad \text{or} \quad s = \sqrt{\frac{rcf}{1.12 \times 10^{-5} \times r}}$$

in which *r* is the radius in centimeters measured from the center of the rotor to the middle of the PCR Kleen spin column, and *s* is the speed of the rotor in rpm.

For reference, the Bio-Rad microcentrifuge (catalog #1660612EDU) has a radius of 5 cm, so the required setting for centrifuging PCR Kleen spin columns is ~3,600 rpm.

Note: Do not use the pulse button to centrifuge the columns. The pulse button overrides the variable-speed setting and can cause column failure.

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- **Purify the PCR product**
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items to purify one PCR product:

Material Needed for Each Workstation	Quantity
PCR product to be cloned	1
PCR Kleen spin column, clear	1
Capless collection tube, 2 ml	1
Microcentrifuge tube, 1.5 ml	1
200 µl adjustable-volume micropipet and filter tips	1
Marking pen	1

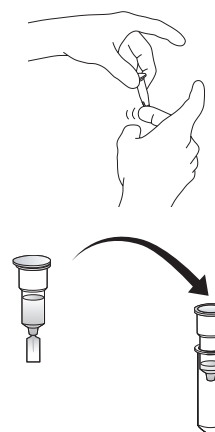
In addition, materials for electrophoresis may be required for an optional step, which will depend on time constraints in the course.

Common Workstation

Material Required	Quantity
Microcentrifuge with variable-speed setting	1
Vortex mixer (if available)	1

Experimental Procedure for PCR Product Purification

1. Obtain the yellow nested PCR reaction tube for the plant gene fragment to be cloned.
2. Label a PCR Kleen spin column and a capless collection tube with your initials. Label a capped microcentrifuge tube with your initials, "purified PCR product," and the plant name.
3. Resuspend the beads in the PCR Kleen spin column by vortexing (or flicking it vigorously if no vortex mixer is available). Return the beads to the bottom of the column with a sharp downward flick.
4. Snap the bottom off the spin column, remove the cap, and place the column in the capless collection tube. Discard cap and bottom of column.



- Place the spin column, still in the capless collection tube, into the microcentrifuge. Make sure that your tube is placed in the rotor with another group's or with a balance tube so that the microcentrifuge is balanced; accommodate classmates' tubes to ensure economic use of the microcentrifuge.

Centrifuge columns at 735 x g for 2 min. **Do not use the top speed of the microcentrifuge for this step.**

This step ensures a uniformly packed column and removes the storage buffer from the column.

- Move the spin column to the labeled microcentrifuge tube. Discard the flowthrough and the collection tube.
- Pipet 30 μ l of the yellow nested PCR onto the top of the column bed in the spin column, without disturbing the resin (or column bed).

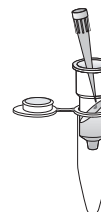
If agarose gel electrophoresis will be performed to analyze the results of the purification, save 5 μ l of this original (unpurified) yellow PCR sample and mix it with 1 μ l of 6x loading dye and stain for the gel analysis.

- Place the column, still in the labeled microcentrifuge tube, into the microcentrifuge. It is best to orient the cap of the microcentrifuge tube downward, toward the center of the rotor, to minimize friction and damage to the cap during centrifugation.

Centrifuge at 735 x g for 2 min. Make sure that another group's sample counterbalances the microcentrifuge.

In this step, unincorporated dNTPs, polymerase, primers, and primer-dimers associate with the beads while the PCR product does not.

- Remove the spin column from the microcentrifuge tube and discard the column. Cap the microcentrifuge tube, which now contains the purified PCR product. Store at 4°C for up to 2 weeks or at -20°C long term.
- (Optional) Add 1 μ l of 6x loading dye and stain into 5 μ l of the purified sample. Electrophorese 5 μ l of this mix next to 5 μ l of the unpurified sample on a 1% agarose gel to visualize differences between the original and the purified sample.

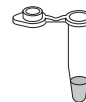


Focus Questions for PCR Purification

1. Why is it necessary to purify PCR products?
2. What type of chromatography is used to purify PCR products?
3. Are there other methods to purify PCR products?

PCR Purification — Quick Guide

1. Obtain the yellow nested PCR reaction tube for the plant to be cloned.



2. Label a PCR Klean spin column and a capless collection tube with your initials. Label a capped 1.5 ml microcentrifuge tube with your initials, “purified PCR fragment,” and the plant name.



3. Resuspend the beads in the PCR Klean spin Column by vortexing or flicking vigorously and return the beads to the bottom of the column with a sharp downward flick.



Flick

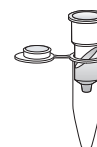
4. Snap the bottom off the column, remove the cap and place the column in a capless collection tube.



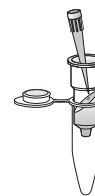
5. Place the spin column in the collection tube into a microcentrifuge and centrifuge at 735 x g for 2 min. **Do not use the top speed of the microcentrifuge for this step.**



6. Move the spin column to the labeled 1.5 ml microtube. Discard the flowthrough and collection tube.



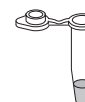
7. Pipet 30 μ l of the yellow nested PCR reaction on top of the beads in the spin column.



8. Place the spin column in the microcentrifuge tube and transfer to a microcentrifuge. Centrifuge at 735 x g for 2 min.



9. Remove the spin column from the microcentrifuge tube and discard the spin columns. Cap the 1.5 ml microcentrifuge tube containing the purified PCR product. Store at 4°C for up to two weeks or -20°C for long term.



CHAPTER 5: LIGATION

Background

Cloning Vectors

Once a gene or part of a gene has been amplified using PCR, the next step in cloning is to insert the DNA into a plasmid or cloning vector so that the fragment can be propagated. In infectious disease, a vector is defined as an organism that transmits an infectious agent from one host to another. In molecular biology, the definition is similar. A vector is an agent (such as a bacteriophage or plasmid) that is used to transfer genetic material into a cell.

Many cloning vectors are derived from bacterial plasmids. Plasmids are extrachromosomal DNA, usually circular DNA molecules 2,000–100,000 base pairs (bp) long, although most plasmids used in cloning are 2,000–10,000 bp. Bacteria may naturally contain many copies of a single plasmid, or single copies of others. Plasmids are able to replicate independently of the host DNA and most plasmids carry at least one gene. Frequently these genes code for a factor or function that helps the bacteria survive. For example, resistance to the antibiotic ampicillin is conveyed by a plasmid carrying an ampicillin-resistance gene. Plasmids are capable of being transferred from one bacterium to another. These characteristics have resulted both in wonderful new uses for plasmids (such as their use in cloning, making many of the techniques of molecular biology possible) and in the emergence of dangerous pathogenic organisms (namely, bacteria resistant to multiple antibiotics).

Plasmids thus already have many of the characteristics needed for use as cloning vectors, and other useful features have been added through genetic engineering. A wide variety of vectors are commercially available for various applications. A plasmid designed to clone a gene is different from a plasmid designed to express a cDNA in a mammalian cell line, which is different again from one designed to add a tag to a protein for easy purification. The primary characteristics of any good vector include:

- Self-replication — plasmids have an origin of replication so they can reproduce independently within the host cell; since the origin of replication engineered into most cloning vectors is bacterial, the plasmid can be replicated by enzymes already present in the host bacteria
- Size — most bacterial vectors are small, between 2,000 and 10,000 bp long (2–10 kilobases or kb), making them easy to manipulate
- Copy number — each plasmid is found at specific levels in its host bacterial strain. A high copy number plasmid might have hundreds of copies in each bacterium, while a low copy number plasmid might have only one or two copies per cell. Cloning vectors derived from specific plasmids have the same copy number range as the original plasmid. Most commonly used vectors are high copy number
- Multiple cloning site (MCS) — vectors have been engineered to contain an MCS, a series of restriction sites, to simplify insertion of foreign DNA into the plasmid. An MCS may have 20 or more different enzyme sites, each site usually unique both in the MCS and in the plasmid. This means that for each restriction site included in the MCS, the corresponding restriction enzyme will cut the plasmid only at its single site in the MCS
- Selectable markers — plasmids can carry one or more resistance genes for antibiotics, so if the transformation is successful (that is, if the plasmid enters and replicates in the host cell), the host cell will grow in the presence of the antibiotic. Therefore, antibiotics can be used as markers to select for positive transformants. Commonly used selectable markers are genes for resistance to ampicillin (*amp^r*), tetracycline (*tet^r*), kanamycin (*kan^r*), streptomycin (*sm^r*), and chloramphenicol (*cm^r*)

- Screening — when bacteria are being transformed with a ligation reaction, not all of the religated vectors will necessarily contain the DNA fragment of interest. To produce visible indicators that cells contain an insert, vectors frequently contain reporter genes, which distinguish them from cells that do not have inserts. Two common reporter genes are β -galactosidase (β -gal) and green fluorescent protein (GFP)

Some newer plasmid vectors use positive selection, in which the inserted DNA interrupts a gene that would otherwise be lethal to the bacteria. If foreign DNA is not successfully inserted into the MCS, the lethal gene is expressed and transformed cells die. If the foreign DNA is successfully inserted, the lethal gene is not expressed and the transformed bacteria survive and divide. Positive selection eliminates the need for reporter genes, as only cells transformed with vector containing an insert will survive

- Control mechanism — most vectors have some control mechanism for transcription of the antibiotic resistance or other engineered gene. One of the best-known control mechanisms is the *lac* operon (an operon is a group of genes). When lactose (a sugar) is absent in the cell, the *lac* repressor protein binds to the *lac* operon, preventing transcription of the gene. When lactose is present in the cell, it binds to the *lac* repressor protein, causing the repressor protein to detach from the operon. With the repressor protein no longer bound to the operon, RNA polymerase can bind and the genes can be transcribed. Lactose acts as an inducer of the *lac* operon. (A compound closely related to lactose, isopropyl β -D-1-thiogalactopyranoside (IPTG), is often used in the lab as an artificial inducer.) Genes from the *lac* operon have been engineered into many cloning vectors
- Size of insert — plasmid vectors have limitations on the size of inserts that they can accept, usually less than the size of the vector. Other vectors have been developed for use if the target DNA is larger, for example, lambda phage (inserts up to 25 kb), cosmids (insert up to 45 kb), bacterial artificial chromosomes (BACs; inserts from 100 to 300 kb), yeast artificial chromosomes (YACs; insert from 100 to 3,000 kb), and bacteriophage P1 (inserts up to 125 kb)

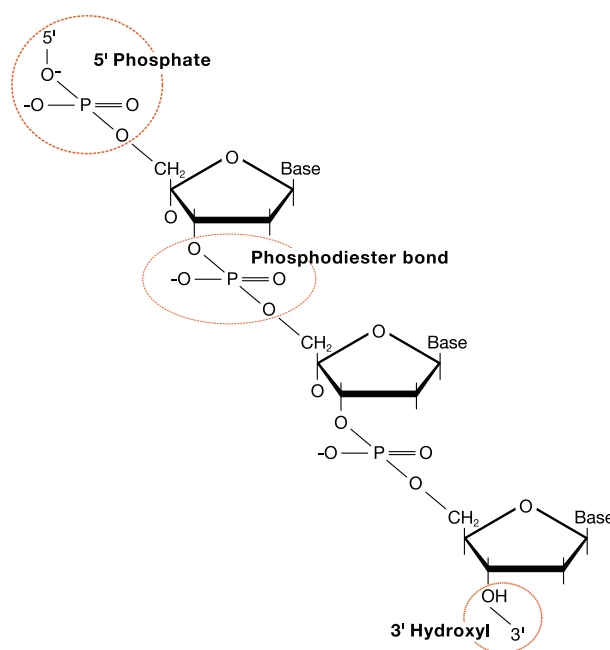
DNA Ligation

Ligation is the process of joining two pieces of linear DNA into a single piece through the use of an enzyme called DNA ligase. DNA ligase catalyzes the formation of a phosphodiester bond between the 3'-hydroxyl on one piece of DNA and the 5'-phosphate on a second piece of DNA.

Cloning and Sequencing Explorer Series

The most commonly used DNA ligase is T4 DNA ligase (named because it originated in a bacteriophage named T4). There are several ways that the efficiency of DNA ligation can be optimized. First, like any enzyme, there are conditions that are optimal for ligase activity:

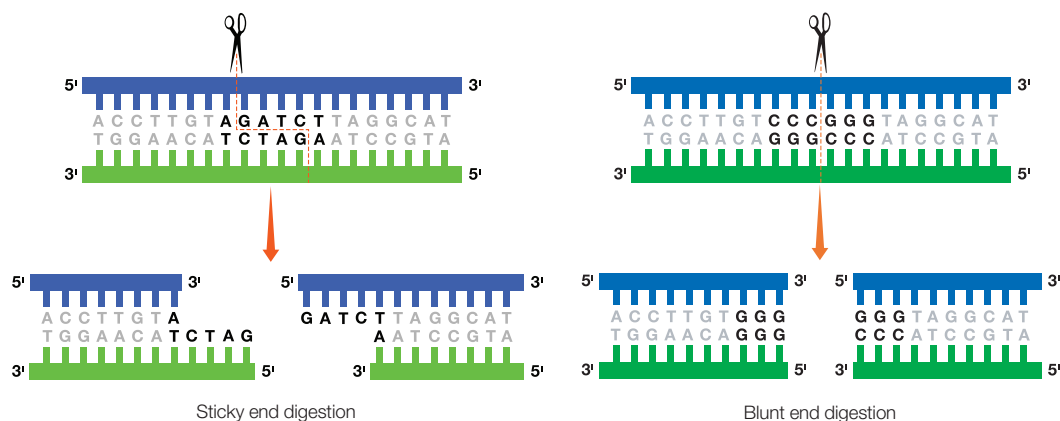
- T4 DNA ligase requires ATP and magnesium ions for activity
- The concentration of vector and insert DNA in solution must be high for efficient ligation
- The molar ratio of insert to vector DNA should be approximately equal, although the optimal ratio may not be 1:1



Chemical structure of deoxyribose nucleic acid (DNA).

Ligation is used to join vector DNA and insert DNA. There are two ways in which DNA can be ligated into a cloning vector, one using DNA with so-called “sticky ends” and the other using DNA with “blunt ends.” Unlike DNA with blunt ends, DNA with sticky ends has one or more unpaired bases at its ends that do not have complementary bases on the other strand, typically referred to as an “overhang.” When a DNA fragment is generated by PCR using Taq polymerase, it typically has sticky ends with a single adenosine (A) overhang. When a DNA fragment is generated by cutting a piece of DNA with a restriction enzyme (an enzyme that cuts both strands of double-stranded DNA), it may have either sticky ends or blunt ends, depending on the restriction enzyme. (For more information on restriction enzymes, refer to the Background section in the Plasmid Purification chapter — Chapter 7.)

DNA ligation with sticky ends — to prepare a cloning vector for ligation with insert DNA, it is cut with a restriction enzyme within the MCS, opening it to receive the inserted DNA. If the insert has sticky ends, that is, overhangs on the ends of the DNA strands, then the vector should be cut with the same enzyme, producing sticky ends that will be complementary to the ends of the insert DNA. For example, if the insert DNA has been prepared by cutting it at both ends with Bgl II, then the vector would also be cut with Bgl II. Having complementary sticky ends improves the efficiency of ligation, whereas mismatches in the sequences make it less likely the ends insertion will take.



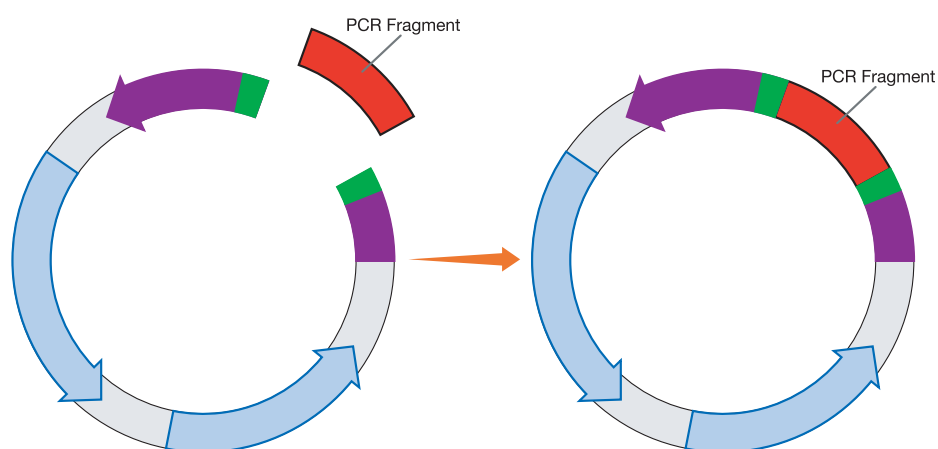
Because the sticky ends on the vector and the insert are complementary, when they come into contact during the ligation reaction they will base-pair with each other. (The base-paired sticky ends of the insert and vector are not stably associated, and they can dissociate prior to ligation.) While the insert and vector are associated, T4 DNA ligase forms a phosphodiester bond, covalently linking the two pieces of DNA. Actually, there are two ligations. The first ligation is intermolecular, between one end of the vector and one end of the insert, resulting in a linear DNA molecule. The second ligation is intramolecular, circularizing the molecule.

One advantage to sticky-end ligation is that it makes directional cloning possible. If it is desirable to have the insert in one orientation only (for instance, in the $A \Rightarrow B$ direction in the vector, but not in the $B \Rightarrow A$ direction), then the insert and vector can both be digested with two different restriction enzymes so that their ends are asymmetric. This is important if the DNA insert is a cDNA to be expressed in the transformed cell. When this is done, only the complementary ends will ligate and the insert will have a single orientation in the ligation products.

Cloning and Sequencing Explorer Series

DNA ligation with blunt ends — blunt-end ligation, in which both the inserted DNA and the vector have blunt ends, has an advantage over sticky-end ligation in that all DNA ends are compatible with all other ends. In other words, it is not necessary to cut the vector and insert with the same restriction enzymes to get complementary overhangs as for sticky-end ligation. Vectors used for blunt-end ligation have a blunt-ended ligation site in the MCS. They still have an MCS, as the restriction enzyme sites are very useful for subsequent manipulation of the inserted DNA.

In PCR, *Taq* DNA polymerase adds a single nucleotide to the 3'-end of the PCR product, usually an A. Since this A overhang would prevent blunt-end ligation, it must be removed prior to ligation. Proofreading polymerases have the ability to correct mistakes by one base pair in the reverse direction. Since an A overhang would look like an error, the proofreading polymerase would go back and excise the A overhang. Treating the PCR product with a proofreading DNA polymerase removes the 3'-A, leaving blunt ends ready for ligation. (Not all thermally stable DNA polymerases used in PCR leave an A overhang; some polymerases, like *Pfu* DNA polymerase, have proofreading ability.)

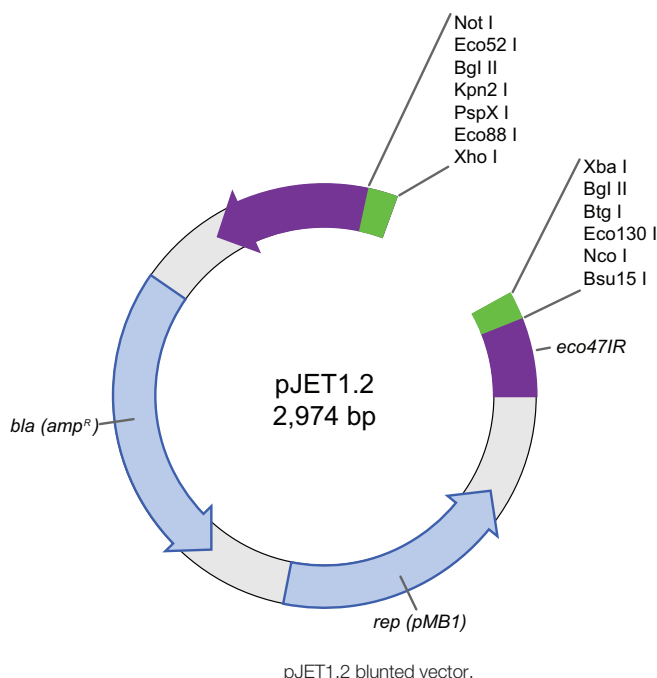


Ligation of PCR fragment into vector.

Features of the pJET1.2 Blunted Vector

The pJET1.2 blunted cloning vector has several features that make it a good choice for this lab:

- It is a vector designed for blunt-end cloning, so is already linearized opened with blunt ends
- Its MCS has restriction enzyme sites that can be used for later manipulation of the DNA
- It is a high copy number plasmid
- It contains the β -lactamase gene *amp^r*, which confers resistance to ampicillin
- It contains the *eco47IR* gene, which allows positive selection of transformants. This gene codes for the *Eco47I* restriction enzyme, which is toxic to *E. coli* when it is expressed. When the *eco47IR* gene is disrupted by the insertion of DNA into the cloning site, the gene will no longer be expressed and the transformed cells will grow and divide on selective media
- It is 2,974 bp in length and its sequence is readily available — the accession number for pJET1.2 is EF694056



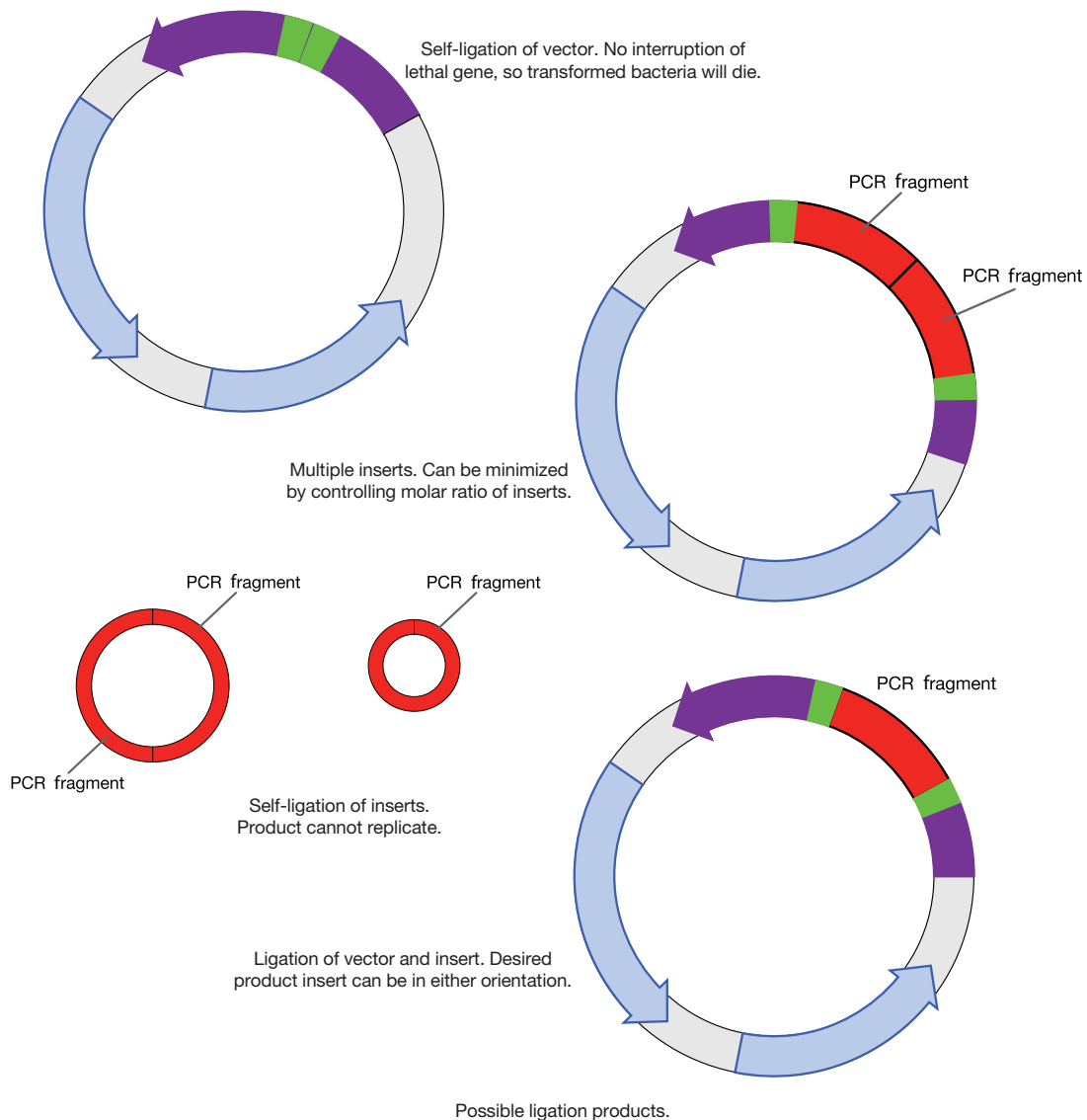
Products of Ligation

Ligation is a very inefficient process; from millions of vectors and inserts, only 1–100 are expected to ligate together as desired and lead to growth of colonies from the desired transformants. There are several possible products from a ligation with a vector such as the pJET1.2 vector:

- Self-ligation of the vector — a self-ligated vector, without any DNA inserted in the MCS, should have an intact lethal gene. The product of the lethal gene should kill any bacteria transformed by these vector-only constructs (however, the *eco47IR* gene is not 100% lethal, and so some bacteria may grow, resulting in tiny colonies, which may be observed in no-insert ligation controls)
- Ligation of a vector with primer-dimers or other short DNA fragments from the PCR reaction — even if these fragments are a small proportion of the total fragments available for ligation, small fragments ligate more easily than larger fragments. It is likely that some plasmid minipreps will appear to grow colonies without any insert ligated, when they have actually ligated small DNA fragments
- Ligation of a vector with multiple inserts — since the inserts all have the same blunt ends, they can ligate to each other (a process known as concatenation) and also ligate to the vector, giving a product with multiple inserts in a row. The number of these products can be reduced by controlling the molar ratio of insert to vector. If the ratio is high (that is, if there are many more insert molecules than vector), then it is more probable that inserts will ligate to each other rather than to the vector. So the molar ratio can be an important factor in setting up ligation reactions
- Self-ligation of insert — it is possible for the insert molecules to self-ligate, forming closed circles. Since these molecules do not have any of the vector genes, they will not be able to replicate to form colonies, so they are not of consequence
- Ligation of one insert into a vector — this is the desired result, and in ligation to a vector such as pJET1.2 blunted vector, there is no directionality in the cloning. In other words, the insert can ligate into the vector in either direction, because all of the ends are identical. So the products should be a 50:50 mix, with half the inserts in one orientation and the other half in the reverse orientation

Instructor's Advance Preparation

At this ligation stage, students will ligate the purified PCR products into a plasmid vector, pJET1.2. This vector is already linearized, or opened, leaving blunt ends ready for ligation to PCR products. Rather than using the traditional *lacZ* gene to assist with colony selection using blue/white colony screening on selective media (LB plates with X-gal and IPTG), pJET1.2 selects successful ligations through the disruption of an otherwise lethal gene, *eco47IR*, which enables positive selection of the recombinants. Before ligation, a 3'-A overhang must be removed from the PCR products, leaving them with blunt ends ready to be ligated to the pJET1.2 plasmid.



In this laboratory, the ligation reaction is fast, complete in 5–10 minutes. Only a minimal increase in the number of transformants is gained by extending the duration of the ligation reaction beyond 10 minutes.

Ligation Checklist

Components from Cloning and Sequencing Explorer Series

	Where Provided	(✓)
Purified PCR product	Previously prepared (PCR Purification Module)	<input type="checkbox"/>
2x ligation reaction buffer*	Ligation and Transformation Module	<input type="checkbox"/>
Proofreading polymerase*	Ligation and Transformation Module	<input type="checkbox"/>
T4 DNA ligase*	Ligation and Transformation Module	<input type="checkbox"/>
pJET1.2 blunted vector*	Ligation and Transformation Module	<input type="checkbox"/>
Sterile water	Ligation and Transformation Module	<input type="checkbox"/>
Microcentrifuge tubes	Ligation and Transformation Module	<input type="checkbox"/>

* Before use, these reagents must be defrosted, thoroughly mixed, and centrifuged to collect contents at the bottom of the tubes.

Required Accessories (Not Provided)

	Quantity	(✓)
Water bath, heating block, or incubator (70°C)	1	<input type="checkbox"/>
Microcentrifuge	1	<input type="checkbox"/>
10 µl adjustable-volume micropipets and tips	12	<input type="checkbox"/>
Ice bath	12	<input type="checkbox"/>
Marking pens	12	<input type="checkbox"/>
–20°C freezer	1	<input type="checkbox"/>

Tasks to Perform Prior to the Lab

1. Defrost the 2x ligation reaction buffer, proofreading polymerase, T4 DNA ligase, pJET1.2 blunted vector, and sterile water on ice. Just before use, mix and centrifuge the reagents to collect contents at the bottom of the tubes. Do not aliquot the ligation reagents for students; the volumes required are too small. Ensure students are familiar with methods to pipet small volumes.
2. In order to complete the lab more efficiently, preparation for the transformation step can be initiated prior to performing the ligation reaction, allowing immediate transformation of competent cells with the products of the ligation reaction. Refer to the Tasks to Perform Prior to the Lab section in the Transformation chapter.

Protocol

Overview

In this stage, you will insert (ligate) the PCR product into a plasmid vector. The plasmid is supplied ready to use, already opened up to receive the DNA fragment. Because this will be a blunt-end ligation, a single adenosine (A) nucleotide left on the 3'-ends of the PCR fragment by *Taq* DNA polymerase must be removed before the ligation so that the fragment will also have blunt ends. The nucleotide is removed and the PCR product blunted by treating the PCR product with a proofreading polymerase. This DNA polymerase is active at 70°C but not at lower temperatures, so it is not necessary to inactivate this enzyme after use.

Once blunted, the PCR product is combined with the plasmid and T4 DNA ligase under conditions optimal for ligation. The ligation reaction will be complete in 5–10 minutes.

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- Purify the PCR product
- **Ligate PCR product into a plasmid vector**
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items to ligate one PCR product.*

Material Needed for Each Workstation	Quantity
Purified PCR product	1–2 µl
2x ligation reaction buffer*	5 µl
Proofreading polymerase*	0.5 µl
T4 DNA ligase*	0.5 µl
pJET1.2 blunted vector*	0.5 µl
Sterile water*	5 µl
Microcentrifuge tube	1
10 µl adjustable-volume micropipet and tips	1
Marking pen	1

* The volumes of reagents required for each student group are provided for your information. However, aliquoting of reagents for students in this exercise is not recommended due to the small volumes required. Please also see Common Workstation.

Common Workstation

Material Required

Ice bucket containing stock tubes of:

- 2x ligation reaction buffer
- Proofreading polymerase
- T4 DNA ligase
- pJET1.2 blunted vector
- Sterile water

Water bath, heating block, or incubator at 70°C

Microcentrifuge

Experimental Procedure for Ligation

1. Label a microcentrifuge tube with your initials, your plant name, and “ligation.”
2. Briefly centrifuge the stock tubes containing the 2x ligation reaction buffer and proofreading polymerase in a microcentrifuge to force contents to bottom of tubes.

Note: Take special care when pipetting very small volumes. Make sure only the soft stop of the pipet is used, when pulling up reagents even though it may feel like a very small movement. Also, look at the end of the pipet tip to be sure that the correct volume of reagent is in the tip. After adding the reagent to the tube, be sure that the pipet tip is empty. Never reuse a pipet tip.

3. Set up a blunting reaction with the following reagents:

Reagent	Amount
2x ligation reaction buffer	5.0 µl
Purified PCR product	1.0 µl*
Sterile water	2.5 µl*
Proofreading polymerase	0.5 µl
Total	9.0 µl

* If the PCR product was not as intense as the 1 kb band in the molecular weight marker in the agarose gel analysis (see results from the Electrophoresis chapter), increase the amount of PCR product added to the blunting reaction to 2 µl. You will need to decrease the volume of sterile water to 1.5 µl to compensate.

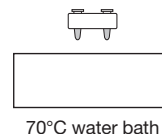
4. Close the cap and mix well. Centrifuge briefly in a microcentrifuge to collect the contents at the bottom of the tube.

This step is essential due to the very small volume used in this reaction.



5. Incubate the tube at 70°C for 5 min.

70°C is the optimal temperature for the proofreading polymerase to blunt the PCR fragment.



6. Cool the tube on ice for 2 min.

This recondenses water vapor to maintain the reaction volume.

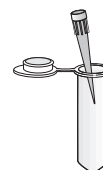
7. Once cooled, centrifuge the tube briefly to collect the contents at the bottom of the tube. Maintain the tube at room temperature.

8. Briefly centrifuge the stock tubes containing the pJET1.2 blunted vector and the T4 DNA ligase in a microcentrifuge to force the contents to bottom of tubes.



9. Set up a ligation reaction with the following reagents:

Reagent	Amount
Blunt reaction (already in microcentrifuge tube)	9.0 µl
T4 DNA ligase	0.5 µl
pJet1.2 blunted vector	0.5 µl
Total	10.0 µl



10. Close the cap and mix well. Centrifuge briefly in a microcentrifuge to collect the contents at the bottom of the tube.



11. Incubate the tube at room temperature for 5–10 min.
12. Store the ligation reaction at -20°C . However, if you are proceeding directly to the transformation, pipet 5 μl of the ligation reaction into a microcentrifuge tube labeled with your initials, plant name, and “transformation” and store it on ice until needed for the transformation.

Focus Questions for Ligation

1. How is the PCR product incorporated into the vector for this procedure?
2. What bond does the ligase enzyme catalyze in DNA?
3. Name one advantage of sticky-end ligation.
4. Name one advantage of blunt-end ligation.

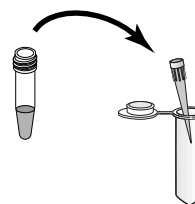
Ligation – Quick Guide

1. Label a microcentrifuge tube with your initials, plant name, and “ligation.”
2. Briefly spin down the stock tubes of 2x ligation reaction buffer and proofreading polymerase to collect the contents at the bottom of the tube.



3. Set up a blunting reaction with the following reagents.

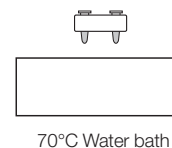
Reagent	Amount
2x ligation reaction buffer	5.0 μ l
Purified PCR product	1.0 μ l
Sterile water	2.5 μ l
Proofreading polymerase	0.5 μ l
Total	9.0 μ l



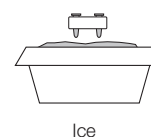
4. Close the cap and mix well. Centrifuge briefly to collect the contents at the bottom of the tube.



5. Place the tube into a water bath at 70°C for 5 min.



6. Place tube on ice to cool for 2 min.



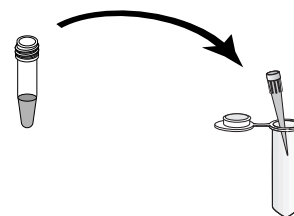
- Once cooled, centrifuge briefly to bring contents to the bottom of the tube and maintain the tube at room temperature.



- Spin down stock tubes of pJET1.2 blunted vector and T4 DNA ligase to collect the contents at the bottom of the tubes.

- Set up a ligation reaction with the following reagents.

Reagent	Amount
Blunting reaction (already in microcentrifuge tube)	9.0 μ l
T4 DNA ligase	0.5 μ l
pJET1.2 blunted vector	0.5 μ l
Total	10.0 μ l

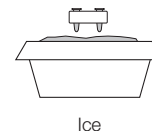


- Close the cap on the tube and mix well. Centrifuge briefly to collect the contents at the bottom of the tube.



- Incubate the tube at room temperature for 5–10 min.

- Store the ligation reaction at -20°C . If proceeding directly to transformation step, transfer 5 μ l of the ligation reaction into a clean microcentrifuge tube and store on ice.

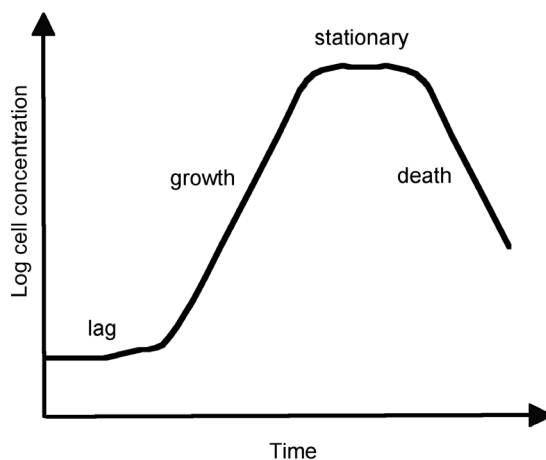


CHAPTER 6: TRANSFORMATION

Background**Transformation**

Once a gene or part of a gene has been amplified using PCR and ligated into a plasmid, the next step in cloning is transformation, introducing the plasmid into living bacterial cells so that it can be replicated. The two methods of bacterial transformation commonly used in the laboratory are heat shock transformation and electroporation. Both methods require competent cells, bacterial cells that can take up DNA. Not all cells are naturally competent. For example, some species, such *Bacillus subtilis*, can be easily transformed, but in other species, such as *Escherichia coli*, only a small number of cells in a culture may be able to take up DNA. Competent cells may be prepared in the laboratory or purchased commercially.

- Heat shock is the most easily accomplished transformation method, as it does not require any equipment other than a water bath. Plasmid DNA and heat-shock competent cells in calcium chloride are mixed together and incubated on ice for several minutes. Although the mechanism is not fully understood, calcium chloride causes DNA to bind to the bacterial cell wall. The cells are then subjected to a brief heat shock by incubation at 42°C for 30–45 seconds resulting in the uptake of DNA into the bacteria. Cells intended for heat shock transformation must be in the exponential growth phase to be highly competent.

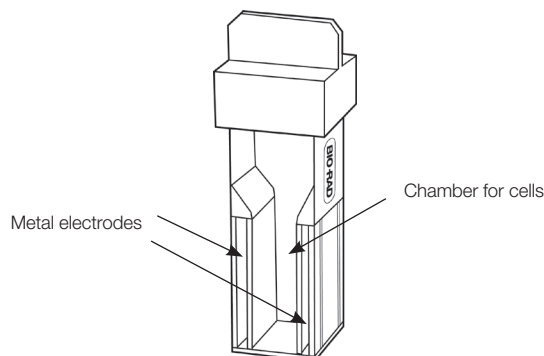


Bacterial growth curve. Bacterial growth follows a regular cycle with four phases. This is described in more detail in Chapter 7, Plasmid Purification Stage.

- Electroporation is also commonly used for transformation, and its mechanism of enabling DNA uptake is somewhat better understood than heat-shock transformation. When bacterial cells are subjected to a brief electrical shock, small pores in their cell walls open, allowing DNA to enter the cells. For electroporation, electrocompetent bacteria and plasmid DNA are mixed and placed in a special type of cuvette, a square test tube with metal electrodes on two sides (see figure). The cuvette is placed in an instrument called an electroporator that delivers an electrical charge of specific strength and duration to the cells. The electricity travels through the cells between the two electrodes, which is why electrocompetent cells must be prepared in a solution of very low ionic strength.

Cloning and Sequencing Explorer Series

For electroporation to be successful, the cells themselves must carry the current across the gap between the electrodes. If there are many ions (like Na^+) in the solution, the ions will carry the current instead of the cells, causing the cells to overheat and die.



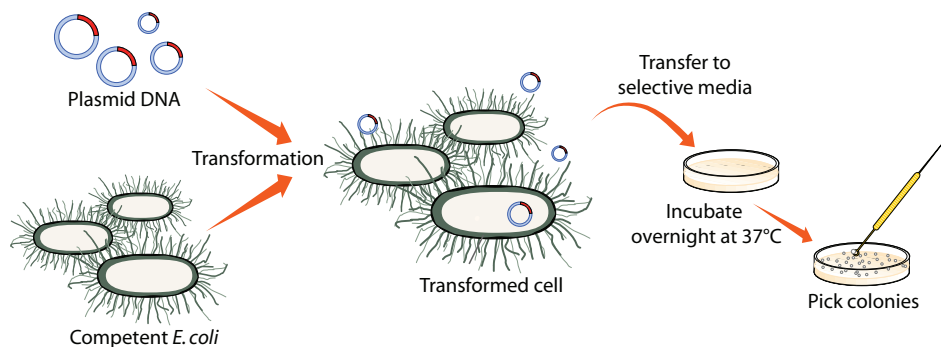
Bio-Rad electroporation cuvette.

There are ways to increase the number of competent cells in a bacterial culture. To prepare competent cells for heat shock transformation, the bacteria must be washed to remove the growth medium, then resuspended in a calcium chloride solution. For electroporation, the cells must be washed repeatedly in a chilled buffer and resuspended in a chilled sterile solution that has very low ionic strength. In both cases, the cells must be in solution and at a high concentration for transformation to be successful. The cells must also be kept cold at all times prior to transformation. The cells are extremely fragile at this stage and the cold keeps them inert. If they are warmed up in transformation solution, they will start to die. Even though the exact process of transformation is still not fully understood, it is thought that the cold temperature stabilizes the cell membranes of the bacteria and increases the interaction between the calcium cations and the negatively charged components in the plasma membrane. A sudden increase in temperature provided by the heat shock creates pores in the plasma, which allows for plasmid DNA to enter the bacterial cell.

Since bacteria have defense mechanisms that use restriction enzymes to degrade foreign DNA, only mutant strains that no longer have restriction activity can be used for transformation. Normal bacteria would degrade the plasmid DNA as soon as it enters the cell. Mutant strains for transformation are widely available and are used in this protocol.

What Happens after Transformation

Following either heat shock or electroporation, the cells are usually incubated in nutrient medium for up to 1 hour to allow them to recover from the stress of the transformation and begin to express the genes on the plasmid (such as an antibiotic resistance gene), although this step may be omitted and is not necessary in this protocol. The cells are plated on a selective medium for growth, usually agar plates containing nutrient medium and the antibiotic for which resistance is carried by the plasmid. For example, if the plasmid contains the *amp^r* gene, providing resistance to ampicillin, the agar plates should also contain ampicillin. This means that only bacteria that have been successfully transformed and now carry the plasmid will be able to survive and divide on the ampicillin-containing plates. The plasmid will replicate in the bacterial cells (using the host cell's replication machinery) and, as the bacteria divide, the plasmids will be passed on to their offspring. The plasmid that is used in this lab, pJET1.2, contains the *amp^r* gene and hence confers ampicillin resistance to any bacteria that are transformed.



Process of bacterial transformation. *E. coli* are made competent and are transformed with plasmid DNA. Only a few bacteria take up the plasmid DNA. The bacteria are plated on a selective medium and incubated overnight. Only bacteria that contain the plasmid will grow and form colonies. Bacterial colonies can then be picked and cultured for minipreps.

An additional level of selection can also be engineered into plasmids used for cloning or expression of genes. For cloning of genes, it is helpful to be able to distinguish colonies that have the gene of interest ligated into the plasmid from those that don't. In order to do this sort of selection, plasmids are designed so the cloning site for the insert is between a promoter and a selectable gene. In plasmids without an insert, the promoter will drive expression of this intact, selectable gene. But in plasmids containing the ligated insert, proper transcription of the selectable gene will be prevented. This process is called insertional inactivation. Some property of the selectable gene is then used to identify the bacteria in which the gene product is no longer made. Traditional "blue-white cloning" uses the *lacZ* gene, which produces an enzyme, β -galactosidase, that catabolizes X-gal (5-bromo-4-chloro-3-indolyl- β -d-galactopyranoside) into blue pigment. If the *lacZ* gene is disrupted by an insert, the bacteria turn white; if there is no insert, the bacteria remain blue. After transformation, bacterial plates should contain both blue and white colonies. The researcher then specifically picks the white "positive" colonies over the blue "negative" colonies.

pJET1.2 plasmid uses a different gene for selection — in this case a gene encoding a restriction enzyme, *Eco47I*. This enzyme is lethal to the bacteria. If the gene encoding this enzyme is not disrupted by an insert, the plasmid should express the gene and any bacteria containing religated plasmid should die. Because *Eco47I* is lethal to bacteria, a second level of control is engineered into the pJET1.2 plasmid to tightly control expression of the *Eco47I* gene. The gene's promoter in the pJET1.2 plasmid system is composed of the *lac* operon. When lactose or its analog IPTG (isopropyl β -D-1 thiogalactopyranoside) is not present, the *Eco47I* gene is not expressed. When IPTG is added, the *Eco47I* gene can be expressed. Because *Eco47I* is lethal to bacteria, cells that express this enzyme in culture will lyse and release the enzyme into solution, where it can kill other bacteria. Therefore, the IPTG is added to the plates, where any *Eco47I* that is expressed will be released locally into the agar, where it will die and therefore not impact growing colonies that contain the desired plasmids with an insert.

Even though the efficiency of bacterial transformation can be optimized using competent cells and determining the best experimental conditions for transformation and selection, transformation is still an inefficient process, with only a small percentage of DNA being taken into a small percentage of competent bacteria. After the transformed bacteria are plated on a selective medium, they will grow and divide on the plate, each forming a colony that is the product of a single transformation event. In other words, all the cells in each colony are clones, hence the origin of the term cloning.

Instructor's Advance Preparation

At this transformation stage, students will transform bacteria with their ligation reactions. Following transformation, pJET1.2 enables positive selection of plasmids with the desired insert due to the disruption of an otherwise lethal gene, *Eco47I*, which allows growth of successful transformants.

Bacterial transformation with ligation mixtures is a very inefficient process (much more inefficient than transformation with circularized plasmid DNA), so students should be encouraged to take special care during this protocol. There are many steps that, if performed improperly, can lead to reduced transformation efficiency or even to no colonies. For this reason, students should also transform the pGAP plasmid as a control to ensure that their transformations have worked and that they managed to make competent cells. Transformation with the pGAP plasmid also provides a backup in case ligations with the plant PCR products were either unsuccessful or too inefficient to produce transformed colonies. In this case, students can make minipreps of the pGAP plasmid and perform sequence analysis on the *Arabidopsis* GAPC gene. Transformation using the pGAP plasmid is expected to produce many more colonies than transformation using the plant gene ligation mixture, but a single positive colony is all that is needed to obtain a novel sequence.

The transformation protocol is a variation on traditional heat shock. It has been optimized to allow production of competent cells and transformation to be completed in less than an hour. Once made, the competent cells must be used immediately or discarded according to local regulations. They cannot be stored for later use. This transformation method permits relatively high transformation efficiency (10^6 transformants per μg of DNA) without requiring a refrigerated centrifuge, commercially purchased competent cells, or a -70°C freezer for storing the cells.

Note: Chemically competent and electrocompetent cells are available from commercial sources should your teaching goals include electroporation or more traditional chemical transformation techniques. The commercially available cells require storage at -70°C and have transformation efficiencies around 10^9 transformants per μg of DNA.

Safety Note: Transformation solution contains dimethyl sulfoxide (DMSO, CAS #67-68-5), an organic solvent. Handle with care and follow standard laboratory practices, including wearing eye protection, gloves, and a laboratory coat to avoid contact with eyes, skin, and clothing. If the solution comes into contact with gloves, change the gloves. DMSO passes directly through latex gloves, readily penetrates skin, and may result in the absorption of toxic materials and allergens dissolved in the solvent. After handling, wash hands and any areas that came into contact with the solution thoroughly. Refer to the Material Safety Data Sheet (MSDS) for complete safety information.

Transformation Checklist

Components from Cloning and Sequencing Explorer Series	Where Provided	(✓)
Sterile water	Ligation and Transformation Module	<input type="checkbox"/>
Control pGAP plasmid	GAPDH PCR Module	<input type="checkbox"/>
1.5 ml microcentrifuge tubes	Ligation and Transformation Module	<input type="checkbox"/>
15 ml culture tube	Microbial Culturing Module	<input type="checkbox"/>
C-Growth medium	Ligation and Transformation Module	<input type="checkbox"/>
Transformation reagent A	Ligation and Transformation Module	<input type="checkbox"/>
Transformation reagent B	Ligation and Transformation Module	<input type="checkbox"/>
IPTG	Ligation and Transformation Module	<input type="checkbox"/>
Sterile inoculating loops	Microbial Culturing Module	<input type="checkbox"/>
LB agar, ampicillin and petri dishes	Microbial Culturing Module	<input type="checkbox"/>
<i>E. coli</i> HB101 stock	Microbial Culturing Module	<input type="checkbox"/>
Required Accessories (Not Provided)	Quantity	(✓)
Incubator at 37°C	1	<input type="checkbox"/>
Shaking water bath or shaking incubator at 37°C	1	<input type="checkbox"/>
Microcentrifuge (refrigerated if available)	1	<input type="checkbox"/>
20 µl adjustable-volume micropipet and tips	1	<input type="checkbox"/>
200 µl adjustable-volume micropipets and tips	12	<input type="checkbox"/>
1,000 µl adjustable-volume micropipets and tips	12	<input type="checkbox"/>
Ice bath	12	<input type="checkbox"/>
Marking pens	12	<input type="checkbox"/>
(Optional) Vortex mixer	1	<input type="checkbox"/>
(Optional) Parafilm sealing film strips	36	<input type="checkbox"/>

Tasks to Perform Prior to the Lab

1. Prepare LB agar and LB amp IPTG agar plates: at least 3 days prior to the transformation, prepare one LB agar plate per team to be used to inoculate *E. coli* starter cultures. Prepare this according to standard protocols (see Appendix A2). Use remaining LB agar to make 2 LB amp IPTG agar plates per team.
2. Prepare *E. coli* starter plate: at least 2 days prior to the transformation, streak LB agar starter plate with bacteria. First, rehydrate the *E. coli* bacteria in the *E. coli* stock vial with 250 μ l of sterile water. Streak 10 μ l of the rehydrated bacteria on the LB plate using standard microbiology techniques to allow formation of single colonies (see Appendix A2). Incubate plate at 37°C overnight. Once colonies have grown, wrap plate in Parafilm sealing film and store at 4°C for up to 2 weeks.

Note: Any strain of *E. coli* bacteria normally used for transformation, such as DH5 or DH10B may be used in place of HB101.

3. Prepare LB broth: at least 2 days prior to the transformation, prepare at least 25 ml of LB broth per team according to standard procedures (see Appendix A2). Each team requires 5 ml of LB broth and 20 ml of LB amp broth.
4. Prepare starter culture: as late as possible the day before the transformation, inoculate a 2–5 ml LB culture with a starter colony from the *E. coli* starter plate (see Appendix A2). Incubate cultures with shaking (275 rpm) overnight at 37°C.

Note: It is important to use a fresh starter culture (<24 hours since inoculation) for the transformation, or transformation efficiency will be severely reduced.

If a shaking water bath or incubator is not available:

- One day prior to the transformation, incubate starter cultures at 37°C, shaking manually as frequently as possible to oxygenate the culture
- To prepare the competent cells, students should incubate the tubes at 37°C and shake them manually every 5 min during the 20–40 min growth phase to oxygenate the culture

Transformation efficiency is likely to be lower with either of these methods.

5. Put 2 LB amp IPTG plates per student team in the 37°C incubator just before lab.
6. (Optional) Just before lab, pipet 1.5 ml C-growth medium into one 15 ml culture tube per team and incubate at 37°C.

Protocol

Overview

During ligation, many different products are produced in addition to the desired ligation product with the PCR fragment inserted into the plasmid vector. For example, the vector may religate or the PCR product may ligate with itself. Relatively few of the DNA molecules formed during ligation are the desired combination of insert and plasmid vector. To separate the desired plasmid from other ligation products and to

have a way to propagate the plasmid, the ligation products are transformed into bacteria. Bacteria naturally contain plasmids, and plasmid vectors are plasmids that have been genetically modified to make them useful for molecular biologists. In this stage, you will transform competent bacteria with the products of the ligation reaction between your PCR product and the pJET1.2 plasmid vector. To transform bacteria with a plasmid, actively growing bacteria are pelleted, chilled, washed, and resuspended in transformation buffer to make them competent. It is vital to keep the bacteria on ice and to treat the competent bacteria very gently at all times — otherwise the transformation efficiency may be severely reduced and the protocol may result in no transformants. The competent bacteria are then mixed with the ligation reaction and incubated on ice to allow association of the DNA with the bacteria. The bacteria are then heat shocked to form pores in the cell membranes that allow transfer of the DNA inside the bacteria. In this protocol, which differs from traditional transformation protocols, the heat shock is performed by plating bacteria directly from ice onto warm agar plates at 37°C. The traditional method is to heat shock the bacteria by moving them from ice to a water bath at 37–42°C prior to adding growth medium, then plating after a recovery period.

Only bacteria expressing an ampicillin resistance gene will grow on LB plates prepared with ampicillin (remember that the pJET1.2 plasmid encodes an ampicillin-resistance gene). The plates are incubated at 37°C overnight to allow colonies to grow. To confirm that the bacteria were made competent, and to allow continuation of the experiment if the ligation fails, bacteria will also be transformed with a control plasmid (pGAP).

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- **Transform bacteria with the plasmid**
- Isolate plasmid from the bacteria
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene

Student Workstations

Each student team will require the following items to transform bacteria with one PCR product:

Material Needed for Each Workstation	Quantity
Ligation reaction from Chapter 5 (on ice)	5 µl
pGAP control plasmid (on ice)	1 µl
Previously prepared fresh starter cultures of <i>E. coli</i> at 37°C	1
15 ml culture tube (containing C-growth medium)	1
C-growth medium (in culture tube) prewarmed to 37°C	1.5 ml
LB amp IPTG plates prewarmed in 37°C incubator	2
Transformation reagent A (on ice)	250 µl
Transformation reagent B (on ice)	250 µl
1.5 ml microcentrifuge tubes	4
Sterile inoculating loops	2
20 µl adjustable-volume micropipet and tips	1
200 µl adjustable-volume micropipets and tips	1
1,000 µl adjustable-volume micropipets and tips	1
Marking pen	1
Ice bath	1
(Optional) Vacuum source	1

Common Workstation

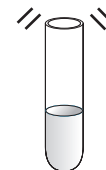
Material Required	Quantity
Shaking 37°C water bath or incubator	1
37°C incubator	1
Microcentrifuge (refrigerated, if available)	1
(Optional) Vortex mixer	1

Experimental Procedure

1. If not already done, pipet 1.5 ml C-growth medium into a 15 ml culture tube. Label tube with your initials and warm it at 37°C for at least 10 min.
2. Label 2 LB amp IPTG agar plates with your initials (on the bottom of the plate, not the lid). Also label one of the plates “pGAP” for the control plasmid and the other for your ligation (pJET + your plant name).
Place plates upside down at 37°C.

- Approximately 20–40 min prior to starting the transformation, pipet 150 μ l of fresh *E. coli* starter culture (inoculated one day prior) into the prewarmed C-growth medium and place in a 37°C incubator or water bath for 20–40 min shaking at 275 rpm.

C-growth medium is a nutrient broth that helps bacteria enter the growth phase efficiently.



Starter culture

Note: It is important to use an *E. coli* starter culture that is fresh (inoculated <24 hours before) to ensure sufficiently high transformation efficiency.

- Label a 1.5 microcentrifuge tube with your initials and “competent cells.”
- Prepare transformation buffer by combining 250 μ l of transformation reagent A and 250 μ l of transformation reagent B into a tube labeled “TF buffer” and mix thoroughly with a vortex mixer (if available). Keep on ice until use. (Note: This mixture must be prepared and used on the day of transformation.)

The transformation buffer contains calcium chloride and DMSO that assist the plasmid DNA to pass through the lipid cell membrane.

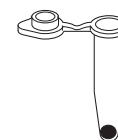
- After bacteria have grown in C-growth medium for 20–40 min at 37°C with shaking, transfer the entire culture to the tube labeled “competent cells” by decanting or pipetting. Do not put the actively growing cell culture on ice at this step.
- Centrifuge the bacterial culture in a microcentrifuge at top speed for 1 min. Accommodate tubes of classmates to ensure economic use of the microcentrifuge. Make sure that the microcentrifuge is balanced. Immediately put the pelleted bacterial culture on ice.



Note: After this step, it is very important to keep the bacteria on ice as much as possible during this procedure. Transformation efficiency will be severely compromised if the cells warm up.

It is very important to treat the bacteria extremely gently during this procedure — the bacteria are very fragile and your transformation efficiency will be compromised unless you are very gentle.

- Locate the pellet of bacteria at the bottom of the tube. Remove the culture supernatant, avoiding the pellet, using a 1,000 μ l pipet or a vacuum source. Keep the cells on ice.
- Pipet 300 μ l of ice-cold transformation buffer into the microcentrifuge tube containing the bacterial pellet. Resuspend the pellet by gently pipetting up and down in the solution above the pellet with a 1,000 μ l pipet, and gradually wear away the pellet from the bottom of the tube. Make sure that the bacteria are fully resuspended, with no clumps. Avoid removing the cells from the ice bucket for more than a few seconds.
- Incubate the resuspended bacteria on ice for 5 min.
- Centrifuge the bacteria in a microcentrifuge for 1 min, then place back in ice bucket immediately.



Note: Ensure that the bacteria are on ice immediately prior to and immediately following centrifugation. If the centrifuge is not close to the lab bench, take the entire ice bucket to the microcentrifuge so that the bacteria are only out of the ice bucket for 1 minute. Use a refrigerated microcentrifuge, if available.



- Remove the supernatant from the pellet using a 1,000 μ l pipet or vacuum source.

13. Pipet 120 μ l of ice-cold transformation buffer onto the pellet and resuspend by gently pipetting up and down with a 200 μ l pipet. Be sure that bacteria are fully resuspended with no clumps. Avoid removing the cells from the ice bucket for more than a few seconds.
14. Incubate the resuspended bacteria on ice for 5 min.

The cells are now competent for transformation.

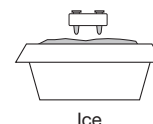
Note: Competent cells made using this protocol must be used on the day of preparation and cannot be stored at -70°C .

Experimental Procedure for Transformation

15. Label one microcentrifuge tube with your initials and “pGAP TF” (for pGAP transformation) and another microcentrifuge tube with your initials, plant name, and “TF” (referred to below as the “plant TF” tube).

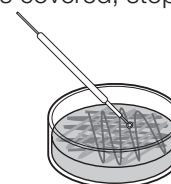
Note: If you are performing the ligation and transformation steps on the same day, use the microcentrifuge tube containing 5 μ l of the ligation prepared and labeled at the end of the ligation step.

16. Pipet 1 μ l of control pGAP plasmid into the microcentrifuge tube labeled “pGAP TF.”
17. If not already done, pipet 5 μ l of the ligation reaction from your ligation reaction tube into the “plant TF” microcentrifuge tube. Store any remaining ligation reaction at -20°C .
18. Using a fresh tip, pipet 50 μ l of competent bacteria directly into the ice-cold “pGAP TF” tube and gently pipet up and down 2 times to mix.
19. Using a fresh tip, pipet 50 μ l of competent bacteria directly into the ice-cold “plant TF” tube containing 5 μ l of your ligation, and gently pipet up and down 2 times to mix.
20. Incubate the transformations for 10 min on ice.
21. Retrieve the warm LB amp IPTG agar plates from the 37°C incubator.



Pipet the entire volume of each transformation onto the corresponding labeled LB amp IPTG plate and, using an inoculation loop or a sterile spreader, very gently spread the bacteria around the plate — remember that the bacteria are still very fragile! Once the plate is covered, stop spreading. Do not spread for more than 10 sec.

It is vital that the LB amp IPTG plates are warm at this step to ensure sufficiently high transformation efficiency. This is the heat shock for the transformation. Spreading the plate until it is dry will also reduce transformation efficiency.



22. Once the volume is absorbed in the agar, cover and place the LB amp IPTG plates upside down and incubate them overnight at 37°C .
23. The next day, analyze the results or wrap the plates in Parafilm and place them upside down at 4°C until required for inoculation of miniprep cultures (see Next Steps).

Next Steps

1. Before the next lab, transformed bacterial colonies need to be grown in liquid culture minipreps. Refer to instructions in Appendix A2 for additional details on growing minipreps.
 - a. Prepare 25 ml of LB amp broth.
 - b. Label four 15 ml culture tubes with your initials and “pJET,” the name of your plant, and #1 through #4.
 - c. Using sterile technique, pipet 3 ml of LB amp broth into each of the four 15 ml culture tubes.
 - d. One day prior to the next lab session, use a sterile loop or a sterile pipet tip to pick a single colony from the LB amp IPTG plate containing the plated bacteria transformed with your plant gene ligation reaction. Inoculate an LB amp culture tube with the colony. Repeat for a total of 4 miniprep cultures (one colony in each culture tube).

Note: It is not necessary to add IPTG to the liquid culture medium.

Note: Occasionally, satellite colonies may grow using this ligation method. Pick the large individual colonies, not the tiny satellite colonies surrounding larger colonies. Be sure that a single colony is picked, or you may isolate multiple plasmids from your miniprep. Multiple plasmids will result in mixed sequencing data that is not able to be deciphered.

- e. Place the miniprep cultures to grow overnight (18–28 hr) at 37°C in a shaking water bath or incubator set to a speed of 275 rpm.

Note: If no colonies grew on your team’s plate from the pJET1.2 + plant gene ligation reaction, either use colonies from another team’s successful transformation, if available, or inoculate your cultures with colonies from the pGAP control plate. Relabel your 15 ml culture tubes accordingly.

2. Prepare a 1% agarose gel and electrophoresis running buffer to analyze the plasmid miniprep restriction enzyme digestion.

Results Analysis of Ligation and Transformation

Count the number of bacterial colonies that grew on the LB amp IPTG agar plates. Occasionally, satellite colonies may grow using this ligation method. Count only the large individual colonies, not the tiny satellite colonies surrounding larger colonies.

Transformation	Number of Colonies
Control pGAP plasmid	
Plant gene ligation	

If the number of colonies is very high and uncountable, enter “TNC” for too numerous to count in the results table.

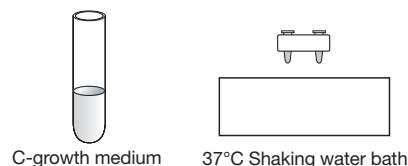
Focus Questions for Transformation

1. What is a vector?
2. What vector is used in this procedure?
3. How do you know which cells took in the desired recombinant plasmid?
4. What is a competent cell?

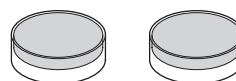
Transformation – Quick Guide

Preparation of Competent Cells

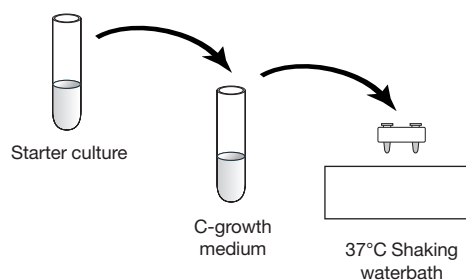
1. If not already done, pipet 1.5 ml C-growth medium into a 15 ml culture tube. Label with your initials and warm to 37°C for at least 10 min.



2. Label 2 LB amp IPTG plates with your initials on the bottom of the plate (not the lid). Also label one plate pGAP and the other for your ligation “pJET plus your plant name.” Place plates upside down at 37°C.

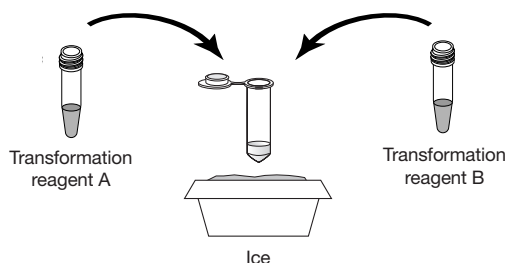


3. Pipet 150 µl of fresh starter culture (inoculated yesterday) into the pre-warmed C-growth medium and place in 37°C water bath shaking at ≥200–275 rpm for 20–40 min.

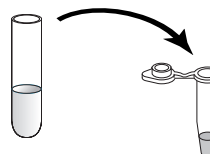


4. Label a 1.5 ml microcentrifuge tube with your initials and “competent cells.”

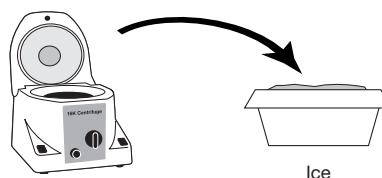
5. Prepare transformation buffer by combining 250 µl of transformation reagent A and 250 µl of transformation reagent B into a tube labeled “TF buffer.” Mix thoroughly and keep on ice.



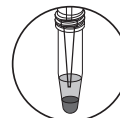
6. After 20–40 min incubation, transfer the actively growing culture in the C-growth medium from step 3 to your competent cells microcentrifuge tube from step 4.



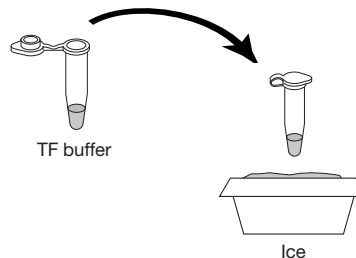
7. Centrifuge the bacteria at top speed for 1 min and immediately put the tube on ice.



8. Use a 1,000 μ l pipet or a vacuum source to remove culture supernatant, avoiding the pellet. Keep the cells on ice.

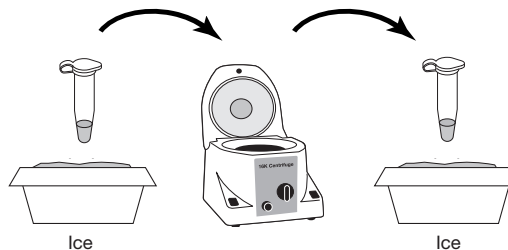


9. Resuspend the bacterial pellet with 300 μ l of ice-cold transformation buffer by very gently pipetting up and down in the solution above the pellet — do not touch the pellet.

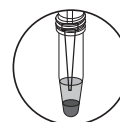


10. Incubate the resuspended bacteria on ice for 5 min.

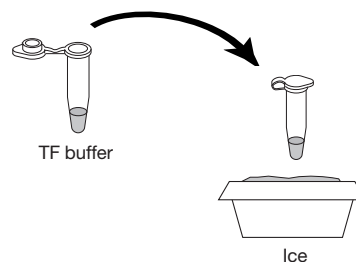
11. Centrifuge the bacteria at top speed for 1 min. Ensure the bacteria are on ice immediately prior to and immediately following centrifugation.



12. Using a 1,000 μ l pipet or a vacuum source, remove the supernatant, avoiding the bacterial pellet.



13. Very gently resuspend the bacterial pellet with 120 μ l of ice-cold transformation buffer. Keep cells on ice.

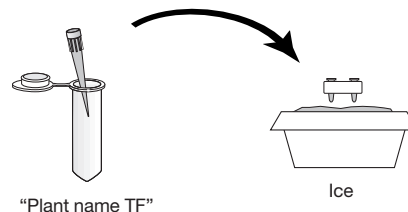


14. Incubate resuspended bacteria on ice for 5 min. The cells are now competent for transformation.

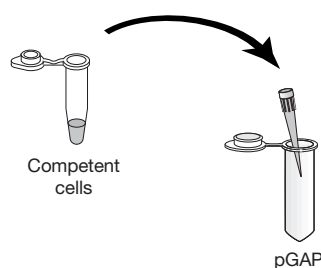
Experimental Procedure for Transformation

15. Label one microcentrifuge tube with your initials and “pGAP TF” for pGAP transformation and another tube with your initials, the plant name, and “TF” for transformation.

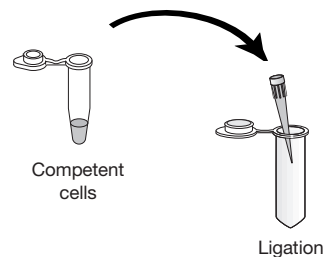
16. Pipet 1 μ l of control pGAP plasmid into the “pGAP TF” labeled tube. Keep on ice.



17. Pipet 5 μ l of your ligation reaction from your ligation reaction tube into the “plant name TF” microcentrifuge tube. Keep on ice.

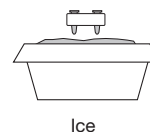


18. Using a fresh tip, very gently pipet the competent cells up and down 2 times, then pipet 50 μ l of competent cells into the “pGAP TF” tube and very gently pipet up and down 2 times to mix and return to ice.

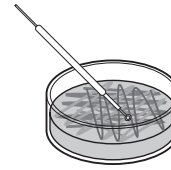


19. Using a fresh tip, pipet 50 μ l of competent cells into the “plant TF” tube containing 5 μ l of your ligation. Very gently pipet up and down 2 times to mix, and then return to ice.

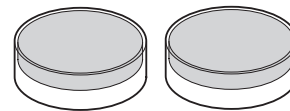
20. Incubate the transformations on ice for 10 min.



21. Retrieve the warm LP amp IPTG agar plates from the 37°C incubator. Using a fresh tip for each transformation, pipet the entire volume of each transformation onto the corresponding labeled agar plate. Use an inoculation loop to very gently spread the bacteria around the plate. Do not spread for more than 10 sec. Replace lid on plates.



22. When the liquid is absorbed in the agar, place LB amp IPTG agar plates upside down at 37°C and incubate overnight.



37°C incubator

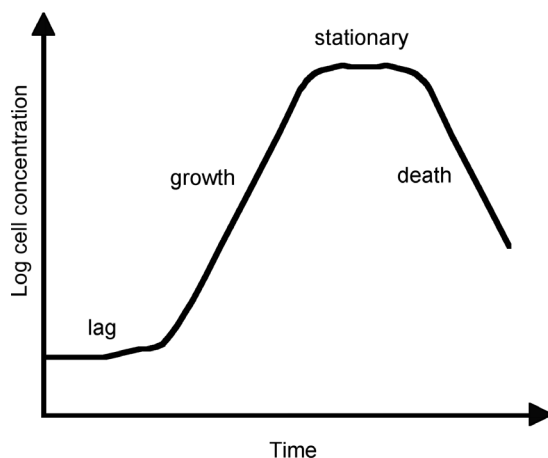
23. The next day, analyze results or wrap plates in Parafilm and place upside down at 4°C until needed.

CHAPTER 7: PLASMID PURIFICATION STAGE

Background**Minipreps of Plasmid DNA**

Once a plasmid has been introduced into competent bacterial cells and the cells have grown into colonies on a medium selective for cells containing the plasmid, the next step is to perform a miniprep of the plasmid DNA in preparation for DNA sequencing. The three steps of this task are: 1) growing cells in liquid culture; 2) purifying the plasmid DNA from the culture; and 3) performing a restriction digest on the purified DNA to determine whether the DNA inserted in the plasmid is the expected size. To start the liquid culture, cells from an isolated bacterial colony are placed in selective medium (nutrient broth with an antibiotic to which the plasmid provides resistance). This placing of the cells into the medium is called inoculation. It is important to choose a single isolated colony from the plate so that the liquid culture will contain cells that all have the same plasmid. If cells from more than one colony are used for inoculation, the miniprep may contain multiple plasmids, and the sequencing results from the mixed DNA samples will not be decipherable.

Growing the cells in liquid culture — for optimal growth, *Escherichia coli* need nutrients, oxygen, and warmth (37°C). The culture medium, Lysogeny Broth (LB), provides the nutrients and is composed of tryptone (peptides), yeast extract, and salt. The bacteria are oxygenated by shaking and this is vital to their growth; a speed of 275 rpm is optimal. If these requirements are not met, the culture will not reach its optimum bacterial density. Bacterial growth follows a regular growth cycle with four phases (see figure). After the culture medium is inoculated with the bacteria, there is a lag phase before the cell numbers begin to increase. After the lag phase, the cells begin an exponential growth phase. Rates of division depend on the growth conditions and the bacterial species. Growth rates of species vary greatly. For example, under optimal conditions, *E. coli*, the bacteria commonly used for cloning, divide every 17 minutes. In contrast, the bacteria that cause tuberculosis, *Mycobacterium tuberculosis*, divide approximately every 15 hours. When the nutrients in the growth medium are consumed, the culture enters a stationary phase during which the numbers of bacteria do not change. Finally, the culture enters the death phase, in which the number of cells decreases. Cells are normally harvested late in the growth phase, when the bacterial numbers are high and the cells are still healthy and dividing. For minipreps, *E. coli* cells are grown for 24 hr at 37°C with shaking at 275 rpm.



Purifying the plasmid DNA from the culture — the cells that have grown overnight are harvested by centrifugation and the plasmid DNA is purified. The purification is a multistep process designed to separate plasmid DNA from genomic DNA (gDNA) and other cellular components. This is a different process than gDNA extraction because the properties of gDNA and plasmid DNA are different. The steps include:

- Cell lysis and alkaline treatment of the lysate — the bacteria must be lysed to release the cell contents (including the plasmid DNA), but the treatment cannot be so harsh that all of the cell components break down. Bacteria normally live in a dilute aqueous environment, with much higher concentrations of solutes inside the cell than outside. Hence, lysing bacterial cells means disrupting or weakening both the cell wall and the cell membrane. The cell lysate is treated with an alkaline solution containing sodium hydroxide, the ionic detergent sodium dodecyl sulfate (SDS), and typically also ethylenediamine tetraacetate (EDTA). SDS helps disrupt the cell membranes and denatures proteins. EDTA, a chelating agent that removes magnesium ions, helps to destabilize the bacterial cell wall, which is a network of disaccharide (sugar) polymers crosslinked by short amino acid chains, and inactivates enzymes such as DNases. When the cell wall is weakened, the cell bursts

Under alkaline conditions (pH 12–12.5), the hydrogen bonds that link the strands of double-stranded DNA are broken, the double helix unwinds, and the two strands separate. Hence, the DNA is denatured, but the molecules remain in high molecular weight form as the strands themselves are intact. The hydrogen bonds connecting the strands of the smaller circular plasmid DNA molecules are also broken under alkaline conditions, but because plasmid DNA molecules are circular and supercoiled, the strands stay linked

- Neutralization — a high salt solution, such as 3–5 M potassium or sodium acetate, neutralizes the alkaline pH of the lysate. As the pH drops, the gDNA renatures and aggregates, and the high salt concentration precipitates the proteins. The cellular debris, including gDNA and proteins, is removed by centrifugation. The solution remaining after removal of the precipitated DNA and proteins, called the cleared or clarified lysate, contains the plasmid DNA. The plasmid DNA strands will also have renatured as the solution was neutralized, but since the molecules are so small, they will remain in solution
- Purification and concentration of plasmid DNA — the cleared lysate has a high salt concentration that would interfere with subsequent experiments, so the final step is usually purification to remove salts and other minor contaminants. This step has the added benefit of concentrating the DNA into a much smaller volume. Historically, alcohol precipitation has been used to concentrate the DNA, but many researchers now use commercial kits. Most kits use column chromatography to purify the plasmid DNA. For example, DNA binds strongly to silica in the presence of high concentrations of salts that disrupt hydrophobic interactions (chaotropic salts). Although not clearly understood, the binding is thought to be due to the exposure of phosphate residues on the DNA as a result of dehydration by the salts. The exposed phosphates adhere (adsorb) to the silica in the column. Contaminants, such as salts, will not bind to silica and can be washed from the column. The DNA can then be eluted from the column by reducing the salt concentration in the elution buffer. Note that although the columns for the gDNA extraction work in a similar way to the columns for plasmid DNA, each column has been optimized for one type of DNA and they are not interchangeable

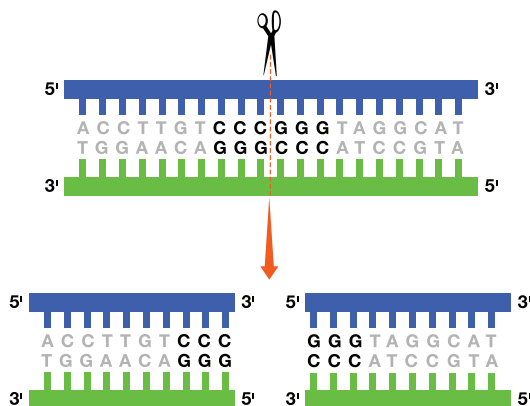
Restriction Digestion of Plasmid DNA

The properties of restriction enzymes — restriction enzymes are naturally occurring enzymes that digest, or cut, both strands of double-stranded DNA. They arose in bacteria as a defense mechanism against invading viruses called bacteriophages (or phages) that inject DNA into bacteria. Bacteria developed enzymes that recognize specific sequences within the phage DNA and then cut the DNA at that site, preventing the phage from destroying the bacteria. (Although the same DNA sequences may also be found in the bacteria's own genome, bacteria protect their own DNA by adding a methyl group to nucleotides in the sequence).

The enzymes used in cloning cut at an interior location in the DNA molecule, so they are called restriction endonucleases. (Enzymes that remove base pairs from the ends of DNA molecules are called exonucleases.) The enzymes that are used in molecular biology, called type II restriction endonucleases, recognize and bind to specific sequences of DNA and cut the DNA either within or very close to the recognition sequence. To date, more than 3,000 different restriction enzymes have been isolated and characterized, recognizing more than 250 different DNA sequences. Of these, more than 300 enzymes are commercially available. Over half of the enzymes that are commercially available have been cloned and are produced by overexpression in *E. coli* or other bacteria. Those restriction enzymes that have not been cloned must be purified from the organism in which they are found naturally, making commercial-level production time-consuming and expensive. The cost of a restriction enzyme may range from pennies to hundreds of dollars per experiment, depending on the difficulty in producing and purifying the enzyme.

Restriction enzymes are named after the bacterial species from which they were isolated. For example, EcoRI was isolated from *E. coli*, SmaI from *Serratia marcescens*, and Bgl II from *Bacillus globigii*. Each restriction enzyme recognizes a specific nucleotide sequence in the DNA (called a restriction site) and cuts the DNA molecule at only that sequence. Restriction sites range from 4 to 8 bp, but most are hexanucleotide sequences (6 bp). Recognition of the restriction sites is very specific. A single base pair difference from the enzyme's restriction site will prevent the restriction enzyme from binding and cutting the DNA.

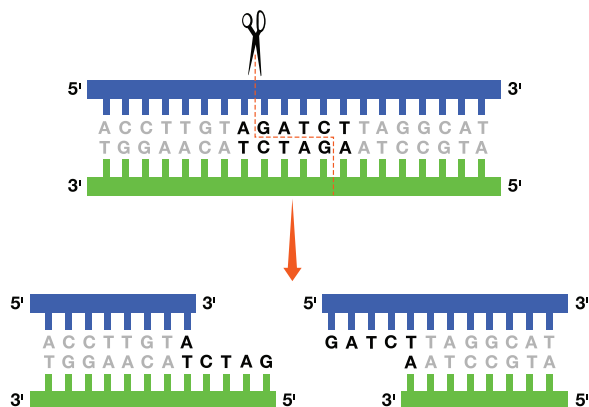
Restriction enzymes cut DNA in one of two ways. In the first, the restriction enzyme makes a cut through both strands between two adjacent base pairs, leaving blunt ends on either side of the cut. For example, the enzyme SmaI digests DNA to produce blunt ends.



Blunt end digestion by SmaI.

Cloning and Sequencing Explorer Series

Other restriction enzymes cut at staggered points on the two DNA strands, leaving a single-stranded overhang on each new DNA molecule. These overhangs are called sticky or cohesive ends, and they are very useful in cloning. A particular enzyme will always leave the same sticky ends. For example, the Bgl II enzyme produces sticky ends with a 5' overhang.



Sticky end digestion by Bgl II.

Each restriction enzyme has optimal conditions for efficient digestion of DNA. Conditions that must be controlled include reaction temperature, presence (or absence) of particular ions, ionic strength, and the pH of the buffer solution:

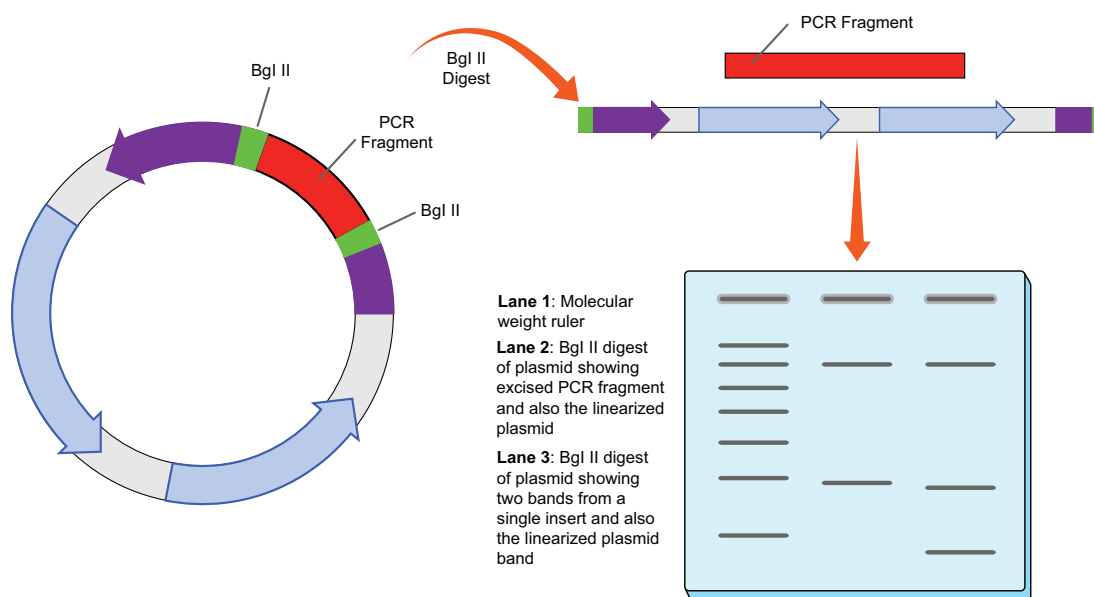
- The optimal temperature for each restriction enzyme reflects the optimal growth temperature for the bacteria in which it was found. Most restriction enzymes have maximum activity around 37°C, although others are more active at lower or higher temperatures; for example, optimal activity of *TaqI*, an enzyme discovered in the thermophilic bacterium *Thermus aquaticus*, occurs at 65°C
- Most restriction enzymes require magnesium ions (Mg^{2+}) for activity. Restriction enzyme buffers also often contain NaCl. The concentrations needed vary for each enzyme
- Some enzymes are sensitive to the presence of other ions, such as manganese ions (Mn^{2+}), and may have reduced or no activity if these ions are present in the buffer
- Most enzymes are most active at pH values slightly above neutral pH (~pH 7.2–8.0)

Companies that sell restriction enzymes provide a specification sheet for each enzyme, describing the optimal conditions needed for good enzyme activity. In fact, most companies provide an appropriate buffer with the enzyme. Problems can arise, however, when cutting DNA with more than one enzyme. Reaction conditions must be such that all enzymes will be active. For some very commonly used enzymes, companies have formulated buffers that are optimal for double digestion. Otherwise, a buffer can be used that may be compatible for both enzymes but not optimal for 100% activity. Alternatively, digests can be performed sequentially, first with one enzyme, followed by a purification to remove the reaction buffer before performing the second enzyme digest.

After DNA has been cut with restriction enzymes, it is often necessary to inactivate the restriction enzymes so they do not interfere with subsequent experiments. Most restriction enzymes can be heat-inactivated by incubation at 65°C for 20 minutes, but some enzymes must be incubated at 80°C for inactivation, while others cannot be heat-inactivated at all (for example, Bgl II).

Determining the size of an insert in a vector — before proceeding with further experiments using the purified plasmid DNA, it is important to verify that the isolated plasmid contains the insert of interest. This is done using restriction digestion followed by agarose gel electrophoresis. Looking back at earlier steps in the experiment, a gene or portion of a gene was ligated into the plasmid vector. From previous work, the size of this insert should be known. By digesting a small amount of the miniprep DNA with a specific restriction enzyme such as Bgl II (see below), the insert should be cut out of the vector. Running the products of the restriction digestion on an agarose gel should give two DNA bands, one the size of the vector and the other the size of the inserted DNA.

If there are more than two bands in any digest, it may mean that the insert itself contains a Bgl II site. Do the sizes of the excised bands add up to the size of the original PCR fragment? Alternatively, two similarly sized fragments may indicate that a mixed culture was used to start the miniprep, instead of an isolated colony.



Restriction enzyme digestion analysis of plasmid DNA. Circular plasmid DNA purified from bacterial minipreps is isolated and digested with Bgl II, a restriction enzyme producing at least two linearized DNA fragments: vector DNA and PCR fragment (lane 2). These fragments can be visualized using agarose gel electrophoresis. If more than two bands appear, this may indicate that the PCR fragment contains a Bgl II restriction site (lane 3). In this case, the sizes of the smaller bands should add up to the size of the entire PCR. A 500 bp molecular weight ruler is shown in lane 1.

Circular plasmid DNA has different properties than linear DNA because it is supercoiled. This means that the DNA circles are wound up very tightly and are very compacted — imagine winding string very tightly. This compacted DNA moves through an agarose gel faster than linear DNA such that supercoiled plasmids can appear much smaller than plasmids cut by restriction enzymes. Occasionally, plasmid DNA can be nicked. Nicked DNA is still circular but one strand of the double helix has been cut. This nick releases the tension and the DNA is no longer supercoiled. Nicked DNA runs through an agarose gel more like linear DNA.

Instructor's Advance Preparation

In this plasmid purification stage, the students will isolate plasmid DNA from transformed bacteria. Prior to the lab, the students (or the instructor) need to pick colonies from the transformation plates to inoculate four minipreps per team in LB amp broth. The miniprep method is an alkaline lysis protocol, with the purification step performed on a silica column. After plasmids are isolated, they will be analyzed by restriction enzyme digestion and agarose gel electrophoresis to determine whether they contain the PCR products.

Plasmid Purification Checklist

Components from Cloning and Sequencing Explorer Series	Where Provided	(✓)
Capless collection tubes, 2 ml	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Microcentrifuge tubes, 1.5 ml	Cloning and Sequencing Series	<input type="checkbox"/>
Aurum plasmid columns, purple*	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Aurum resuspension solution	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Aurum lysis solution	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Aurum neutralization solution	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Aurum wash solution, 5x concentrate	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Aurum elution solution	Aurum Plasmid Mini Purification Module	<input type="checkbox"/>
Bgl II restriction enzyme	Ligation and Transformation Module	<input type="checkbox"/>
10x Bgl II reaction buffer	Ligation and Transformation Module	<input type="checkbox"/>
Sterile water	Ligation and Transformation Module	<input type="checkbox"/>
1% agarose gel	Previously prepared (Electrophoresis Module)	<input type="checkbox"/>
Electrophoresis running buffer	Previously prepared (Electrophoresis Module)	<input type="checkbox"/>
UView 6x loading dye and stain	GAPDH PCR Module	<input type="checkbox"/>
500 bp molecular weight ruler	GAPDH PCR Module	<input type="checkbox"/>

* The mini columns in both the Nucleic Acid Extraction module and the Aurum Plasmid Mini Purification Module are purple but they are functionally different. Be sure to keep them in their separate labeled bags so they do not get mixed or confused.

Required Accessories (Not Provided in Kit)	Quantity	(✓)
95–100% ethanol	100 ml	<input type="checkbox"/>
Microcentrifuge with variable speed setting $\geq 12,000 \times g$	2	<input type="checkbox"/>
(Optional) Vortex mixer	1	<input type="checkbox"/>
Water bath, incubator, or heating block at 37°C	1	<input type="checkbox"/>
Ice bath	12	<input type="checkbox"/>
1,000 μ l adjustable-volume micropipet and tips	12	<input type="checkbox"/>
200 μ l adjustable-volume micropipet and tips	12	<input type="checkbox"/>
10 μ l adjustable-volume micropipet and tips	12	<input type="checkbox"/>
Horizontal electrophoresis chambers and power supplies	12	<input type="checkbox"/>

Tasks to Perform Prior to the Lab

1. Add 100 ml of 95–100% ethanol to the Aurum wash solution and mix well.
2. Ensure that the day prior to the lab, LB amp miniprep cultures have been inoculated by students (or instructor) and incubated at 37°C with shaking at 275 rpm (see Next Steps section of Transformation chapter and Appendix A2).
3. Thaw the Bgl II restriction enzyme and 10x Bgl II reaction buffer on ice. Just before use, mix and centrifuge the reagents to collect contents at the bottom of the tubes. Do not aliquot the restriction digestion reagents for students; the volumes required are too small. Also ensure students are familiar with methods to pipet small volumes.
4. Prior to the lab, students or instructor should prepare 1% agarose gels and electrophoresis running buffer.

Protocol

Overview

In this stage, you will analyze plasmids that have been successfully propagated to verify that they have the PCR product inserted. Purifying the plasmid from a small culture of transformed bacteria (a miniprep) produces a sufficient amount of plasmid DNA. The plasmid is analyzed by restriction enzyme digestion to assess the size of the fragment inserted and to compare it to the size of the PCR product ligated into the plasmid.

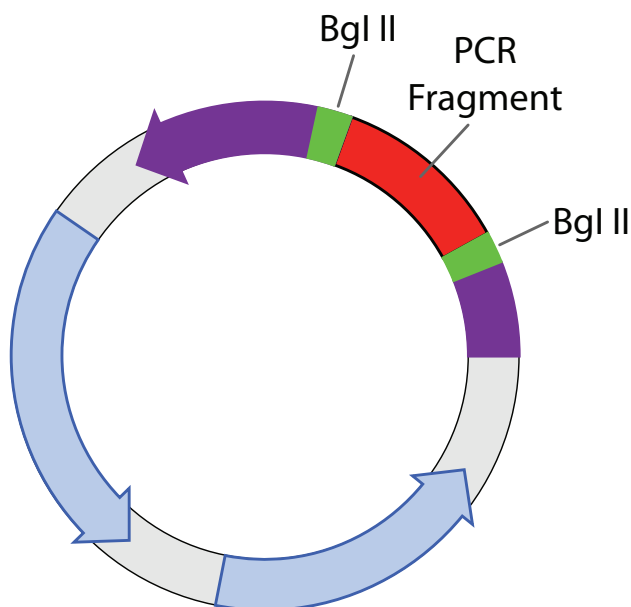
The blunted PCR product was inserted into the plasmid vector pJET1.2 (see figure). pJET1.2 contains a Bgl II restriction enzyme recognition site on either side of the insertion site. Thus, once the plasmid DNA has been isolated, digestion with this restriction enzyme and subsequent electrophoresis will allow determination of the size of any fragment inserted into the vector. Transformed bacteria containing plasmids should have been grown to saturation in LB amp medium prior to this lab — see Next Steps at the end of the Protocol in the Transformation chapter.

Transformed bacteria containing plasmids should be grown to late growth phase in LB amp broth prior to this lab — see Next Steps at the end of the protocol in the Transformation chapter.

Safety Note: The lysis solution contains sodium hydroxide, which causes burns. Handle with care and follow standard laboratory practices, including wearing eye protection, gloves, and a laboratory coat to avoid contact with eyes, skin, and clothing. Please refer to the Material Safety Data Sheet for complete safety information.

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- **Isolate plasmid from the bacteria**
- Sequence DNA
- Perform bioinformatics analysis of the cloned gene



Novel pJET1.2 plasmids containing PCR fragment in between two Bgl II sites.

Student Workstations

Each student team requires the following items to purify and digest 4 plasmids*:

Material Needed for Each Workstation	Quantity
Miniprep cultures	4
Microcentrifuge tubes, 2 ml	17
Capless collection tubes, 2 ml	4
Aurum plasmid mini columns, purple	4
Aurum resuspension solution	1 ml
Aurum lysis solution	1 ml
Aurum neutralization solution	1.5 ml
Aurum wash solution (ensure that ethanol has been added to wash solution)	3 ml
Aurum elution solution	500 µl
Bgl II restriction enzyme*	4 µl
10x Bgl II reaction buffer*	8 µl
Sterile water	30 µl
1% agarose gel, electrophoresis running buffer, horizontal electrophoresis chambers, and power supply	1
UView loading dye and stain, 6x	45 µl
500 bp molecular weight ruler (ensure loading dye has been added to MWR)	10 µl
1,000 µl adjustable-volume micropipet and tips	1
200 µl adjustable-volume micropipet and tips	1
10 µl adjustable-volume micropipet and tips	1

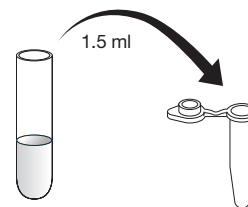
Common Workstation

Material Required	Quantity
Microcentrifuge	2
(Optional) Vortex mixer	1
Water bath, incubator, or heating block at 37°C	1
Ice bath containing stock tubes of:	
• Bgl II restriction enzyme	
• 10x Bgl II reaction buffer	

* These reagents must be stored on ice. It may be easier to keep them on ice at a common workstation rather than in a separate ice bucket at each student workstation. See common workstation.

Experimental Procedure for Plasmid Purification

1. Label one capless collection tube, one plasmid mini column, and two microcentrifuge tubes for each miniprep culture with your initials and the information on the labels of the 1.5 ml cultures. On the microcentrifuge tubes, include "miniprep DNA" in the name.
2. Place each column in the appropriate capless collection tube.
3. Transfer 1.5 ml of each miniprep culture into one of the appropriately labeled microcentrifuge tubes by pipetting or decanting.
4. Centrifuge the microcentrifuge tubes for 1 min at top speed ($>12,000 \times g$) to pellet the bacteria. Accomodate tubes of classmates to ensure economic use of the microcentrifuge. Make sure that the microcentrifuge is balanced.



5. Locate the bacterial pellet and remove the supernatant from each tube using a 1,000 μ l pipet or a vacuum source; avoid touching or extracting the pellet. Dispose of the supernatant according to your local environmental health and safety regulations.
6. Add the remaining 1.5 ml of the appropriate miniprep culture to the correct microcentrifuge tube containing the bacterial pellet, centrifuge for 1 min and remove supernatant.
7. Add 250 μ l of resuspension solution to each tube. Resuspend bacterial pellet by pipetting up and down or vortexing. Use a fresh tip each time. Ensure that no clumps of bacteria remain.
8. Pipet 250 μ l of lysis solution into each tube and mix by gently inverting 6–8 times. Do not pipet or vortex this lysate or you risk shearing (fragmenting) the bacterial gDNA, which could contaminate your plasmid preparation.

This step chemically breaks the cells open.

9. Within 5 min of adding lysis solution, pipet **350 μ l** of neutralization solution into each tube and mix by gently inverting 6–8 times. Do not pipet or vortex this lysate. A white precipitate should form.

This step precipitates out the gDNA, proteins, and cellular debris.

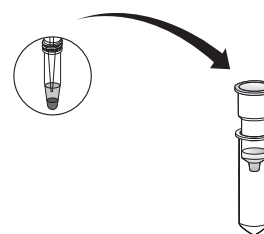
10. Centrifuge the tubes for 5 min at top speed. Make sure the microcentrifuge is balanced.

This step purifies the plasmid DNA from the bacterial lysate.

11. Decant or pipet supernatant from the centrifuged tubes onto the appropriately labeled column. Avoid transferring any precipitate. If necessary, recentrifuge the microcentrifuge tubes. Discard microcentrifuge tubes.

This step binds plasmid DNA to column matrix.

12. Centrifuge the columns, still in the capless collection tubes, in the microcentrifuge for 1 min at top speed.



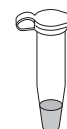
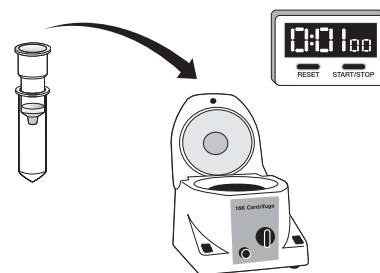
13. Discard flowthrough from the collection tube and replace the column in the collection tube.
14. Pipet 750 μ l of wash solution onto each column (ensure that the wash solution has had ethanol added to it prior to use).
15. Centrifuge columns in the capless collection tubes for 1 min at top speed.
16. Discard the flowthrough from the collection tube.
17. Replace columns into collection tubes and centrifuge for an additional 1 min to dry out the column.

It is vital that this step be performed, or wash solution and ethanol may contaminate the purified plasmid.

18. Transfer contents of each column to the appropriately labeled capped “miniprep DNA” microcentrifuge tube and pipet 100 μ l of elution solution onto the column.
19. Let the elution solution be absorbed into the column for 1–2 min.
20. Place the column in the microcentrifuge tube into the centrifuge. It is best to orient the cap of the microcentrifuge tube downward, toward the center of the rotor, to minimize friction and damage to the cap during centrifugation.
21. Centrifuge the columns for 2 min.

This step elutes the purified plasmid DNA.

22. Discard the columns and cap the tubes containing the eluted sample.
23. Store miniprep plasmid DNA at 4°C short term (up to 1 month) or –20°C long term.

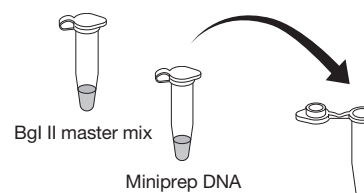


Experimental Procedure for Restriction Digestion Analysis

24. Label one microcentrifuge tube with your initials and “Bgl II master mix.”
25. Prepare a 2x master mix for Bgl II restriction digestion reactions according to the following table, using stock reagents from the common workstation. Use a fresh tip for each reagent.

Reagent	Volume for 1 Reaction	Volume for 4 Reactions
10x Bgl II reaction buffer	2 μ l	8 μ l
Sterile water	7 μ l	28 μ l
Bgl II enzyme	1 μ l	4 μ l
Total	10 μ l	40 μ l

26. Label a microcentrifuge tube for each plasmid miniprep.
27. Prepare digestion reactions by combining 10 μ l of the Bgl II master mix and 10 μ l of each plasmid DNA in the appropriately labeled microcentrifuge tubes.



28. Mix tube components and spin briefly in a microcentrifuge to collect the contents at the bottom of the tube.
29. Incubate reactions at 37°C for 1 hr. If the reactions will not be analyzed by agarose gel electrophoresis immediately, store them at –20°C until analysis.

Experimental Procedure for Electrophoresis

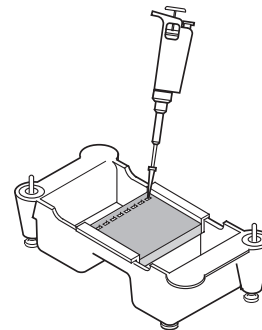
30. Plan your gel electrophoresis experiment.

Lane Number	Sample	Sample Volume (μl)	Resulting Band Sizes (bp)
1	500 bp molecular weight ruler	10 μl	500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000

(Optional) It is recommended that undigested DNA samples are loaded and run next to the corresponding digested samples.

31. If you are running undigested DNA, combine 5 μl of miniprep DNA with 15 μl of sterile water.
32. Briefly centrifuge samples to force the contents to the bottom of the tube.
33. Add 3.5 μl of 6x loading dye and stain to each of the digested and undigested samples.

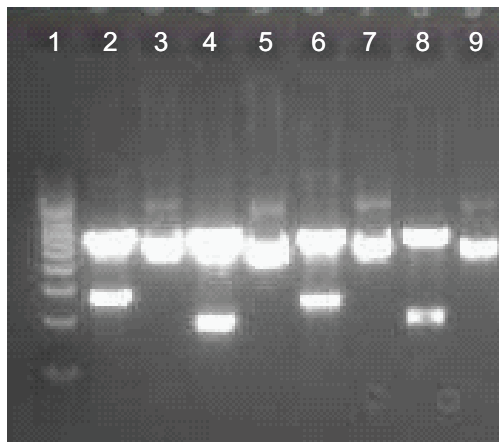
34. Put a 1% agarose gel in the electrophoresis chamber and add electrophoresis running buffer to just cover the gel.
35. Load 20 μl of each sample and 10 μl of the 500 bp molecular weight ruler according to your plan on the 1% agarose gel.
36. Connect the electrophoresis chamber to the power supply and turn on the power. Run the gel at 100 V for 30 min unless otherwise instructed.
37. Visualize your bands (using appropriate safety equipment) and acquire an image of your gel. Paste image below and label it accordingly.



Results Analysis from the Restriction Digest

Construct a table of your results for each sample that describes the sample's name, the reaction conditions, what bands are present, the size of the bands, and the relative intensity of the bands.

Note: Undigested plasmid DNA is usually supercoiled, meaning that the DNA runs through the gel faster than unsupercoiled (digested or nicked) DNA because it is compacted. Thus, it is expected that the undigested plasmid samples will appear to be significantly smaller than the digested plasmids.



Example of miniprep digestion. Four miniprep clones of pJET1.2 GAPDH plasmids derived from a single ligation with maize *GAPDH* PCR product were digested with Bgl II. Digests and undigested DNA were electrophoresed on a 1% TAE agarose gel. Lane 1 = 500 bp molecular weight rule; lanes 2, 4, 6, and 8 = minipreps digested with Bgl II enzyme; lanes 3, 5, 7, and 9 = undigested minipreps. **Note:** Different sizes of inserts suggests different *GAPDH* genes were cloned in this ligation.

CHAPTER 7 PROTOCOL

1. Did your ligation generate plasmids containing the PCR product? Refer to your results from the electrophoresis stage to compare band sizes from the PCR product and Bgl II digested miniprep fragments.
2. What size is the pJET1.2 plasmid?
3. Do you have any digests with more than two bands? What could cause this? What do the sizes of the bands add up to?
4. Do you have any digests with no insert or with an insert that does not correspond to your PCR product? What could cause this?
5. Which plasmids contain fragments that you suspect are *GAPC* gene fragments? These are the ones that you will go on to sequence.

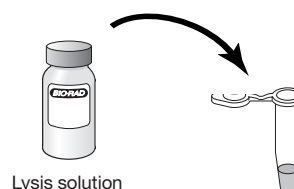
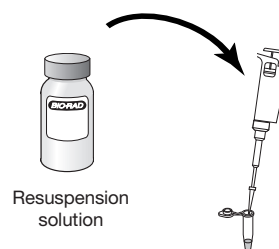
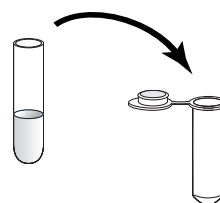
Focus Questions for Plasmid Purification

1. What forms of DNA are in transformed bacteria?
2. How is the plasmid purified in a miniprep?
3. How can you verify that the plasmid is the recombinant you wished to create?
4. What is supercoiled DNA and how does supercoiling affect the migration of DNA on an agarose gel?

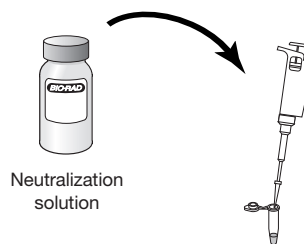
Plasmid Purification – Quick Guide

Plasmid Miniprep Purification

1. Label one capless collection tube, one column, and two microfuge tubes for each miniprep culture with your initials and the same label as each of your miniprep cultures.
2. Place each column in the appropriate capless collection tube.
3. Transfer 1.5 ml of each miniprep culture into the appropriately labeled microcentrifuge tube by pipetting or decanting.
4. Spin microcentrifuge tubes for 1 min at top speed to pellet the bacteria.
5. Locate the bacterial pellet and remove supernatant from each tube using a 1,000 μ l pipet or a vacuum source, avoiding the pellet. Discard the supernatant.
6. Add the remaining 1.5 ml of the appropriate miniprep culture to the correct microcentrifuge tube containing the bacterial pellet, centrifuge for 1 min and remove supernatant.
7. Using a fresh pipet tip, pipet 250 μ l of resuspension solution into each tube. Resuspend bacterial pellet by pipeting up and down or vortexing. Ensure no clumps of bacteria remain.
8. Pipet 250 μ l of lysis solution into each tube and mix by gently inverting 6–8 times. DO NOT pipet or vortex this lysate.



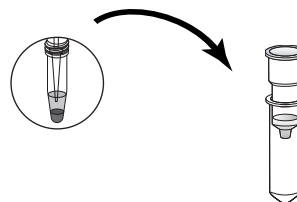
9. Within 5 min of adding lysis buffer, pipet **350 μ l** of neutralization solution into each tube and mix by gently inverting 6–8 times. A white precipitate should form.



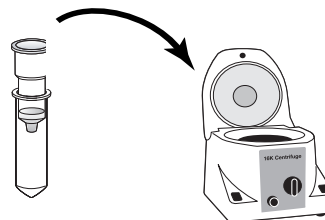
10. Spin the tubes 5 min at top speed in the microcentrifuge.



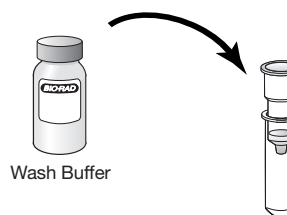
11. Decant or pipet supernatant from the centrifuged tubes onto the appropriately labeled column. Avoid transferring any precipitate. If necessary, recentrifuge the microcentrifuge tubes.



12. Spin columns in the capless tubes in the microcentrifuge for 1 min at top speed.

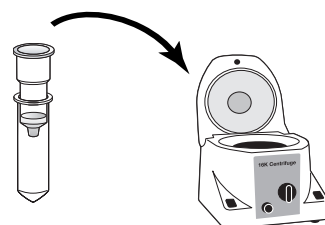


13. Discard flow-through from the collection tube and replace column in collection tube.



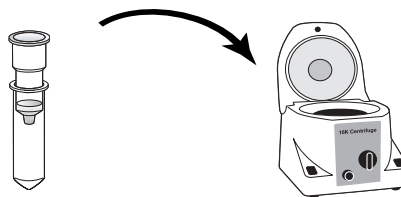
14. Ensure the wash buffer has had ethanol added to it prior to use. Pipet 750 μ l of wash buffer onto each column.

15. Spin columns in the capless collection tubes in microcentrifuge for 1 min at top speed.

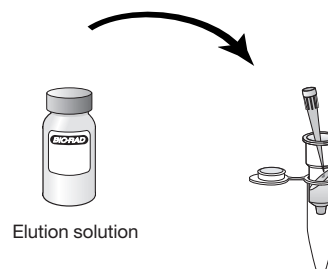


16. Discard flowthrough from the collection tube.

17. Replace the column in the collection tube and spin for another 1 min to dry out the column.



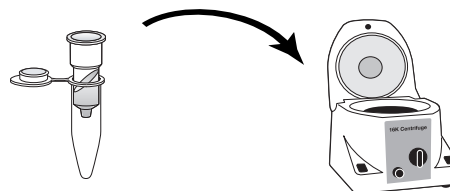
18. Transfer each column to a fresh, appropriately labeled capped “miniprep DNA” microcentrifuge tube and pipet 100 μ l of elution solution onto the column.



19. Let the elution solution absorb into the columns for 1–2 min.



20. Place the columns in their microcentrifuge tubes into the centrifuge.



21. Spin the columns for 2 min at top speed.

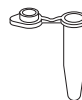
22. Discard columns and cap the tubes.



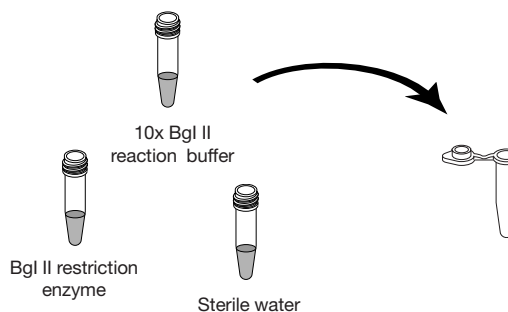
23. Store miniprep plasmid DNA at 4°C short term (up to 1 month) and -20°C long term.

Restriction Digest Analysis

24. Label one microcentrifuge tube with your initials and "Bgl II master mix"

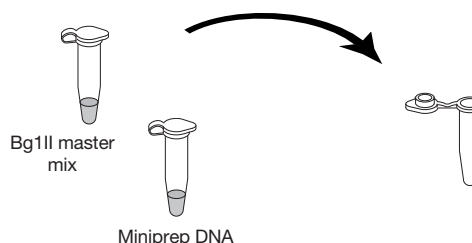


25. Prepare a master mix for Bgl II restriction enzyme digestion. Pipet 8 µl of 10x Bgl II reaction buffer, 28 µl of sterile water and 4 µl of Bgl II enzyme into the tube and mix by pipetting up and down



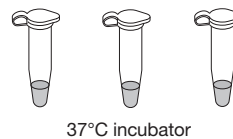
26. Label a microcentrifuge tube for each plasmid prep with your initials, "Bgl II" and the name of the plasmid prep.

27. Pipet 10 µl of Bgl II master mix into the digestion tubes. Using a fresh tip for each miniprep DNA, pipet 10 µl of each miniprep DNA into the appropriately labeled digestion tube.



28. Close the cap and mix well. Centrifuge briefly to collect the contents at the bottom of the tube.

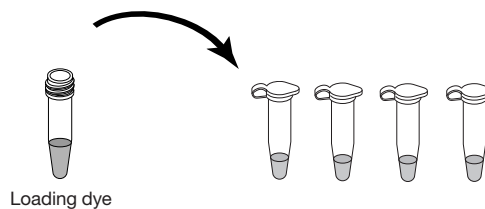
29. Incubate reactions at 37°C for 1 hr. If they will not be analyzed by agarose gel electrophoresis immediately, store reactions -20°C until the next laboratory session.



Gel electrophoresis of Restriction Digestion Reactions

30. If you are running undigested DNA samples on the gel, combine 5 μ l of undigested miniprep DNA with 15 μ l of sterile water. Pipet up and down to mix.

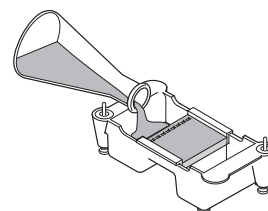
31. Using a fresh tip for each, add 3.5 μ l of 6x loading dye and stain to the restriction digestion samples. Pipet up and down to mix.



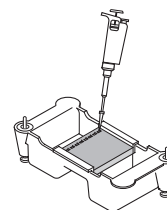
32. Briefly spin all samples in a microcentrifuge to collect contents to the bottom of the tube.



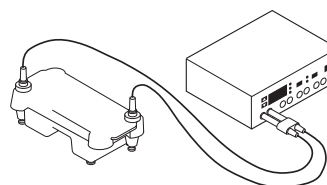
33. Place a 1% agarose gel in the electrophoresis chamber. Pour electrophoresis buffer into the chamber until it just covers the gel by 1–2 mm.



34. Load 10 μ l of the 500 bp molecular weight ruler and 20 μ l of each sample into the gel.



35. Connect the electrophoresis chamber to the power supply and turn on the power. Run the gel at 100 V for 30 min unless otherwise instructed.



36. On completion of electrophoresis, visualize your DNA according to your instructor's directions.

CHAPTER 8: DNA SEQUENCING

Background

The Development of DNA Sequencing

Sequencing DNA means determining the exact order of nucleotide bases (guanine (G), adenine (A), thymine (T), and cytosine (C)) in a DNA molecule. DNA sequencing began in the 1970s when two research groups developed different methods for sequencing, the Maxam-Gilbert method and the Sanger method, at almost at the same time. Although we take DNA sequencing for granted now, when researchers started sequencing, it was a laborious process requiring the use of hazardous chemicals. After days of work, the results were relatively short sequences.

Today, most researchers send their samples to core laboratory facilities where, for a fee, their DNA is sequenced for them using an automated sequencer. The researchers receive the sequence data in a day or two. To date, researchers have sequenced the complete genomes of almost 700 organisms. Most of the completed genomes are from bacteria, but other organisms with completed genome sequences include several yeasts, fungi, plants, fruit flies, mosquitoes, zebrafish, and mammals such as the mouse, rat, opossum, chimpanzee, and dog (a boxer named Tasha). Many more genomes are currently undergoing sequencing.

Maxam-Gilbert Sequencing Method

Allan Maxam and Walter Gilbert, working in the U.S., developed a chemical degradation method for DNA sequencing. The steps in sequencing by chemical degradation are:

1. Label the 5' end of the DNA to be sequenced with a radioisotope tag. (The labeling can also be performed at the 3' end, but the end selected for labeling must be consistent.)
2. Divide the labeled DNA into four test tubes, each containing different chemicals that cleave the DNA strands after a particular base.
3. Use polyacrylamide gel electrophoresis to separate the radioactive fragments by size. The four separate reactions are placed in adjacent lanes in the gel.
4. Place the radioactive gel against X-ray film (a technique called autoradiography). When developed, the X-ray film will have dark bands corresponding to the radioactive bands on the gel, while the cleaved ends cannot be seen since they do not contain the radioactive label.
5. Derive the DNA sequence from the X-ray film image of the gel. The shortest fragments are at the bottom of the gel and the largest at the top. DNA fragments that are adjacent in size are one base different in size.

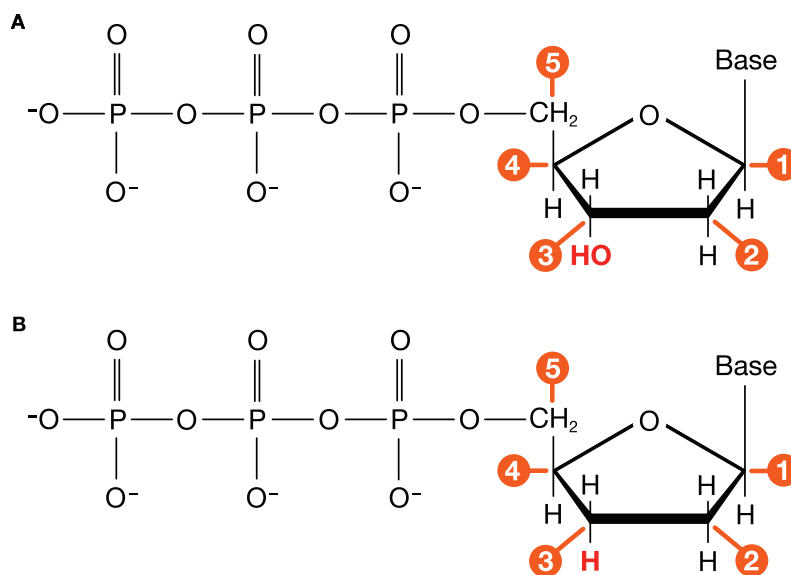
The reaction is relatively inefficient and cleaves the DNA at a small percentage of the occurrences of the bases, not at every single one. This produces “a nested set of radioactive fragments extending from the labeled end to each of the positions of that base” (Maxam and Gilbert 1977).

The Maxam-Gilbert method could sequence about 100 bases into a DNA fragment, but it used very hazardous chemicals. Another negative was that the method was not easy to automate as technologies improved.

Sanger Sequencing Method

In Europe Sanger and Coulson developed the chain termination method for DNA sequencing or, as they called it, the “plus and minus” method (Sanger et al., 1977). Since the mid-1980s chain termination has been the predominant method used for sequencing, in large part because the technique could be automated. (Frederick Sanger received a Nobel Prize for his work.) The steps in chain termination sequencing are:

1. Prepare a single-stranded template of the DNA to be sequenced.
2. As in Maxam-Gilbert sequencing, divide the DNA into four test tubes and add:
 - DNA primer that will start DNA synthesis at the area to be sequenced. Sequencing primers, like primers for PCR, must be specifically designed for each specific sequencing reaction. Luckily when sequencing DNA cloned into plasmids, the plasmid sequence is known and primer sequences known to anneal to the plasmid are usually available and can sequence from each side of the plasmid cloning site. However, if the cloned fragment is long primers may need to be designed to bind internally to the cloned fragment itself to ensure generation of the complete sequence. This can be challenging when the sequence is unknown
 - DNA polymerase
 - Labeled nucleotides — these are deoxynucleotide triphosphates (dNTPs: dGTP, dATP, dTTP, and dCTP) and they are always in excess in the reaction. In early sequencing the dNTPs were labeled with a radioisotope tag, but now they are usually labeled with one or more fluorescent tags
3. Add a modified nucleotide called a dideoxynucleotide (ddNTP) to each reaction tube. (This sequencing method is also sometimes called dideoxy sequencing because of the use of ddNTPs.) ddNTPs lack the 3'-hydroxyl group needed for elongation of the DNA molecule (see figure). Each reaction tube gets a different ddNTP, either ddGTP, ddATP, ddTTP, or ddCTP.

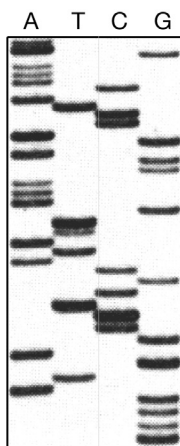


Structures of nucleotide triphosphates (NTPs) used in chain termination sequencing. A) dNTPs have a 3'-hydroxyl (–OH) group (at position 3), necessary for elongation of a DNA molecule as the 3'-hydroxyl forms a phosphodiester bond with the 5'-phosphate group on the next nucleotide; **B)** ddNTPs do not have a 3'-hydroxyl group. The position has been modified so there is a hydrogen (–H) at that position. Therefore, when a ddNTP is incorporated into a DNA molecule the synthesis will end at that nucleotide. In other words, the DNA chain will terminate.

4. Allow DNA synthesis to proceed in each reaction tube. During synthesis, almost all of the nucleotides that are incorporated into the new DNA strand are labeled dNTPs as dNTPs are in excess. However, when a dideoxynucleotide is incorporated, DNA synthesis will stop on that strand as there is no 3'-hydroxyl to form the next phosphodiester bond. For example, the ddNTP incorporated into the new DNA strand is ddATP, then that DNA fragment will end with an A.

Because the sequencing reactions are always set up with both template DNA and dNTPs in excess, DNA synthesis will continue until each strand incorporates a ddNTP and synthesis stops, meaning that the four sequencing reactions produce labeled DNA fragments of all lengths. If either dNTPs or template DNA were limiting in the reaction, then not all possible fragments would be produced and the sequence would be incomplete.

5. As with Maxam-Gilbert sequencing, Sanger sequencing uses polyacrylamide gel electrophoresis and autoradiography to separate the radioactive fragments by size. The sequence is read from the X-ray film.



Example of X-ray film derived from Sanger sequencing. Sequence read from bottom to top: GGGGATGAGCCTCGCATATTGAAAGGAGACCTACAAAGAA.

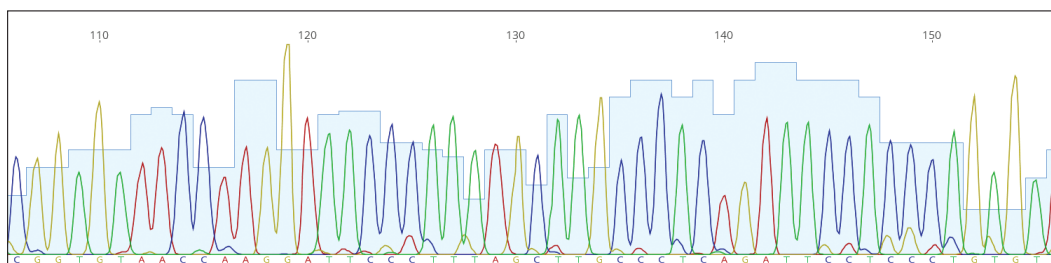
Using the Sanger method with dNTPs labeled with radioisotopes, it is possible to read up to several hundred bases from an autoradiograph, but the process is both time and labor intensive. It was now possible to sequence DNA in the research lab, using either the Maxam-Gilbert or the Sanger method, but both methods were woefully inadequate when scientists began to consider sequencing entire genomes.

Modifications to Chain Termination (Sanger) Sequencing That Allowed Automation

Modifications to the Sanger procedure have made it possible to sequence more DNA with much less effort. One of the first modifications was to tag the DNA with a fluorescent tag instead of a radioactive tag and to use capillary electrophoresis rather than a standard polyacrylamide slab gel to separate the DNA fragments. Although it was still necessary to run four separate sequencing reactions (one for each base), the sequence could be read automatically by detecting the fluorescent tag as the DNA fragments came off the gel.

Labeling the four ddNTPs with four different fluorescent dyes was the next step in the evolution of DNA sequencing, one that led to total automation of sequencing. This modification of the Sanger method is called dye-terminator sequencing. Because each dideoxynucleotide is labeled with a different dye (each of which fluoresces at a different

wavelength), the sequencing can be done as a single reaction. As the DNA fragments exit the capillary electrophoresis gel, the dyes are excited by lasers and the emitted light is detected. The result is a graph called a chromatogram or electropherogram where bases are represented by a sequence of colored peaks. The peak height indicates the intensity of the fluorescent signal. The automated sequencer interprets the results, assigning G, A, T, or C to each peak. If the software cannot determine which nucleotide is in a particular position it will assign the letter N to the unknown base.



Sample DNA sequencing chromatogram. Each peak on the graph represents one base and each of the four colors represents a different base. For example, a red peak represents an A.

Although not described here further, improvements in sequencing methods and in automation, such as cycle sequencing, have greatly increased the rate at which DNA can be sequenced. A modern automated sequencer can sequence close to one million bases a day. Imagine where genome science would be today without these advances. Using the early Maxam-Gilbert or Sanger sequencing methods meant that 1,000 bases of sequence was a good day's work. At that rate sequencing the 3 billion bases of the human genome would have taken over 8,000 years rather than the 13 years it actually took.

Instructor's Advance Preparation for Sequencing

Students will combine DNA and sequencing primers and mail them to a sequencing service and then analyze sequences using your Geneious Prime account. Eurofins MWG/Operon is offering a discounted rate for educators sequencing samples from this module. Alternately, there are other commercial sequencing services or local sequencing services that may be utilized. Most major universities have core labs that will sequence for a fee.

If you are using a university or commercial sequencing facility, ensure you determine and follow the specific requirements for submitting samples to that facility. Each sequencing facility has different requirements for receiving samples. Instructions for sample submission can be obtained from the facility itself and are usually available on the facility's website. Differences in sample submission may include concentrations of primer or DNA template, whether or not primers and DNA should be combined, and how the samples should be shipped — 96-well plate or microtubes. Also, please inform the sequencing facility that the primers you are sending (if you are using the ones provided in this kit) contain colored dyes. The colored dyes have been tested with standard sequencing reactions and do not interfere with the fluorescence detection of the sequencing instrument.

DNA Samples

The protocols outlined in this instruction manual are based on using clean, homogeneous plasmid DNA samples such as minipreps of PCR products or cDNA cloned into a plasmid. If not sequencing at Eurofins MWG/Operon, please confirm the required plasmid DNA concentration for the sequencing facility you will be using. The concentration can be estimated by comparing the intensity of a band of plasmid DNA sample with the Bio-Rad EZ Load Precision Molecular Mass Ruler on an agarose gel. With 5 μ l loaded per lane, the bands contain the following masses of DNA: 1,000 bp = 100 ng, 700 bp = 70 ng, 500 bp = 50 ng, 200 bp = 20 ng, and 100 bp = 10 ng.

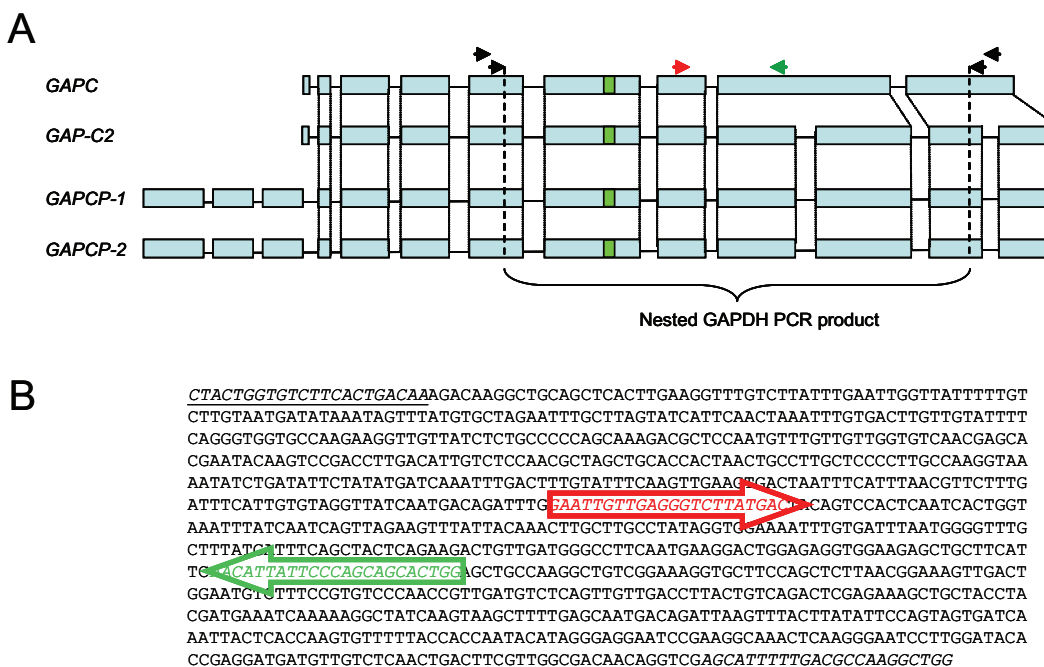
Sequencing Primers

Four sequencing primers are provided with this module. The primer names and sequences are as follows:

- pJET SEQ F, forward sequencing primer, 200 μ M (blue)
sequence: CGACTCACTATAGGGAGAGCGGC
- pJET SEQ R, reverse sequencing primer, 200 μ M (yellow)
sequence: AAGAACATCGATTTTCCATGGCAG
- GAP SEQ F, GAPC forward sequencing primer, 200 μ M (red)
sequence: GGHATTGTTGAGGGTCTNATGAC
- GAP SEQ R, GAPC reverse sequencing primer, 200 μ M (green)
sequence: CCAGTGGTGCTRGAATGATGTT

The pJET SEQ F and pJET SEQ R sequencing primers are designed to anneal to the pJET1.2 plasmid outside of the multiple cloning site. The primers can be used to sequence any gene that has been cloned into the pJET1.2 plasmid.

The GAP SEQ F and GAP SEQ R sequencing primers are designed to anneal to plant *GAPC* genes. The locations of the GAP SEQ primers are depicted below.



Location of PCR and sequencing primers for *Arabidopsis thaliana* GAPC genes. **A.** *Arabidopsis thaliana* has four *GAPC* genes with different intron/exon structures. The location of the initial and nested PCR primers from the *GAPDH* PCR module are depicted as the outer arrows. The location of the GAP SEQ F primer and the GAP SEQ R primer as the depicted forward and reverse inner arrows respectively. **B.** the locations the GAP SEQ F and GAP SEQ R sequencing primers anneal within the *Arabidopsis thaliana* *GAPC* gene are shown as the forward and reverse arrows respectively. The underlined sequences are the sequences of the nested PCR primers from the *GAPDH* PCR module.

If you have cloned a plant *GAPC* gene using the *GAPDH* PCR module or intend to sequence the pGAP control plasmid, both the GAP SEQ F and GAP SEQ R primers may be used. If you are sequencing an entirely different gene, you may also design your own primers that will anneal to your gene if this is wanted for greater depth of coverage or to sequence a longer gene.

Tips for Sequencing

It is vital that the number identifying the plate found next to the barcode on the sequencing plate label (for example A150936) be recorded in a secure place. This is the information to access the class's bioinformatics Geneious Prime account.



Barcode and plate number on 96-well plate.

It is highly recommended that students mix their DNA and sequencing primers in microcentrifuge tubes prior to pipetting them into the 96-well plate. This should reduce the likelihood of students pipetting their samples into the wrong wells. Using lab tape to temporarily cover completed wells also may help prevent pipetting errors.

A map of a 96-well plate has been provided to plan the location of the students' samples. It is recommended that each student team be assigned specific wells, for example a numbered column, and that the information gets recorded on the diagram. It may also be a good idea to dedicate rows A–D to the forward sequencing primers and rows E–H to reverse sequencing primers. This may simplify later analysis of the 96 sequence files.

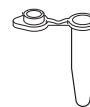
Tasks to Perform Prior to the Sequencing Lab

The requirements for this laboratory will change depending on your learning objectives, DNA samples and the requirements of your sequencing facility. The protocol described is based on each student sequencing two pJet1.2 plasmids containing novel DNA inserts. It is also advisable to have at least one student group set up reactions with the pGAP control plasmid and all four sequencing primers.

1. If using Eurofins MWG/Operon for sequencing, make sure you have a purchase order from your institution to pay for the reactions.
2. Educators not using Eurofins MWG/Operon
 - Locate sequencing facility
 - Determine required format of samples. It may be different than the instructions provided here
3. Retain the plate barcode number in a safe place. The plate barcode number is required for both Eurofins MWG/Operon and for a subscription to your Geneious Prime account.
4. Retain the foam shipping box for shipping the samples to the sequencing facility.
5. Place the ice pack in the freezer until ready for shipping.
6. Activate your Geneious Prime account at least **one week** prior to receiving your DNA sequences back from your sequencing facility. Each Sequencing and Bioinformatics module includes a subscription for a three month Geneious Prime account. The three month time period begins when you log in to your account for the first time. Please follow the instructions in the “Tasks to Perform Prior to the Bioinformatics Lab” section under Instructor’s Advance Preparation for Bioinformatics to log in and change the passwords for your account.

DNA Sequencing Quick Guide

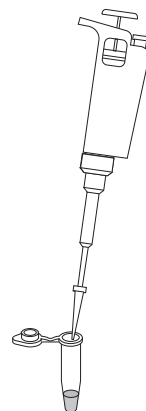
1. Label microcentrifuge tubes with the well numbers your instructor has assigned for your samples.



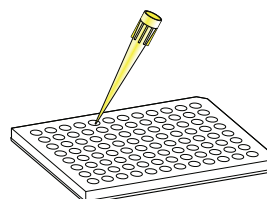
2. Fill in the following table for each sample.

Well Identifier	DNA Sample Name	Sequencing Primer

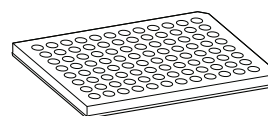
3. Combine 10 μ l of DNA sample with 1 μ l of the appropriate sequencing primer in a microcentrifuge tube. Pipet up and down to mix.



4. Pipet 10 μ l of each DNA/primer mixture into the appropriate well of the 96-well plate.



5. Once the entire class has added their samples, seal the plate with the sealing film. Record the barcode from the 96-well plate.



6. Carefully pad and pack the plate in the foam cooler with a frozen plastic ice pack and ship overnight to the sequencing facility.

Student Protocol

Overview

For DNA sequencing, you will combine DNA from your selected samples with the sequencing primers needed to obtain the sequence. Like PCR, sequencing reactions rely on the basic principles of DNA replication and as such require primers to initiate DNA replication. However, sequencing is performed in just one direction, so instead of a primer pair sequencing uses single oligonucleotide primers. Each sequencing reaction will sequence in a single direction. At least two sequencing reactions should be set up for each DNA sample: at least one forward sequencing reaction and at least one reverse sequencing reaction. This will ensure that as much of the cloned fragment as possible is sequenced. If the fragment is sufficiently short it will allow overlap of the sequencing reads, permitting both assembly of the two sequences and increased confidence in the sequence since it has been confirmed by two different sequencing reactions. A single sequencing run typically generates a read length of 600–800 base pairs (bp). If your DNA region of interest is too long, additional internal primers for sequencing may be required to ensure the entire cloned fragment is covered. Depending on what is known about the sequence of the cloned fragment, it may or may not be possible to design internal sequencing primers at this time.

The sequencing primers that will be combined with the plasmid DNA to be sequenced will depend on the particular samples being sequenced:

- pGAP control plasmid or novel pJet1.2 derived plasmids containing cloned plant *GAPC* genes can be combined with all four sequencing primers: pJET SEQ F (blue) and pJET SEQ R (yellow) that anneal to either side of the pJet1.2 multiple cloning site, and the internal sequencing primers GAP SEQ F (red) and GAP SEQ R (green) that are designed to regions of homology within plant *GAPC* genes
- PCR products cloned into pJet1.2 can be combined with the two plasmid-based sequencing primers: pJET SEQ F (blue) and pJET SEQ R (yellow) that anneal to either side of the pJet1.2 multiple cloning site
- PCR products cloned into plasmids other than pJet1.2 will require different sequencing primers which will be provided by your instructor — these may be to the plasmid vector, or individual primer sequences that match those used for the PCR primer pair

Once primers and plasmid DNA have been combined, the samples will be mailed to a sequencing facility that will provide electronic files containing the sequencing data. These data will be uploaded into a bioinformatics tool called Geneious Prime. The sequences are then assessed and analyzed.

Student Workstations

Each student workstation will require the following items to set up both a forward and reverse sequencing reaction for two DNA samples.

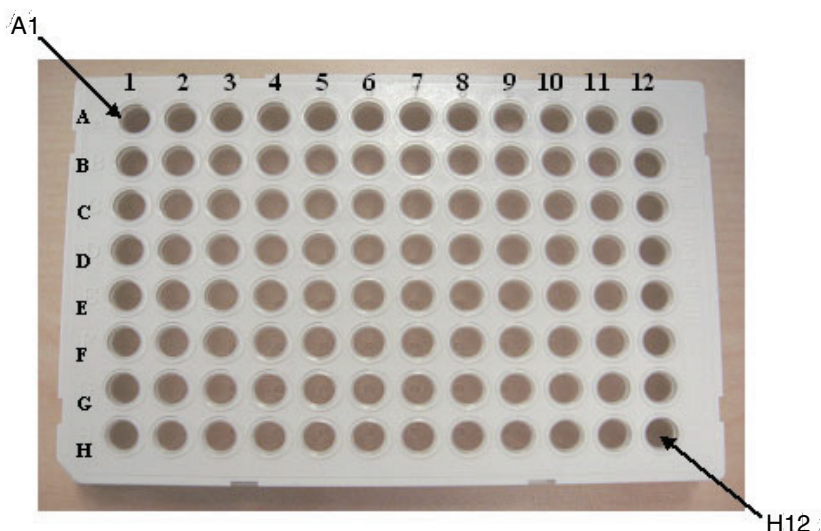
Materials Needed for Each Workstation	Quantity	(✓)
DNA sample cloned into the pJET1.2 plasmid	2	<input type="checkbox"/>
Colored microcentrifuge tubes, 2.0 ml	4	<input type="checkbox"/>
10 µl adjustable-volume micropipet and tips	1	<input type="checkbox"/>

Common Workstation

Material Required	Quantity	(4)
Forward sequencing primer (pJET SEQ F)	1	<input type="checkbox"/>
Reverse sequencing primer (pJET SEQ R)	1	<input type="checkbox"/>
Barcoded 96-well plate	1	<input type="checkbox"/>
Sealing film	1	<input type="checkbox"/>
Foam shipping box	1	<input type="checkbox"/>
Ice pack	1	<input type="checkbox"/>
96-well plate map specifying reaction locations (Note: Make sure to write the barcode number of the plate on the sheet)	1	<input type="checkbox"/>
pGAP control plasmid, GAP SEQ F, and GAP SEQ R sequencing primers for control sequencing reactions	1	<input type="checkbox"/>

Preparation for Setting Up Sequencing Samples for Sequencing

- Your instructor will assign each student team a group of wells on the class 96-well plate. Positions on a 96-well plate are identified by a row letter (A–H) and a column number (1–12). For example, the top left well is designated A1 while the bottom right well is H12.



2. Choose DNA samples to sequence. You can also sequence the pGAP control plasmid. At least one team should prepare the four control sequencing samples. The pGAP control plasmid should be combined with each of the sequencing primers individually.
3. Choose sequencing primers to be used. Preferably, this will include at least one forward and one reverse primer.
4. Plan your experiment. You will combine each DNA sample with each sequencing primer individually.

In the table below, record which plasmid combined with which primer will go into each well. When you name your sequences (sequence wells), make sure you use the same names when submitting samples to the sequencing facility.

Well Identifier	DNA Sample Name	Sequencing Primer

Detailed Protocol for Setting Up Sequencing Samples

1. Label your microcentrifuge tubes for the well into which the samples will be placed.
2. In your microcentrifuge tubes, combine 10 µl of DNA sample with 1 µl of sequencing primer. Pipet up and down to mix.
3. Pipet 10 µl of the DNA sample/primer mixtures into the assigned wells of the 96-well plate.

Write down the barcode number from the 96-well plate: _____

4. Once the entire class has added samples to the plate, seal the plate using the sealing film provided. Ensure a secure seal by rubbing extensively over the top of the plate with a gloved finger. It is essential to seal the plate completely so that the precious samples are not lost or cross-contaminated during transit.

If sequencing with Eurofins, go to eurofinsgenomics.com and follow these instructions.

- Follow the steps to create an account and enter 3070457 into the price agreement # field. If you already have an account, contact Eurofins customer service to have pricing agreement 3070457 added to your account.
- Click DNA Sequencing > Plate Sequencing > Standard Plates > Submit Seq Plate.

- Download and complete the Excel submission form.
 - Rename tab “Plate01” as your plate barcode number. This is the only way that Eurofins will be able to connect your physical plate with your order.
 - Enter Sample Type as “Plasmids”, Construct type as “1.00-5.99 kbp”, and Plate Layout as “By Columns”.
 - Enter the sample names according to their actual locations on your 96-well plate (e.g. A1). Select “Premix” for all samples in the Primer 1 column.
 - Upload the completed Excel file and click next.
 - Proceed through the ordering steps to place your order. Your discount will be applied directly in the cart.
5. Express mail the plate, well-packaged in a foam shipping box with a plastic ice block to maintain the reactions at 4°C, to your chosen sequencing facility. Pack the plate well to prevent shifting during transit and keep the foam box in its original cardboard box for added protection.

96-Well Plate Map

Record 96-well plate barcode number _____

	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C												
D												
E												
F												
G												
H												

Sequencing Focus Questions.

1. What is DNA sequencing?
2. Briefly explain the role of dideoxynucleotides in the traditional Sanger method of DNA sequencing.
3. How does automated sequencing that uses Sanger principles differ from traditional Sanger sequencing?
4. Since a single sequencing run generates only 600–800 base pairs of sequence (and eukaryotic genes are much larger than that), what are some strategies that can be used to acquire more sequence data?

DNA Sequencing – Quick Guide

1. Plan your experiment. Fill in the following table for each sequencing sample.

96-Well Plate Identifier	DNA Sample Name (10 μ l)	Sequencing Primer (1 μ l)	Primer Color

2. Label microcentrifuge tubes with the well numbers your instructor has assigned for your samples.



3. Combine 10 μ l of DNA sample with 1 μ l of the appropriate sequencing primer in a microcentrifuge tube. Pipet up and down to mix.



4. Pipet 10 μ l of each DNA/primer mixture into the appropriate well of the 96-well plate.

5. Once the entire class has added their samples, seal the plate with the sealing film. Record the barcode from the 96-well plate.



6. Carefully pad and pack the plate into the foam shipping box with a frozen plastic ice pack and ship it overnight to the sequencing facility.

BIOINFORMATICS

Background

Analysis of DNA Sequences Using Bioinformatics Tools

The wealth of information obtained through DNA sequencing of genes and the polymerase chain reaction (PCR), two biotechnological breakthroughs developed in the 1970s and 1980s, necessitated the development of an electronic repository for the many genes being discovered. This database, called GenBank, is operated by the National Center for Biotechnology Information (NCBI) and funded by the U.S. National Institutes of Health (NIH). GenBank is accessible via the Internet to scientists, teachers, and students worldwide free of charge.

Major efforts to completely sequence entire genomes were initiated in the 1990s and have now been completed for humans and for numerous model organisms studied by scientists, like the bacterium *Escherichia coli*, the common yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, and the murine family of rodents such as the house mouse, *Mus musculus*, and the brown rat, *Rattus norvegicus*. The speed and accuracy of gene isolation and sequencing have grown quickly. From 1982, when GenBank was at Release 3, to Release 242 in February of 2021, the number of nucleotide bases in GenBank doubled every 18 months. Release 3 contained only 606 sequences while Release 242 contains more than 776 billion sequences!

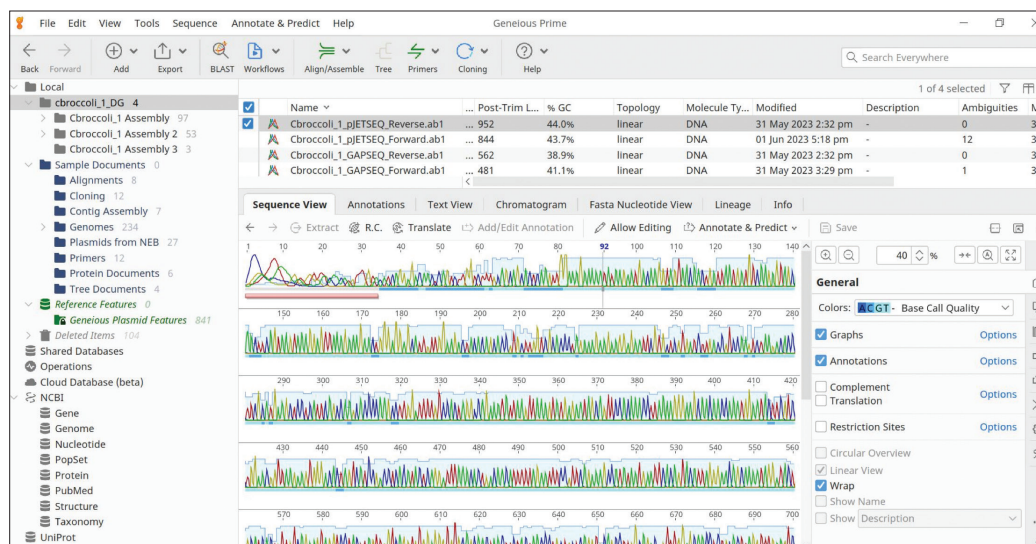
The challenge of analyzing all the DNA sequences deposited in GenBank spurred the development of numerous computer programs for interpreting DNA and protein sequence data. This computer-aided analytical approach is called bioinformatics. In addition to GenBank, other databases storing sequence information are available, as is a wide range of software programs and tools designed to obtain, analyze, and organize this information. The primary tools that will be used in this module to analyze sequence information are:

- Geneious Prime: a bioinformatics software platform from Biomatters, Inc. that offers data management, data analysis, and the ability to view DNA sequencing data
- BLAST (basic local alignment search tool), an online tool from the NCBI for comparing primary sequence data

DNA Sequencing Data

Once the sequencing reaction has been performed and the samples have been run on a sequencing instrument, the end result is a data file that contains a chromatogram. A chromatogram is a representation of the DNA molecules generated from the Sanger chain termination sequencing protocol, where the sequence of peaks represents the sequence of bases. A chromatogram provides information on the peak intensities, the time course in which they eluted, and the base calls that the instrument made for these peaks. The data can be analyzed manually by opening the data file in a program such as Geneious Prime. An example of a chromatogram is shown below.

The trace shows the peaks for each base in the order they eluted off the sequencing instrument. Above each peak is the letter code for the base that the sequencing instrument called for each peak (hence the term “base call”). This chromatogram also has information on the quality of the base calls. The colored boxes outlining them represent the quality of each base call. The cursor can be scrolled over each base call to display the quality score assigned to that base in the lower right-hand corner of the trace.

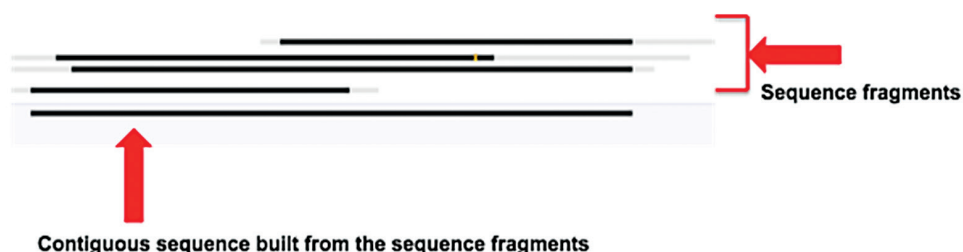


Example of a chromatogram viewed in Geneious Prime.

Geneious Prime Software Tools for Sequence Trimming and Assembly

Geneious Prime is a powerful piece of software that allows you to view the chromatograms as well as search for sequence data that need to be removed. Because we are sequencing a region of DNA that is contained within a plasmid, data from the cloning vector pJET1.2 needs to be removed. Another functionality of Geneious Prime is to remove low-quality sequence data from the 5' and 3' ends of a sequence, which is known as trimming. Sequencing reactions yield unreliable sequence data when near the priming sites and these low-quality data need to be removed. By cleaning up sequence data before further analysis, the best possible contiguous sequence (also called a "contig") can be generated using the sequence assembly function.

Currently, the average length of a sequence generated by a Sanger sequencing reaction is about 700 bases. Since most genes are kilobases in length, many overlapping sequences must be assembled to build the sequence of a single gene. This task is much like solving a jigsaw puzzle. To do it manually would be laborious and time consuming. Therefore, the computational capability to assemble many, many sequences is critical for determining the sequence of an entire gene. Geneious Prime can assemble a series of sequence fragments and, by incorporating quality score information, can generate the most likely full-length sequence from all the fragments (consensus sequence). This allows one long sequence for a gene to be constructed from pieces generated through overlapping individual sequencing reactions.



Generation of a contiguous sequence. Individual shorter sequences are compared, aligned, and assembled by programs such as Geneious Prime to generate longer contiguous consensus sequences. This methodology is used to generate continuous sequences that are longer than current sequencing instruments are capable of generating in a single sequencing reaction.

BLAST Searches

One of the initial steps in analyzing a novel sequence is to determine whether the sequence is like any others that have been sequenced before. To do this, the user-entered (query) sequence is compared to a database containing other sequences and a best match is determined.

The most commonly used tools for this analysis are the BLAST family of search tools, which are designed to find short (local) regions where pairs of sequences match. The BLAST family of programs and information on them can be found on the NCBI webpage (blast.ncbi.nlm.nih.gov/Blast.cgi).

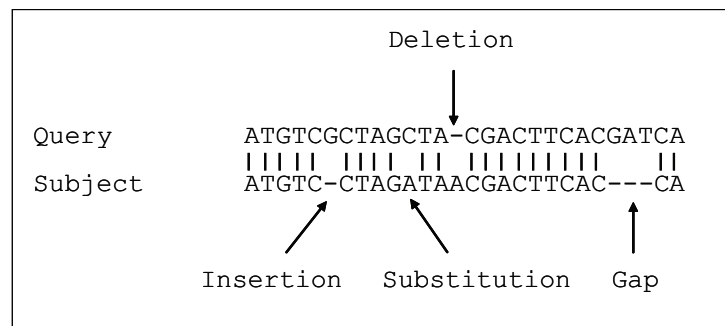
The BLAST search that is used usually depends on the type of sequence data that has been generated. For example, a *blastx* search translates a nucleotide sequence into predicted amino acid sequences and compares these to a database of protein sequences. If a protein sequence is the only information that is available, it can be used to search the protein database (protein blast) or a translated nucleotide database (*tblastn*).

Members of the BLAST family work in similar ways, but for now this discussion will focus on *blastn*. The *blastn* program is used to compare a user-entered nucleotide sequence (query sequence) to a database of nucleotide sequences. To do this comparison, *blastn* breaks the query sequence into “words” of a defined length. Then *blastn* compares each word to a database of words found in a user-determined set of nucleotide sequences. If all the letters in the words match perfectly, *blastn* looks at each end of the word pair to see if the matching region might be extended, trying to make the longest matching region that it can.

The set of user-entered nucleotide sequences should be chosen to give the best possible chance of a meaningful match. For example, the subset to be searched for an unknown plant *GAPDH* gene will be most productive when it contains only plant genomic sequences rather than human or mouse genomic sequences.

After the database has been searched, *blastn*, like all the BLAST programs, returns several statistics that, when used together, can help determine which sequence or sequences (known as subject sequences) in the database have the highest degree of alignment with the query sequence. Some of these statistics include the max score, total score, query coverage, max identity, and E value. These statistics will be explained in detail in Section 2 of the protocol.

The Geneious Prime software enables you to perform BLAST searches directly through its user interface using its “BLAST search” function. In much the way many researchers around the world access BLAST programs, this bioinformatics workflow will allow you to experience BLAST through both the Geneious Prime software and the NCBI’s BLAST website directly.

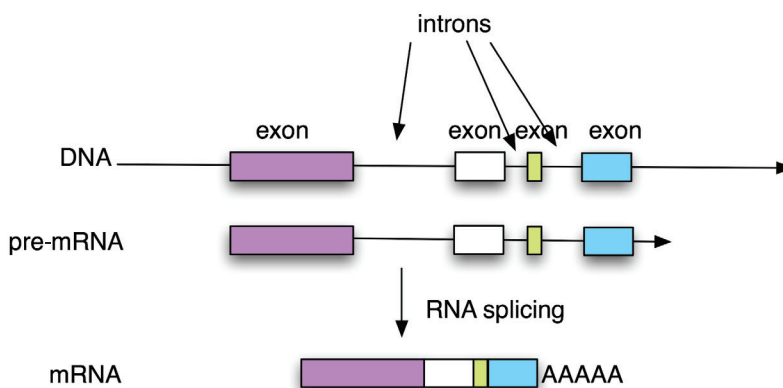


BLAST alignment. BLAST programs compare a user-entered (query) sequence with subject sequences in a database. It scores the match depending on the sequence identity and the number of differences — deletions, insertions, substitutions, and gaps — between the sequences.

Predicting an mRNA Sequence

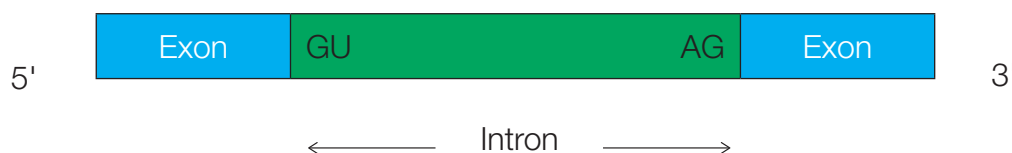
Most eukaryotic genes contain some sequence that does not code for protein. During gene transcription, RNA splicing removes these sequences, known as introns, and fuses (or splices) the sequences that contain coding information (exons) together at what are known as splice sites to form the mRNA sequence. This process is shown below.

After obtaining a genomic DNA (gDNA) sequence, a gene model that shows where the exons are likely to be located within the sequence can be constructed to predict the likely amino acid sequence for the encoded protein.



Gene splicing. Gene structure is composed of introns and exons. Introns are spliced out of the pre-mRNA to make mRNA.

Predicting splice sites in gDNA is an active area of research. Gene models are built by aligning mRNA sequences to gDNA, but not enough is understood about splicing signals yet to predict splice sites accurately by computation alone. There are some simple guidelines for determining splice sites, such as that the sequence of the mRNA at the 5' end of the intron tends to be GU and the sequence of the mRNA at the 3' end of the intron tends to be AG.



Pre-mRNA. Introns that are to be spliced out tend to start with GU and end with AG..

However, the presence of a GU or an AG is not enough to signal that an intron is present; other signal sequences within the intron sequence are also needed. Determining the different sequences that regulate splicing is a very active area of investigation, since this information can help clarify the multitude of splice variants for different proteins expressed in different cells or at different developmental stages.

In this lab, mRNA sequences will be aligned first to the four reference mRNA sequences for *Arabidopsis GAPDH* genes in order to help predict intron/exon positions. This alignment will then be further refined by aligning your putative mRNA sequence (query sequence) with mRNA sequences found in GenBank databases.

Predicting a Protein Sequence

Amino acids are specified by groups of three nucleotides called codons, which is what search programs use to compare protein sequences. Each DNA sequence can potentially be translated into codons via any of six reading frames, three for each strand (positive and negative). Each reading frame “frames” a consecutive nonoverlapping group of three nucleotides, or codons, in a sequence (for example, AGGTGACA in reading frame 1 = AGG | TGA, in reading frame 2 = GGT | GAC, and in reading frame 3 = GTG | ACA), and each frame must be read to determine which codons it encodes. A blastx search translates a nucleotide sequence in all six reading frames before it compares the resulting amino acid sequences to a database of protein sequences. Usually only one frame has any significant matches. A blastx search is thus very helpful for predicting the correct reading frame.

Instructor's Advance Preparation

In this bioinformatics stage, students will perform a series of analyses on their DNA sequences. The DNA sequences will be obtained either from Eurofins or from a local DNA sequencing service. The bioinformatic procedures used in this instruction manual are not automated and formulaic. The purpose of this exercise is to stimulate student understanding of the unique nature of real research data and the challenges this brings. The methodologies outlined in this portion of the series are a general framework for analyzing sequencing data. However, due to the novel and real nature of each dataset, it is impossible to predict a generic outcome for each analysis. Best efforts have been made to provide guidelines for general data analysis. However, you may find that certain aspects of the analysis need to be investigated in more depth and may require students to apply the skills they have learned in novel ways not directly specified in this manual.

The analysis portion of the lab is quite open-ended; the level of complexity and the depth of the analyses are entirely up to the instructor. Time constraints may not allow all steps in the process to be performed, but the following types of analyses are suggested.

1. Use Geneious Prime to look at the quality of individual reads.
2. Assemble sequences into a contig and correct sequencing errors with Geneious Prime.
3. Verify which *GAPDH* gene was cloned by conducting a blastn search on the contig sequence against the GenBank nr/nt database.
4. Annotate a gene by conducting a blastn search against the GenBank nr/nt database to predict gene structure (that is, intron/exon boundaries) and mRNA sequence.
5. Translate the predicted mRNA sequence into a protein sequence and verify that there are no stop codons.

To ensure accuracy of the data for those who wish to submit sequences to GenBank, we recommend that students assemble the final contig sequences for each plant using the same genes (GAPC or GAPC-2) derived from different clones. This will increase the depth of coverage for the gene and provide more confidence in the final sequence. After analysis, see Appendix C for instructions on GenBank's sequence submission policy and how to submit the class sequence information to the GenBank database.

Before teaching the lab you, the instructor, should become familiar with the protocols and software. We recommend that instructors go through the portions of this instruction manual they intend to teach using the example data for cbroccoli (bio-rad.com/CSresources).

These analyses are designed to be self-paced. Students may proceed through all of the protocols at once, or they may stop and save their data and then resume from where they stopped at a later time. While the protocols are divided into six main sections, it is possible to stop and restart in the middle of any of the steps if time constraints make that necessary.

Skills Required for the Bioinformatics Lab

This portion of the lab covers programs and techniques that are commonly used in bioinformatics. It does not address basic computer skills. Therefore, before beginning the analyses, it may be helpful to have students review the following techniques and software:

1. Geneious Prime software — students should familiarize themselves with the Geneious Prime user interface prior to performing the bioinformatics steps. See the section entitled, A tour of the Geneious Prime platform, as a general guide.
2. Internet searches — students should be able to use Internet search engines to look up the definition of a term.
3. Context-sensitive menus (menus that depend on what task is being performed or a particular location on a desktop or website. These menus are typically accessible by clicking on the right mouse button on a PC) — students should be able to:
 - Open contextual menus either by clicking the right mouse button or by using the Ctrl+click command
 - Locate the desktop
 - Navigate to find files

Tasks to Perform Prior to the Bioinformatics Lab

1. Make sure computers have the minimal system requirements to run Geneious Prime. See Table in Section 1.1.
2. Download Geneious Prime installer and install software on student computers (approx. 1 hr, depending on the number of computers) and submit your request for the license key from Geneious Prime.
3. Activate Geneious Prime software using the license key emailed to you from Geneious Prime.
4. Obtain DNA sequences and decide how to distribute the class's sequencing data (30 min, depending on Internet connection speed).
5. Set up Custom BLAST search services on each student computer. This step is required if this is the first time you are installing Geneious Prime onto the computers.
6. Enable the FASTA view custom feature to facilitate viewing sequences in FASTA format and GenBank submission.
7. Read and run through the activity prior to class using Chinese broccoli (*Brassica oleracea*) sample data (6–10 hr).

1. Check for minimum requirements for computers (approx. 30 min).

- 1.1** For Geneious Prime version 2021, check to make sure that computers have one of the following operating system versions before installing Geneious Prime:

Operating System	System requirements
Windows	7/8/10
Mac OS	10.11 - 10.16
Linux	Ubuntu Desktop LTS, last 2 supported versions

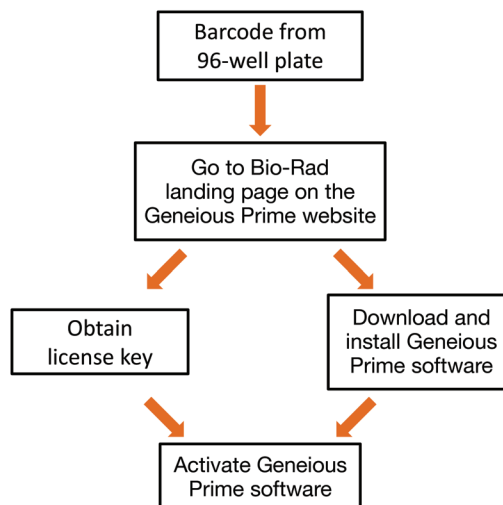
- 1.2** It is also recommended that computers have at least the following specifications for running Geneious Prime:

- Processor: Intel x86 / x86_64
- Memory: 2,048 MB or more
- Hard disk: 2 GB or more free space
- Video: 1,024 x 786 resolution or higher

1.2.1 To check your computer's hardware:

- On a Windows operating system:
 - o Go to **Start > All Programs**
 - o Open the Accessories folder, then the System Tools folder
 - o Select **System Information**. Here you will find your operating system name, processor, whether your system is 32-bit or 64-bit (system type), and memory (RAM)
 - o To check free space on your hard disk, click **Components > Storage > Drives**
 - o To check your display resolution, click **Display** under Components
 - o To check your Java version, go to **Start > All Programs**, then open the Java folder. Click **About Java** to find the build version
- On a Mac operating system:
 - o Click the **Apple** icon in the menu bar and go to **About This Mac**. Here you will find your Mac OS version, processor, and memory (RAM)
 - o To check free space on your hard disk, click **More Info > Hardware > Storage**
 - o To check your display screen resolution, click **Graphics/Displays**

- 5.3** Perform steps 5.1–5.2 for each student computer. Alternatively, you can copy the installation files from a computer that is already set up and upload them onto the rest of the student computers. This method may be speedier if your Internet connection is slow. See Appendix G for further instructions.



2. Obtain the Geneious Prime software license key, download the Geneious Prime installer, and install software onto student computers (approx. 1 hr).

Download the Geneious Prime software about one week before your DNA sequences are due back from your sequencing facility. Each Sequencing and Bioinformatics Module includes one 25-person license key for full, unrestricted use of the Geneious Prime software for 120 days. This license will comfortably outfit two computers per workstation (12 workstations in total), with one copy for the instructor's computer. The 120 day time period begins when the license key is emailed to you from Biomatters. After the license ends, Geneious Prime can still be used in a restricted mode, but some features will no longer be available.

2.1 Have the barcode from your 96-well sequencing plate ready.



2.2 Go to biorad.com/geneious and follow the instructions to request your license key and to download Geneious Prime.

3. Activate your Geneious Prime license on student computers.

Once the software has been installed, use the license key to activate software on all computers. Be aware that your 120-day countdown will begin on the day that your license key is emailed to you. Therefore, please plan accordingly — allow 3 business days from the day you contact the Geneious Prime team to email the license key to you, and build in any extra days you need to run through the sample sequences using the software prior to beginning the laboratory bioinformatics activities with your class.

Important! If using Geneious Prime on shared network computers, be sure to consult with your local IT department for installation and activation.

- Geneious Prime must be activated on each user account.
- Geneious Prime must have internet access to activate and to connect with NCBI servers. Be sure Geneious Prime is not behind a firewall, and adjust proxy server settings if necessary.

3.1 If this is the first time you are installing Geneious Prime onto student computers:

Open the Geneious Prime program. You should again see the dialog box asking about a license. Click **Activate a License**. If you have previously downloaded Geneious Prime, navigate to the Help button on the main toolbar (not the one with the orange question mark), and select **Activate License**. A new dialog box will appear.

- Select **Use license key**
- Enter a valid email address in the “Your Email” section

- Copy the license key from the text file that was emailed to you and paste it into the License Key field. Click **OK**

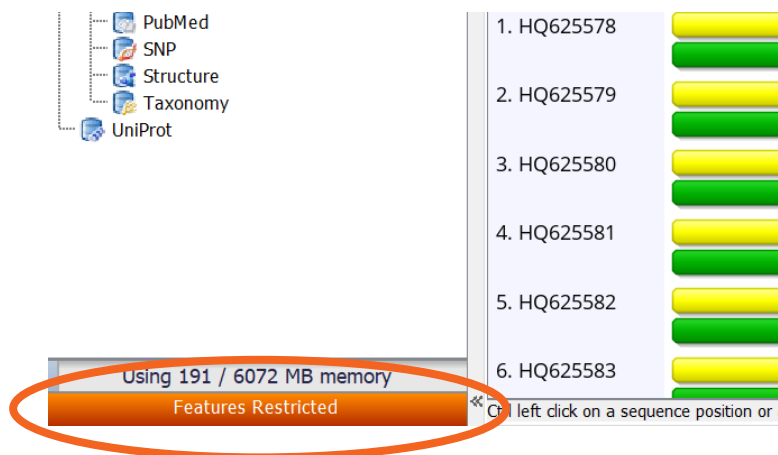
NOTE: The top navigation bar may have a different appearance in the previous versions of Geneious Prime software.



A new dialog box will appear to indicate that your registration was successful. You will now have full, unrestricted access to the Geneious Prime software. You may be required to install the Flexnet software license management program. Follow dialog to install.

3.2 Check to make sure that the licenses are working.

Open the Geneious Prime program and look on the bottom left corner of the Geneious Prime window. If you see an orange flag saying Features Restricted, there is a problem with your license key:



Check for successful license activation. If you see a Features Restricted orange flag in the bottom left corner of your Geneious Prime window, this indicates that there is something wrong with the license key. Contact Bio-Rad Technical Support for help resolving software activation problems.

If you see this orange flag, contact Bio-Rad Technical Support.

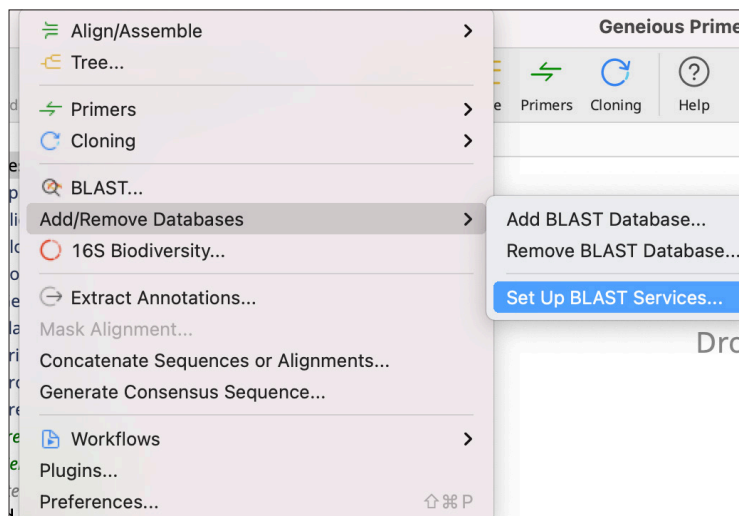
4. Distribute the class's sequencing data (approx. 30 min).

If the files are stored on the sequencing facility's database, or if they were emailed, you can download the data onto the hard drive of each individual classroom computer. Alternatively, you can give the web link and downloading instructions (or a memory stick or other type of portable storage media containing all the data) to your students so they can download all the sequencing data onto their computers themselves. Note that the latter strategy will take up class time, so if your class sessions are short, you may want to perform this step yourself.

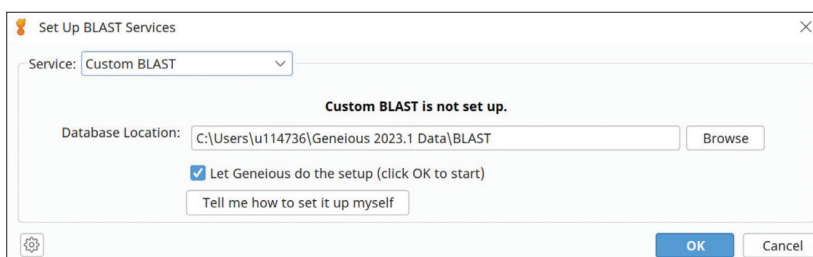
5. Set up Custom BLAST search services on each student computer.

Before students can trim vector sequences or run BLAST searches, BLAST services will need to be set up within Geneious Prime.

- 5.1 Open Geneious Prime on your computer. In the main toolbar of the Geneious Prime window, click the **Tools** tab, select **Add/Remove Databases**, then select **Set Up BLAST Services**.



- 5.2** A new dialog box will appear. For Service, choose **Custom BLAST** from the dropdown menu and check the box to let Geneious Prime do the setup for you.



Click **OK**. You will see a window with a progress bar and a time estimate. The download time will depend on your Internet connection speed (3–15 min). Geneious Prime will let you know when the BLAST search service setup is complete. Please see Appendix G for troubleshooting help.

- 5.3** Perform steps 5.1–5.2 for each student computer. Alternatively, you can copy the installation files from a computer that is already set up and upload them onto the rest of the student computers. This method may be speedier if your Internet connection is slow. See Appendix G for further instructions.

IMPORTANT NOTE: Regarding BLAST searches using Geneious Prime: In general, the amount of time it takes to retrieve BLAST results will vary depending on how many searches NCBI BLAST is running at that moment from researchers around the world. In some cases, searches performed through Geneious Prime are not as fast as performing the BLAST searches directly from NCBI.

If you have short class periods (50 min or less) and need the BLAST results as soon as possible to enable data analysis, you have the option of having your students perform the BLAST searches directly from the NCBI website for Sections 3 and 4. Please refer to Appendix G for protocol steps to export sequences as FASTA files for BLAST searching directly on the NCBI website.

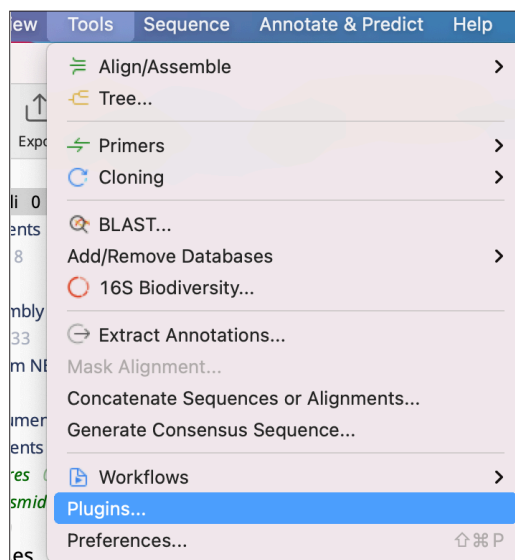
6. (Optional) Enable the FASTA View custom feature in the Geneious Prime program.

NOTE: This step is optional, but would be very helpful if your students plan to submit sequences to GenBank for publication.

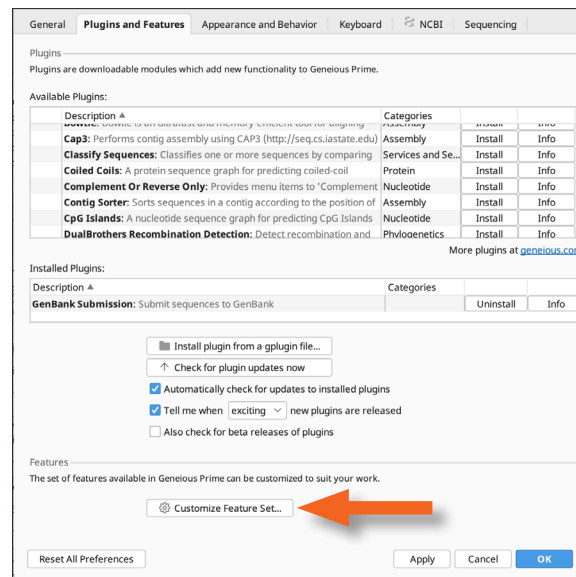
FASTA is a text-based format for representing sequences such that nucleotides (or amino acids) are denoted using single-letter codes. This format also allows for a single line of description for the sequence name and comments about the sequence itself. FASTA is a widely used sequence format in the field of bioinformatics.

GenBank requires that sequences be either cut and pasted or uploaded as a file in FASTA format. Thus, enabling the Fasta View feature in Geneious Prime will facilitate your GenBank submission process without requiring you to manually export sequence files.

- 6.1** Open the Geneious Prime program. Click the **Tools** tab in the main Geneious Prime window and select **Plugins**. A new Preferences dialog box will appear.



- 6.2** In the Plugins and Features section of the window, click **Customize Feature Set**. A Customize Feature Set window will open (see orange arrow in the figure below).



Description ▲	Categories		
EMBL importer: Provides EMBL import		<input checked="" type="checkbox"/> Enabled	Info
EMBOSS Tools: Provides various tools from the EMBOSS package		<input checked="" type="checkbox"/> Enabled	Info
Endnote import/export: Provides import and export functions for		<input checked="" type="checkbox"/> Enabled	Info
Export Consensus: Provides a single operation that extracts consensus		<input checked="" type="checkbox"/> Enabled	Info
Expression Analysis: Provides operations to calculates expression levels		<input checked="" type="checkbox"/> Enabled	Info
Extract Annotations: Extracts named or typed annotations from		<input checked="" type="checkbox"/> Enabled	Info
FASTA Importer/exporter: Provides FASTA file importing/exporting		<input checked="" type="checkbox"/> Enabled	Info
FASTQ Exporter: Provides FASTQ file importing facilities		<input checked="" type="checkbox"/> Enabled	Info
FASTQ Importer: Provides FASTQ file importing facilities		<input checked="" type="checkbox"/> Enabled	Info
Fasta View: Provides a fasta format view of sequences and alignments.		<input checked="" type="checkbox"/> Enabled	Info
Feature Finder: Annotate sequence with specified annotation sequences		<input checked="" type="checkbox"/> Enabled	Info
Find Duplicates: Provides the Find Duplicates function for finding		<input checked="" type="checkbox"/> Enabled	Info
GC/AT Content Graph: GC/AT Content Graph Plugin		<input checked="" type="checkbox"/> Enabled	Info
GCG importer: Provides GCG import		<input checked="" type="checkbox"/> Enabled	Info
GEL Plugin: A virtual GEL viewer, and a basic GEL analysis toolkit		<input checked="" type="checkbox"/> Enabled	Info
GFF import/export: Adds support for importing/exporting Sanger GFF		<input checked="" type="checkbox"/> Enabled	Info
Geneious Assembler: Geneious Assembler		<input checked="" type="checkbox"/> Enabled	Info

- 6.3** Click on the Description header to sort the features in alphabetical order. Click the checkbox for Fasta View to enable, and then click **OK**. The Customize Feature Set window will close. Click **OK** on the Preferences window, which will then close as well.
- 6.4** The Fasta View is now enabled. You will see the tabs named Fasta Alignment View and Fasta Nucleotide View in various sequence document windows when viewing sequences or alignments during the bioinformatics workflow.

You should now be ready to perform your bioinformatics analyses. To familiarize yourself and your students with the software being used and the tasks to be performed, you may want to perform all analyses on sequences generated from a Chinese broccoli (*Brassica oleracea*) clone and compare results to those already obtained for this sample. This clone includes four sequence files with which to perform these bioinformatics analyses.

Note: all files you generate will be stored on your computer and not on a server. When your 120 day account subscription expires, you can still run Geneious Prime in restricted mode to access your sequence files.

Protocol

Overview

After sequencing a gene of interest, bioinformatics tools can be used to gain additional information from the results. In this portion of the lab, you will obtain and analyze your DNA sequences. First, upload to your computer the raw DNA sequences returned by the sequencing facility so you can import it into the Geneious Prime software. You will analyze individual gene sequences and then perform a series of bioinformatics analyses with the resulting sequence information. Time constraints may not allow all steps in the process to be performed, but the following types of analyses are suggested:

Cloning the *GAPC* gene

- Identify and extract gDNA from plants
- Amplify region of *GAPC* gene using PCR
- Assess the results of PCR
- Purify the PCR product
- Ligate PCR product into a plasmid vector
- Transform bacteria with the plasmid
- Isolate plasmid from the bacteria
- Sequence DNA
- **Perform bioinformatics analysis of the cloned gene**

- Use Geneious Prime to look at the quality of individual reads
- Assemble sequences into a contig and correct sequencing errors with Geneious Prime
- Verify which *GAPDH* gene was cloned using BLAST (blastn) on the contig sequence against the GenBank reference genomic sequence database
- Annotate the gene by predicting gene structure (intron/exon boundaries) and mRNA sequence using BLAST (blastn) against the GenBank nr sequence database
- Translate the predicted mRNA sequence into a protein sequence, verify that there are no stop codons, and verify the sequence with BLAST (blastx)

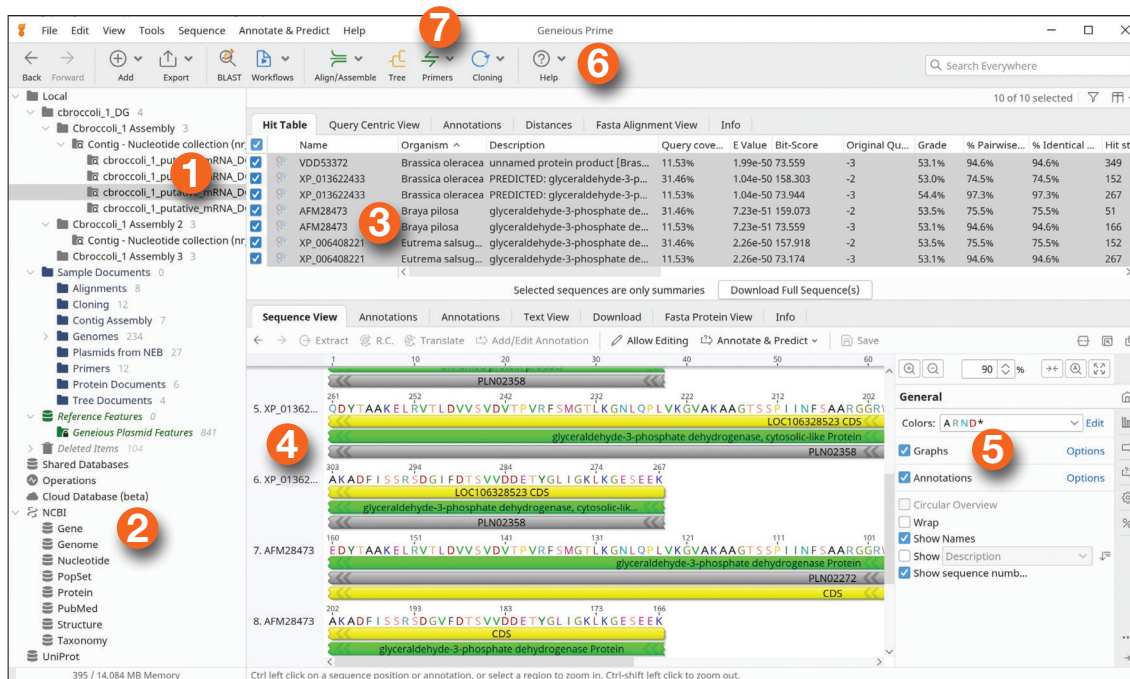
Assembling sequences from the entire class is an optional additional step. After analysis, the gene sequences can be submitted to the GenBank database (see Appendix C).

Materials Required

- Computer with Internet access
- DNA sequence files imported into Geneious Prime software

A tour of the Geneious Prime platform

Here is a quick orientation to the Geneious Prime platform and interface. Keep this reference handy as you go through your bioinformatics steps.



1, 2 – Sources Panel

This is where (1) all of your data are stored, and (2) you can designate which public database you would like to query.

3 – Document Table

The document table displays summaries of downloaded and imported data such as DNA sequences, protein sequences, journal articles, sequence alignments, and trees. This information is presented in table form. By clicking on the search icon you can search data for text or by sequence similarity (BLAST). You can enter a search string into the Filter (search) box located at the right side of the toolbar; this will hide all documents that do not contain the matching search string.

While search results usually contain documents of a single type, a local folder may contain any mixture of documents, including sequences, publications, and other types. If you cannot see all of the columns in the document table, you may want to close the Help panel (6) to make room for more panels. Selecting a document in the document table will display its details in the document viewer (4). Selecting multiple documents will show all the selected documents if they are of similar types; that is, selecting two sequences will show them both side by side in the Sequence View tab of the document viewer. Note that different files generate different tabs in the document viewer (4).

The easiest way to select multiple documents is by clicking on the checkboxes down the left-hand side of the table. To view the actions available for any particular document or group of documents, right click on a selection of them. These options vary depending on the type of document you are working with.

4 — Document Viewer

The document viewer panel shows the contents of any document highlighted in the document table, allowing you to view sequences, alignments, trees, 3-D structures, journal articles, abstracts, and other types of documents in a graphical or plain text view. Many document viewers allow you to customize settings such as zoom level, color schemes, layout, and annotations (nucleotide and amino acid sequences); three different layouts, branch and leaf labeling (tree documents); and many more.

The document viewer panel contains two tabs that are common to most types of documents: Text View and Info. Text View shows the document's information in text format. The exception to this rule occurs with PDF documents, in which the user needs to either click the View Document button or double click the file itself to view it. Most viewers have their own small toolbar at the top of the document viewer panel.

5 — Options Panel

The available options vary with the document being viewed. Examine the selections that run vertically in the options panel to explore the different ways you can display your sequence in Sequence View.

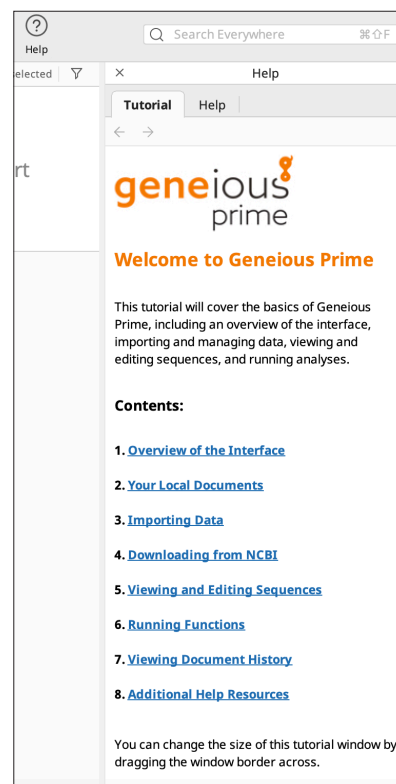
6 — Help Panel

The Help panel is accessed by clicking on Help and selecting “Quick Help”. The Help panel has two tabs: Help and Tutorial. The Help tab provides information about the service you are currently using or the viewer you are currently viewing. The Tutorial is aimed at first-time users of Geneious Prime and has been included to provide a feel for how Geneious Prime works. It is highly recommended that you work through the tutorial if you haven't used Geneious Prime before.

7 — Menu Bar

The menu bar contains several icons that provide shortcuts to common functions in Geneious Prime, including BLAST; Workflows, which search databases for new content even while you sleep; Align/Assemble; Tree building; and Help. The Back and Forward options help you move between previous and subsequent views in Geneious Prime and are analogous to the back and forward buttons in a web browser.

For more useful Tips and Tricks on personalizing and navigating the Geneious Prime software, please see Appendix H.



Note: Additional information and resources to navigate Geneious Prime interface is available at **Geneious Academy**.

1. View sequence traces and review the quality of the sequencing data.

Using the Geneious Prime software for bioinformatics

In this part of the lab, you will use many of the features of the Geneious Prime software to manually review your sequencing results. When chromatogram files are uploaded to Geneious Prime, several programs act in sequence to analyze and process these data. You will explore some of these data.

Extracting sequences and assessing their quality

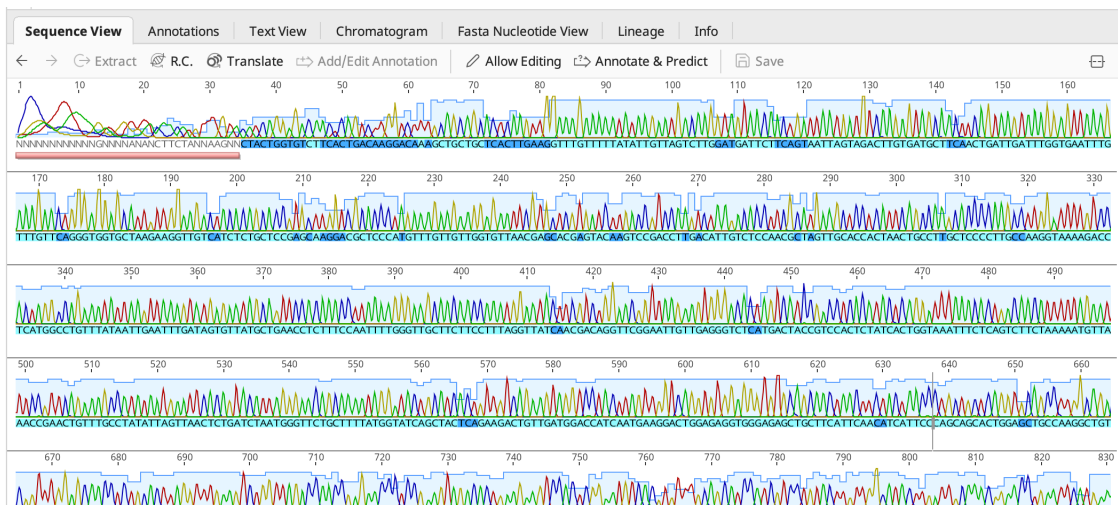
Upon import of .ab1 files, Geneious Prime will read your data files automatically and will display elements like quality scores and DNA sequence information from the file containing the sequencing chromatogram. This chromatogram was generated by the sequencing software after the DNA was sequenced.

During DNA sequencing, fluorescently labeled molecules of DNA are separated by size through capillary electrophoresis. As the DNA moves through the capillary, it passes in front of a fluorescence detector that measures the intensity of the light signal. Software in the sequencing instrument processes that signal and associates the signal with a nucleotide base. The chromatogram file also includes information about the presence of signals corresponding to alternate bases at each position and their intensities.

When this information is presented in a graph, it is called a trace. At the top of the trace are the base calls (the bases the sequencing software identifies as belonging to each peak in the chromatogram). Each letter represents a base call (A, T, C, or G), with any unidentified bases assigned an N. For each base, the height and shape of the peak corresponds to the signal intensity, and the spacing of the peaks shows the relative times at which the signals were measured.

Many DNA sequencing instruments contain additional software that can evaluate the signal intensity, the time between signal peaks, and whether any peaks overlap. This software provides additional information about the quality of each base. The ability of base-calling software to accurately interpret raw peak traces varies with the quality of the sequence data. Advances in base callers led to the evolution of the “quality” or “confidence” score, originally called a phred score. The phred score assigns a quality value to each called base. Quality scores are numeric values corresponding to each base call that define the likelihood that the base call is incorrect. The most common scale is from 1 to 60, where 60 represents a $1/10^6$ chance of a wrong call, 50 a $1/10^5$ chance, 40 a $1/10^4$ chance, etc. A higher quality score means a greater confidence that the base call is correct, and a lower quality score suggests that the base call has a lower chance of being reliable. Depending on the program used to generate the confidence value, the quality score may be based on peak height, the presence of more than one peak, and/or the spacing between the peaks.

A base is considered to be of high quality when its identity is unambiguous. A high-quality region of sequence has evenly spaced peaks that do not overlap and has signal intensity in the proper range for the detection software. A quality score of ≥ 20 is considered the threshold for reasonable confidence in the data.

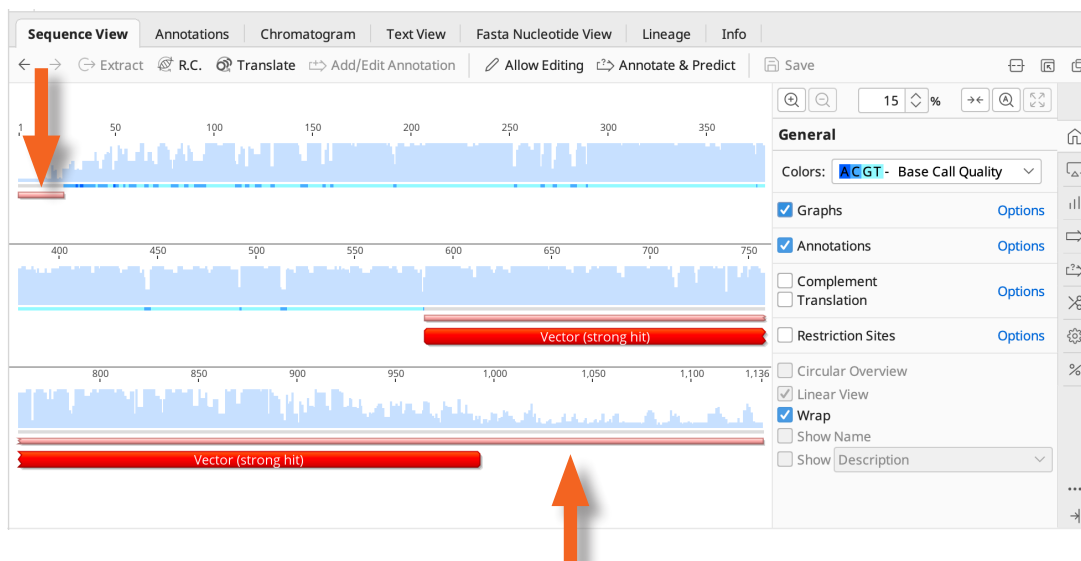


Example of a chromatogram viewed in Geneious Prime. Compared with the bases at either end of the chromatogram, the bases in the middle of this read are high quality because the peaks do not overlap, are evenly spaced, and are all approximately the same height.

Quality trimming

The sequence of bases from each chromatogram is called a read. Once the read and quality information have been extracted from the chromatogram or determined by the base-caller program, other analytical programs can be used. Since the data at the 5' and 3' ends of reads are often poor quality, a standard step in DNA sequence analysis is quality trimming. In this process, software examines the quality of each base at either end of the read. When quality scores for the bases at each end of the read fall below a certain threshold, usually 20, the trimming program marks those positions and measures the length between trim points, which defines the read length.

Later, when we prepare the DNA sequence for alignment of the nucleotides with other sequences, we can elect to have portions of low-quality sequence trimmed or hidden. In Geneious Prime, the trimmed regions are indicated by a light pink annotation.

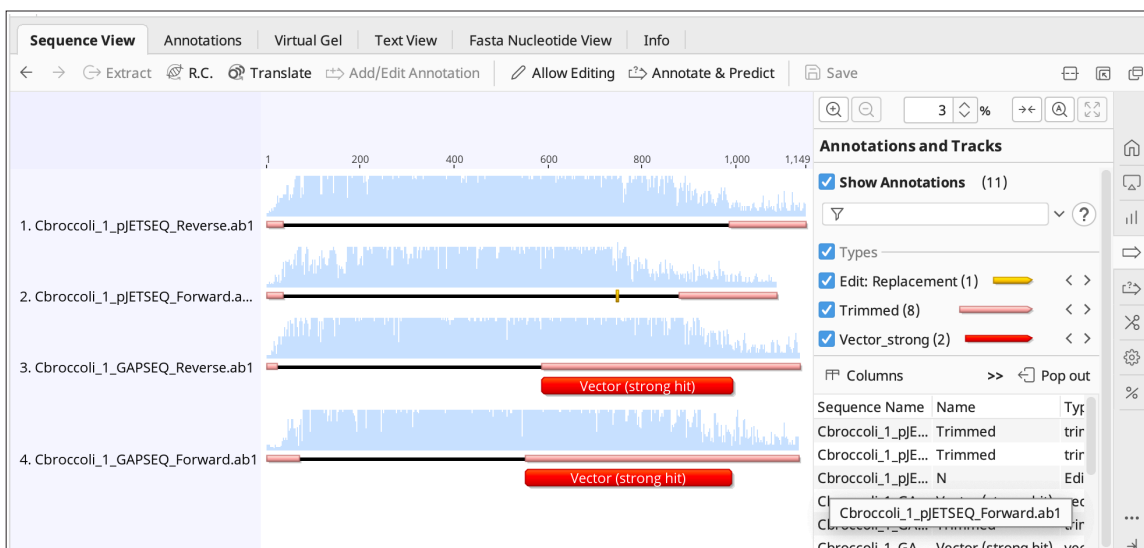


Example of a trimmed region. The light pink lines beneath the sequence indicate low-quality bases at the 5' and 3' ends of a sequence in Geneious Prime.

Vector identification and trimming

If the distance between the sequencing primer and the 5' end of the cloned fragment is short, the resulting chromatogram may include sequence from the cloning vector. A vector sequence can also be present at the 3' end of sequence traces if the sequencing reaction runs for large numbers of bases. This may interfere with subsequent analyses. One feature of Geneious Prime is the ability to automatically provide vector identification and masking. The vector sequence identification step involves comparing the read sequence to a database of cloning vector DNA sequences, determining the percentage that match, and obtaining a score. This analysis allows a determination of which parts of a read are similar to the cloning vector, allowing you to quickly determine how much of the sequence obtained is part of the gene itself and how much is the cloning vector, thereby enabling you to trim the vector sequence off.

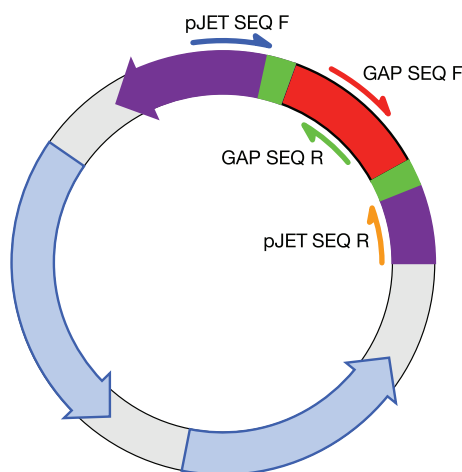
Vector masking is an optional step that can take place when data are imported into Geneious Prime for use in an external program such as BLAST. If this option is selected, the trimmed vector regions in Geneious Prime are also indicated by a red line that will hide them from other DNA analysis programs. This simplifies future analysis of the gene of interest by eliminating interference from the vector sequence.



Chromatograms in Geneious Prime showing vector sequence annotated by a dark red line. Just like the quality-trimmed annotations in pink, these vector sequences are hidden from other DNA analysis programs to prevent their interfering with subsequent analyses.

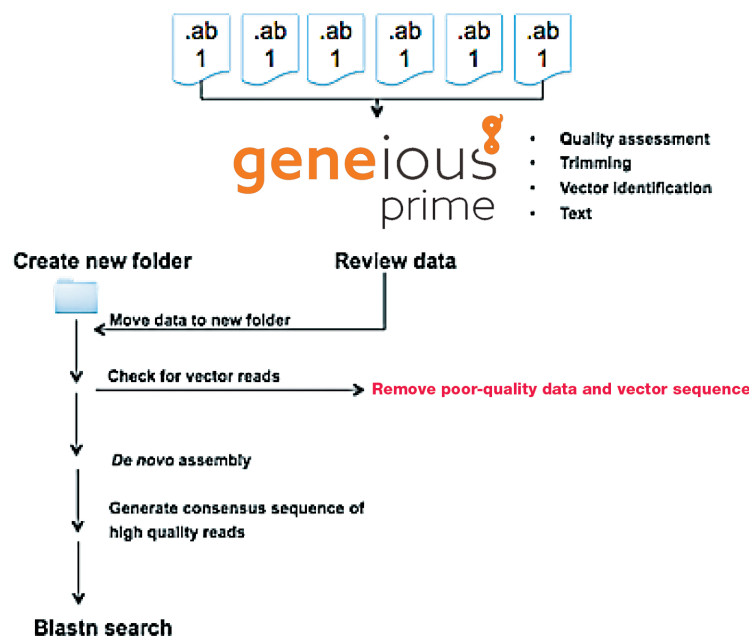
1.1 View sequence traces and review the quality of the sequencing data.

In this part of the project, you will review and identify high-quality sequences from your dataset and work with your own sequencing data to prepare them for further analysis with programs such as BLAST. At the end of the hands-on part of the cloning lab activity, miniprep plasmids or clones that you generated were combined with four different sequencing primers and sent for sequencing. The sequencing primers are designed to sequence different parts of the cloned gene because reading from a single primer would sequence only part of the gene.



Forward and reverse sequencing primers for pJET1.2 plasmids with *GAPDH* inserts.

Refer back to your notes on the miniprep clones that were sequenced, the sequencing primers used, and their positions on the 96-well plate (if used). This information will help in locating your data. Within Geneious Prime you will create one folder for each miniprep clone your team sent out for sequencing and then discard poor-quality data and chromatograms that are composed mainly of vector sequences so that they will not interfere with future steps. The workflow for this stage is shown below.



Workflow for processing GAPDH sequencing data.

1.2 Downloading .ab1 sequence files onto your computer.

If this step has not already been done by your instructor, you can retrieve your sequencing files in a few quick steps.

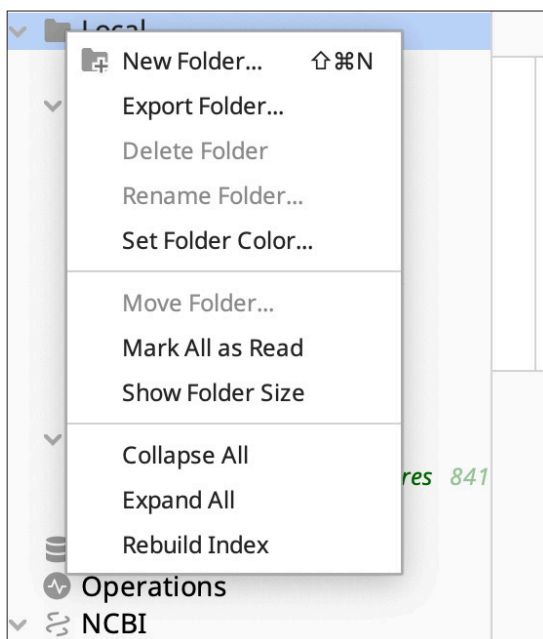
- If the files are stored on the sequencing facility's database, ask your instructor for the link to the website and any instructions that accompany the retrieval of your files.
- If your instructor set up a database server or a web-based data storage site, obtain the link to the website or database to download your files directly to your computer.
- If your instructor has all the sequencing files on a memory stick or other type of portable storage media, connect this to your computer for a direct download of the files.
- If the files provided are individual files and not compressed into a .zip file, it is recommended that you save disk space by compressing the files into a single .zip-formatted file. Mac OS X computers can create .zip files and Microsoft Windows computers use a program called WinZip to create .zip files. If desired, files may also be uploaded one at a time, but this can be time consuming. Please note, for Eurofins files, zip only the files with the suffix .ab1, as these are the correct format to be used by Geneious Prime. Geneious Prime will automatically unzip your files when you import them.

1.3 In Geneious Prime, create a new folder for your data.

- 1.3.1 In the Sources panel, click to select the Local folder at the top of your directory (see figure).
- 1.3.2 Right click your mouse and select New Folder from the shortcut window. Once the new folder is created, rename it.

Working with your data will be easier if it is well organized. Create folders for each clone of your plant and place the sequence results data in the appropriate folders. Name your folders with the name of your plant, the number of the clone and your initials. For example: **cbroccoli_1_DG**.

Tip: Right click on the new folder and assign it a custom color to further organize your sequence data.



Right click on the Local folder and select New Folder to create a new folder for your data.

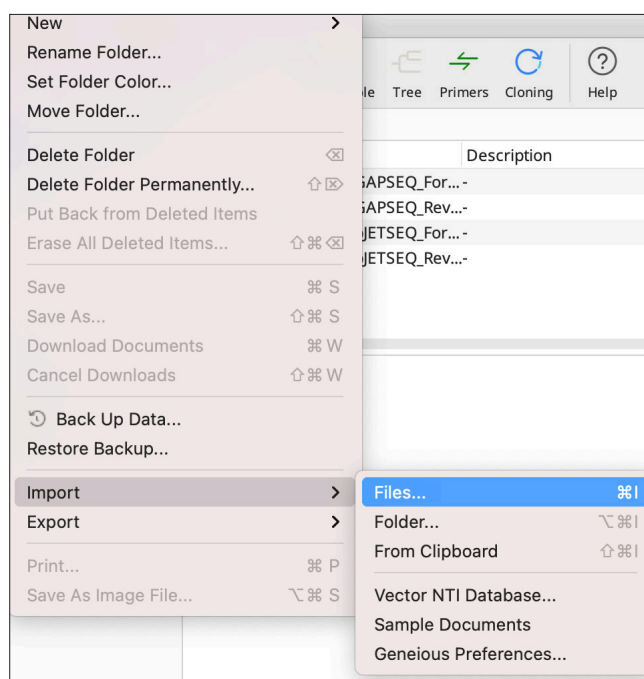
1.4 Import sequence files and view chromatograms.

There are two ways to import your .ab1 file data into Geneious Prime: drag and drop your files or use the File tab at the top of the Geneious Prime program's toolbar.

- 1.4.1 To drag and drop your .ab1 files directly into your new folder in Geneious Prime:
 - Click to highlight and select your new folder in the Sources panel
 - Locate your .ab1 sequencing files on the local hard drive of your computer
 - Click on a file name to select it. To select several files at once, hold Ctrl and click each file name. Drag the selected files to the Document table in Geneious Prime to drop them into your folder. Your .ab1 files should now be in your folder

1.4.2 Using the File tab from the main toolbar:

- Click to highlight the folder you created. The document table shows you the contents of the highlighted folder. Your folder should currently be empty
- On the toolbar for the main Geneious Prime window, click the **File** tab, and then hover your mouse over **Import**. A window will appear. Select **Files...**
- Locate your .ab1 files and click Import to continue. Your sequence files will now appear in your folder in the Sources panel, and in the Document Table.






1.5 Examine your chromatogram traces and evaluate data quality.

Import your data using the File tab in the menu bar.

The document table contains information about your sequencing data in a column format that includes the name of your file, the sequence length, the percent of the sequence of a particular quality (for instance, HQ% = percent high-quality bases, LQ% = percent low-quality bases), etc.


- 1.5.1 Examine your chromatograms. Click a file once to open it in the document viewer panel. To look at multiple sequences at once, click the checkboxes down the left-hand side of the document table.

The file data will be shown in the document viewer in a tab labeled Sequence View. Geneious Prime will display the chromatogram traces superimposed on the sequence. If the chromatograms are not visible, go to the Graphs tab and check the Show Graphs and Chromatogram boxes. When several sequences are selected, you can scroll through all sequences at once by using the scroll bar at the bottom of the page.

Tip: To zoom in on your sequence on the x-axis, which spreads out the peaks, use the magnifying glass  tool at the top of the options panel. To zoom in on the y-axis, go to the options panel and click the Graphs tab . In the Chromatograms section, increase the value in the box on the right to stretch out the y-axis . This function is particularly useful when viewing multiple sequences at once, when the chromatograms appear smaller.



Viewing a chromatogram trace in Sequence View. All four sequences are selected for viewing from base 1, zoomed in at 86%. The files in the document table are all grayed out because they are all selected but the mouse is active in a different panel in Geneious Prime. Check the box for Show Names to see the sequence names to the left of the chromatograms.

To set up your column data to match the image above, right-click on any of the column headings or click on the small data table icon  on the upper right of the document table. The image on the right shows a partial view of the options that will appear on your screen, and the blue checked options are the most useful data to display: Name, Sequence Length, LQ% (% low-quality bases), HQ% (% high-quality bases), and Molecule Type.

Do your data include long sequence lengths?

Do all your sequences have a large number of bases that are high quality?

<input checked="" type="checkbox"/> Small Rows
Manage Columns...
<input checked="" type="checkbox"/> % GC
<input checked="" type="checkbox"/> % HQ
<input checked="" type="checkbox"/> % LQ
% MQ
<input checked="" type="checkbox"/> Ambiguities
At least Q20
At least Q30
At least Q40
Bin
Created
<input checked="" type="checkbox"/> Description
Failed Binning Fields

Partial list of column data options for your sequences in the document table. These options can be accessed by right clicking any column header or clicking the small data table icon on the upper right of the document table.

1.5.2 View the quality scores for your sequences. The quality of the base calls is indicated in two general ways. First, the quality scores are indicated by a varying blue color scheme on the sequence itself. The software automatically assigns a shade of blue to each base according to its quality (confidence) score for all alignments of all chromatograms. Confidence scores are represented as dark blue for <20, medium blue for 20–40, and light blue for >40. Second, a quality score histogram is superimposed on the chromatogram itself. This histogram is light blue in color and gives a quick survey of the distribution of high-quality base calls in your sequence. The taller the bar, the better the quality.

- When mousing over the sequence or chromatogram, a line follows your mouse to indicate which base the mouse is pointing at. At the bottom of the window, text tells you which base your cursor is fixed on and what base you are mousing over and its quality score:




Quality scores for base calls. The text at the bottom of the window (orange rectangle) reports the quality score of the base being marked by the line cursor (marked by an orange arrow).

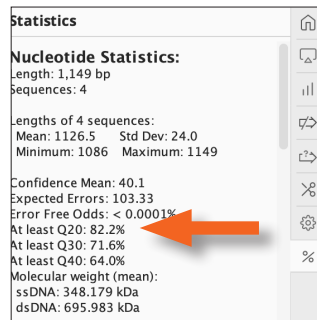
What do multiple peaks at the same location and blunt peaks indicate about a base call's quality score?

Are the quality scores of medium blue bases higher or lower than those of the light blue bases?

Are the bases with lower quality scores toward the middle of the sequences or at the 5' and 3' ends?

What are the differences, in terms of sharpness of peaks, number of overlapping sequences, and number of individual bases with low quality scores, between sequences with low versus high overall quality scores?

- The Statistics tab  in the options panel will show the percentage of sequence(s) that fall under certain quality scores:



The Statistics tab in the options panel. The Statistics tab displays the percentage of your sequences that have specified quality scores (bottom orange arrow).

- The Graphs tab  in the options panel includes a checkbox that lets you see the quality score histogram:



Toggling the view of the quality score histogram. In the Graphs tab of the options panel, the quality score histogram can be toggled off (top panel) and on (orange arrow pointing to light blue histogram) using the "Qual" checkbox (circled in orange).


1.6 Sequence data cleanup.

In this part of the project, you will be cleaning up your sequences using Geneious Prime in order to prepare for alignment searching with BLAST and for generating contiguous sequence for further analysis. You will:

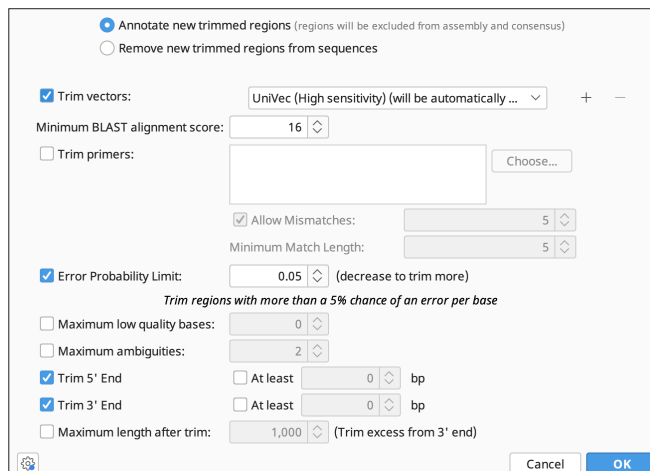
- Trim sequences to remove low-quality data
- Check for pJET1.2 vector sequence and remove it from your sequences
- Identify any sequence files that consist of all low-quality data and move them to the Deleted Items folder

- 1.6.1 Trim the low-quality regions of your fragments and remove pJET1.2 vector sequences from your chromatogram files. Since the data at the 5' and 3' ends of reads are often of poor quality, a standard step in DNA sequencing is quality trimming. In Geneious Prime, the quality of each base at either end of the read is examined, and the point at each end of the read where a predetermined fraction of the bases have quality scores above a certain threshold is marked. The sequence between the trim points will be the new read length if the low-quality bases are trimmed. The trimmed annotations are light pink in color and are called a “soft trim”; the low-quality bases are not removed, but the underlying nucleotide sequence will be hidden (ignored) for downstream analysis. Geneious Prime still shows these data because researchers will sometimes go back and reanalyze the untrimmed data. So, even if you can still see the trimmed bases, remember that they won't interfere with the next steps in your workflow.

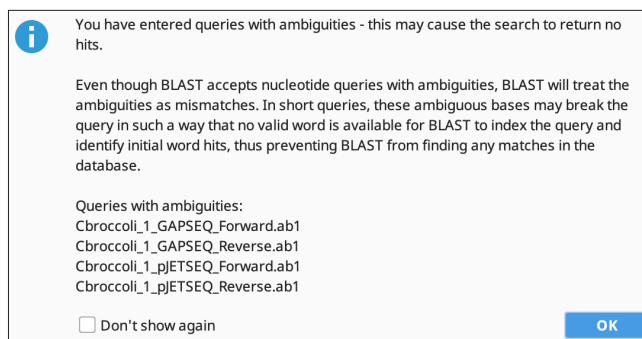
Removing pJET1.2 vector sequence. When using sequencing primers that anneal to the plasmid, such as the pJET SEQ F and pJET SEQ R sequencing primers, the resulting chromatogram may include sequence from the cloning vector. This sequence could immediately follow the sequencing primer or, if the sequenced clone is not particularly long, sequencing reactions could continue beyond the end of the gene and sequence the plasmid at the far end of the insert.

1.6.1.1 Sequences can be trimmed one at a time or in a batch. Click to select one sequence (or multiple) from the document table. In the Sequence Viewer toolbar, click the Annotate & Predict button  Annotate & Predict and then choose **Trim Ends**. A new dialog box will appear:

- Select **Annotate new trimmed regions**



- Check the box for **Trim vectors**
- Click **OK**. A new window may open titled Ambiguous Query:



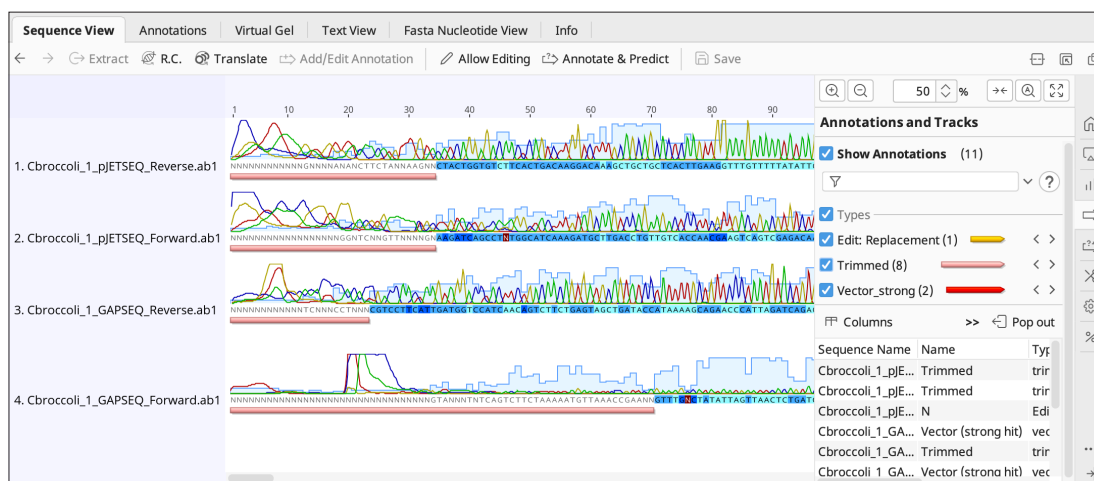
Appearance of this window indicates that the sequence still contains nucleotide(s) assigned as N that did not get trimmed away and will be treated as mismatches by BLAST when it tries to match up vector sequences from the UniVec database. Click OK and proceed. If the trimming step fails, it is a good idea to go back and check the quality of any N base(s), and their neighbors, near the ends of the sequence to see if an extra manual trimming step will resolve the problem (see step 1.6.2). However, if these Ns are in the middle of your sequence, it is best NOT to trim or delete them, as they may represent something other than errors from low-quality sequence

- In the Sequence View toolbar, click **Save** to preserve the new trim regions. In Sequence View, you will see that some portions of your sequences are grayed out or may have a strikethrough. This indicates that this part of the sequence has been trimmed away



Trimmed sequences have a strikethrough and are grayed out when annotation boxes are not checked.

Geneious Prime has processed your data and placed trim annotations under the sequence where it determines the error rate is too high. A trimmed annotation feature will also be added to the Annotations and Tracks tab in the options panel. Click on the box for **Types** (make sure Show Annotations is enabled) to display the annotations on your sequences. The Trimmed annotation appears in light pink, representing the quality-trimmed regions. The Vector_strong annotation appears in bright red, representing the sequences that have very strong matches to vector sequences and are unlikely to be part of your plant sequence. If there are no strong matches to vector sequences, the annotation will not appear.



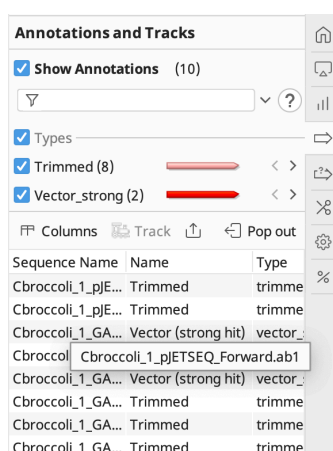
Quality-trimmed regions are annotated as light pink lines, and vector matching sequences are annotated as dark red lines beneath the sequence. These regions are hidden from downstream processes and will not interfere with further bioinformatics steps.

Note: Keep in mind that even though you can still see the trim annotations, they are actually hidden away from any subsequent bioinformatic steps. You can always go back to reanalyze your data or manually trim your sequences if you disagree with the automatic trim regions. For example, if there are stretches of Ns that have not been trimmed near the ends, you may want to manually trim these away.

What's an annotation?

Nucleotide and protein sequences downloaded from curated sequence databases often come decorated with annotations, which identify the locations of various features of a gene to help give context to what those genes do. These annotations are also called metadata, and may be viewed in Sequence View with your sequence if they are available. Annotations may be either placed directly on a sequence in Sequence View or grouped into tracks that can be expanded or hidden.

A **track** is a collection of one or more annotation types. Tracks are stacked vertically underneath the sequence, with a separate line for each track and its annotations. For example, when you trimmed your sequences for quality and vector masking in Section 1, these defined sections of your sequence became annotations in a track beneath your sequence. If your sequence data have annotations and tracks, the options panel will include the Annotation and Tracks tab, denoted by the open arrow icon (\Rightarrow) in the options panel:

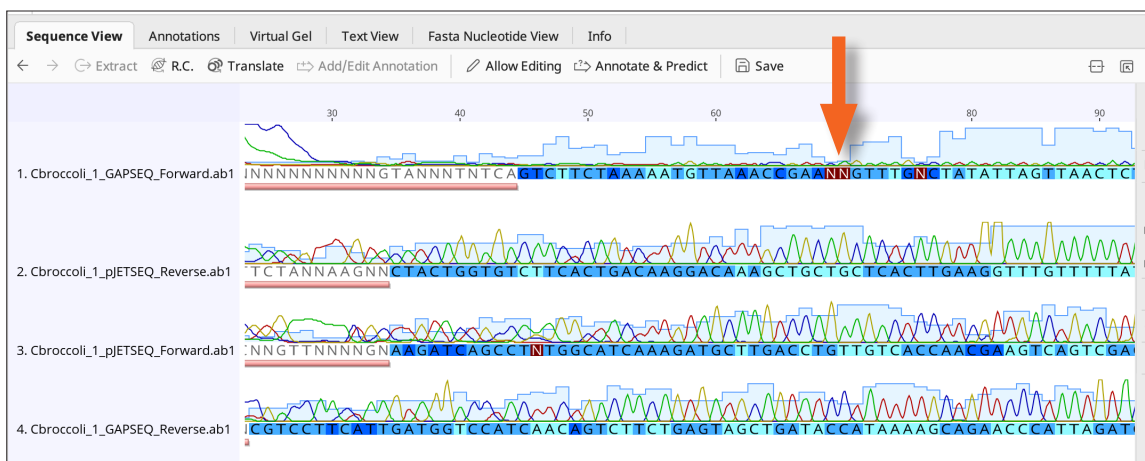


- Beneath Show Annotations is a text field where you can search for annotations
- Beneath the text field is a list of annotation types for the tracks present in the current sequence. These annotations are either directly annotated on the sequence or are present in multiple tracks
- Clicking on the left or right arrowheads $\leftarrow \rightarrow$ will take you to the part of the sequence where the annotation resides in the Sequence View window

1.6.2 To manually trim your sequences:

Note: You will want to trim the Ns only from the ends — not the middle! — of your sequence. The middle of your sequence should have fairly high-quality scores, so a base call of N here would mean something different than simply a low-quality error.

- 1.6.2.1 Click the single sequence you would like to work with and click Allow Editing in the Sequence View tab toolbar. Identify the region you would like to manually trim away. In the cbroccoli example below, the problematic bases are the two Ns at positions 69 and 70.
- 1.6.2.2 Click the trim annotation. Hover your cursor over the right edge of the pink bar and it will change to a vertical bar with arrows to the right and left.



Manual trimming two N base calls near the end of a sequence. The two problematic Ns are indicated by the orange arrow. These will be trimmed away by manually extending the light pink line to cover them. Note that only one sequence is selected (Cbroccoli_1_GAPSEQ_Forward.ab1) and shown in Sequence View.

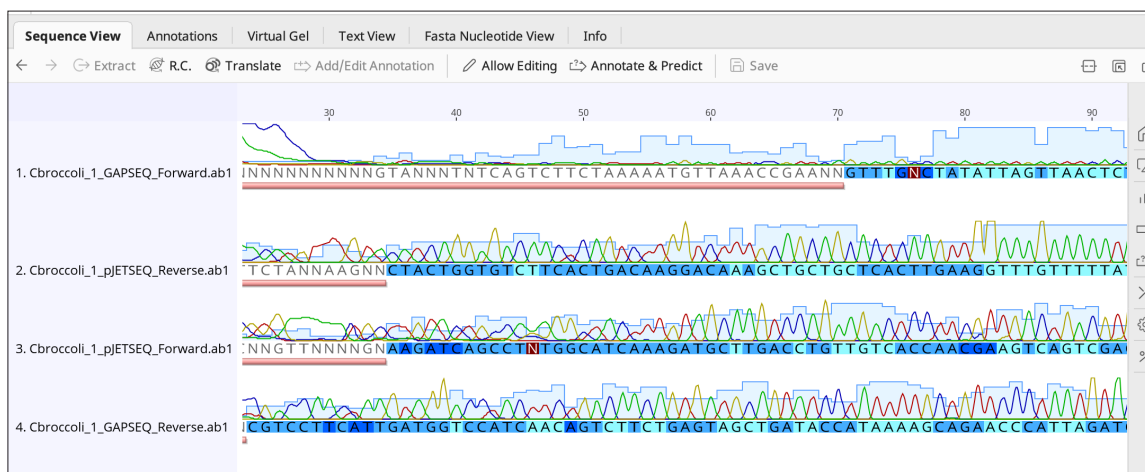
1.6.2.3 Click and drag the vertical bar to extend the light pink trim annotation and cover the Ns:

1.6.2.4 In the Sequence View toolbar, click **Save** to preserve the new trim region. The trim annotation now covers the problematic Ns:





Extending the trim annotation to cover the N base calls. The green bar marks the bases that are covered by the trim annotation as it is being dragged.

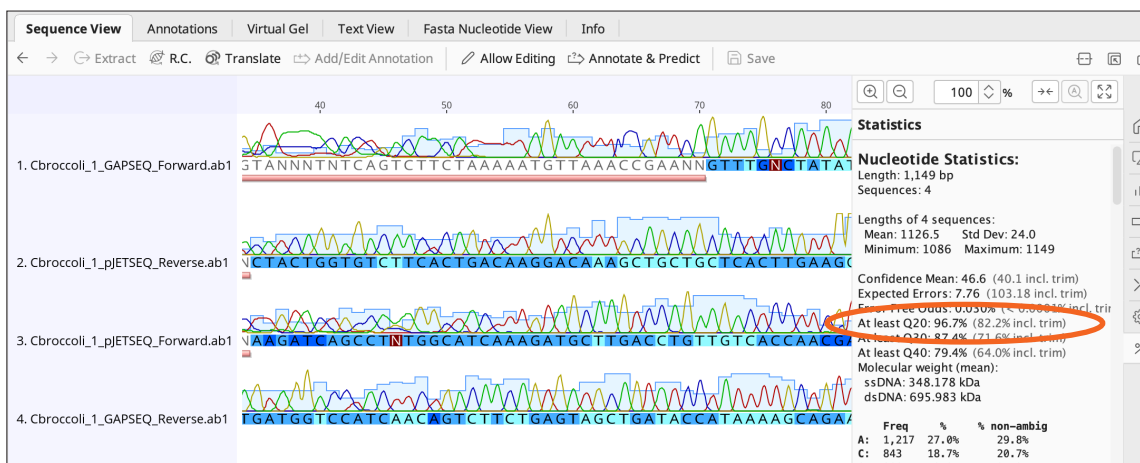
Note: If you prefer to permanently delete the low-quality regions rather than keep them as annotated trimmed regions that are hidden from downstream bioinformatics steps, click to highlight the annotation itself and press the Delete key on your keyboard. Note that this will permanently delete these bases from the original sequence files as well.



New trim region is preserved. The Save icon in the Sequence View toolbar will be grayed out once the new trim region is saved. When the annotation is selected, dark blue highlighting indicates the sequence covered by the trim region.

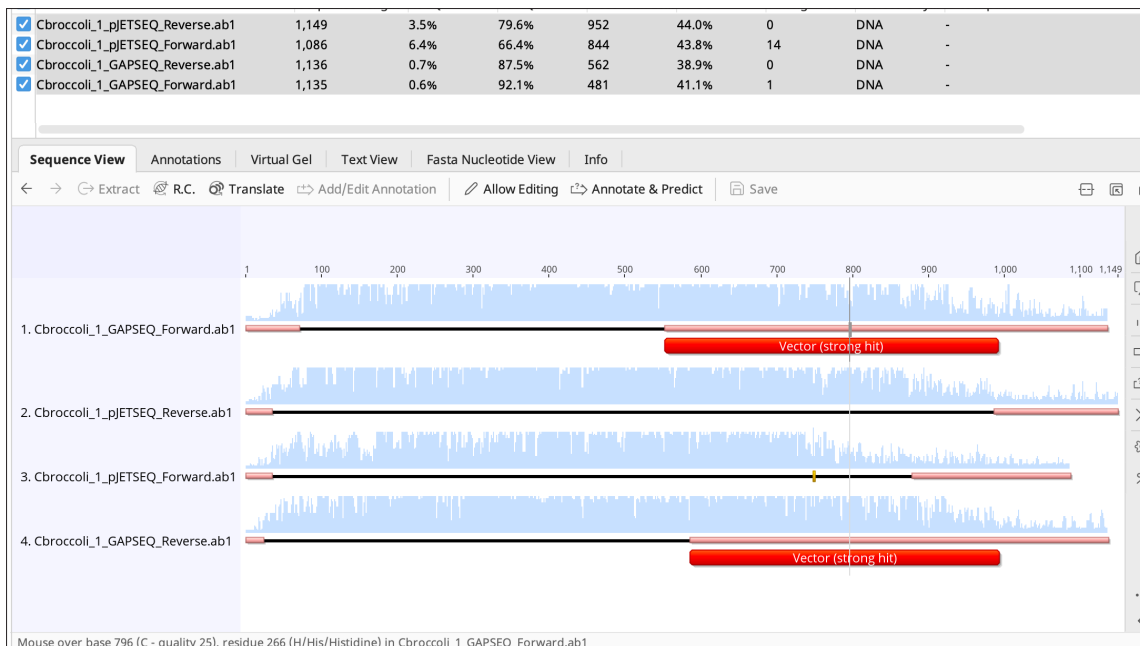
CHAPTER 9 PROTOCOL

- 1.6.2.5 Uncheck Allow Editing when you are finished.
- 1.6.2.6 If needed, perform the manual trimming steps for all four of your sequences. Note that manual trimming may delete some high-quality bases as well. It is a balancing act to trim as much poor-quality data as necessary while keeping as much high-quality sequence as possible.
- 1.6.3 Move the low-quality and vector matching chromatograms to the Deleted Items folder. Chromatogram files that have fewer than 50 bases with a quality score ≥ 20 should be moved to the Deleted Items folder  Deleted Items (under the Local folder in the Sources panel) so as to not confuse any of your good sequences with poor sequences.
 - 1.6.3.1 Check your sequences by viewing the Statistics tab  in the options panel. The length of your sequence will be at the top with the percentages of your sequence that have a minimum Q20, Q30, and Q40 scores listed underneath. In the example shown below, the sequence length is 1,135 bases with at least 83.4% (946.59 bases) of the trimmed sequence containing scores of at least Q20. Because this is more than 50 bases, this sequence will be kept for further analysis.



This sequence consists mostly of high-quality data, so it will be kept for further analysis.

- 1.6.3.2 Highlight the file names of any chromatogram files that have fewer than 50 bases of at least Q20 and drag them to the Deleted Items folder. For mostly vector matching sequences, restriction digest analysis of the minipreps should have screened out clones that did not have an insert but may occasionally miss clones of very small PCR products or a self-ligated vector. In cases like these, sequencing primers to the vector will still anneal, but the sequencing reaction will read right through to the other side of the plasmid vector. If the sequence is predominantly vector sequence, move the chromatogram to the Deleted Items folder.



Identifying sequences that contain mostly vector sequence. Sequences consisting mainly of vector sequence are not useful for further analysis and should be moved to the Deleted Items folder. The sequences from cbroccoli shown above contain more than 50 bases of high-quality, non-vector-matching sequences so they will be kept for further analysis.

1.7 Final assessment of the reads.

Your folder should now have reads that have high-quality sequence and are not predominantly vector sequence. If you have miniprep clones without any high-quality chromatograms, it is recommended that you do not proceed any further with your own sequences and instead work with clones that have high-quality data; either alternative clones of your own, ones from classmates, or the sample data provided. Please consult with your instructor to determine the best option.

1.8 Record the miniprep clone and team folder names.

For each of your miniprep clones, note the following file information and whether each chromatogram will be used for further analysis.

Miniprep clone name: _____

Geneious Prime folder name (and folder color, if assigned): _____

Sequence File Name	96-Well Plate Location	Sequencing Primer	Chromatogram To Be Used for Further Analysis? (Yes/No)
		pJET SEQ F	
		pJET SEQ R	
		GAP SEQ F	
		GAP SEQ R	

1.9 Results analysis of sequence data.

- Did you get data from the primers that anneal to the plasmid (pJET SEQ F and pJET SEQ R) and from the primers that anneal to the cloned *GAPDH* insert (GAP SEQ F and GAP SEQ R)?

2. If you did not get any good data from the primers that anneal to the cloned *GAPDH* insert (GAP SEQ F and GAP SEQ R), what might be the cause?

3. If other classmates started with the same plant as your group, did they get good data using the GAP SEQ F and GAP SEQ R primers?

The pJET SEQ F and pJET SEQ R primers are designed to anneal to the pJET1.2 plasmid vector on either side of the cloning site. Therefore, these primers should produce sequence irrespective of the gene cloned into them. The GAP SEQ F and GAP SEQ R sequencing primers are homologous to sequences within the *GAPC* gene itself. They are degenerate primers made to be a “best match” to conserved *GAPC* sequences of many plants — similar to the initial *GAPDH* PCR primers. Thus there is a chance that the primers will not match the *GAPC* gene of your plant and will not bind to the cloned DNA. If you and your classmates who worked on the same plant did not get any good data using the GAP SEQ F and GAP SEQ R primers, it may be that these primers did not match your sequence.

However, by working with other teams you may be able to use the sequences generated from the pJET SEQ F and pJET SEQ R sequencing primers to confirm your sequences. Moreover, if your *GAPC* gene is relatively short, you may still be able to assemble the sequences. More of this will be discussed in Section 2 when you will be assembling sequences.

1.10 Section 1 Focus Questions

1. If a base has a quality value of ≤ 20 , what might this tell you about the identity of the base?
2. What are the characteristics of a high-quality base?

2. Assemble the sequences and correct mistakes in the basecalls.

In this part of the procedure, you will be assembling your individual sequences to form one large contiguous sequence before comparing to sequences in a selected database (BLAST search).

Note: If you would like to perform the BLAST searches on your individual sequences before assembly, see Appendix H for instructions. This would be a great way to make a preliminary determination of which plant *GAPDH* genes most closely resemble the gene you cloned. This will also help you become familiar with BLAST and how to understand alignment data.

In nature, DNA molecules are found in a variety of sizes, many of them quite large. Even chromosome 21, the smallest human chromosome, is 47 million nucleotides in length. The smallest *Arabidopsis* chromosome is 18.5 million nucleotides long. Sanger DNA sequencing technology, however, rarely produces sequences longer than 1,000 bases. Consequently, it is possible to find the sequence of a longer piece of DNA only by reconstructing it from smaller pieces. This process is called assembly or sequence assembly. In this lab, we will use Geneious Prime to assemble sequence reads from the clones.

The steps used by many assembly programs, including Geneious Prime, are:

1. Compare all the sequences to each other.
2. Calculate a score for each pair of sequences.
3. Merge the sequence pairs together, working from the highest scoring pair to the lowest scoring pair until all possible pairs have been merged.

The contiguous sequences that result from merging shorter sequences are called **contigs**. A diagram of a contig is shown below. Some assembly programs, including Geneious Prime, can also use quality information, when available, to help guide the assembly process. If there are positions where the sequences disagree, these programs choose the higher quality base for the contig.



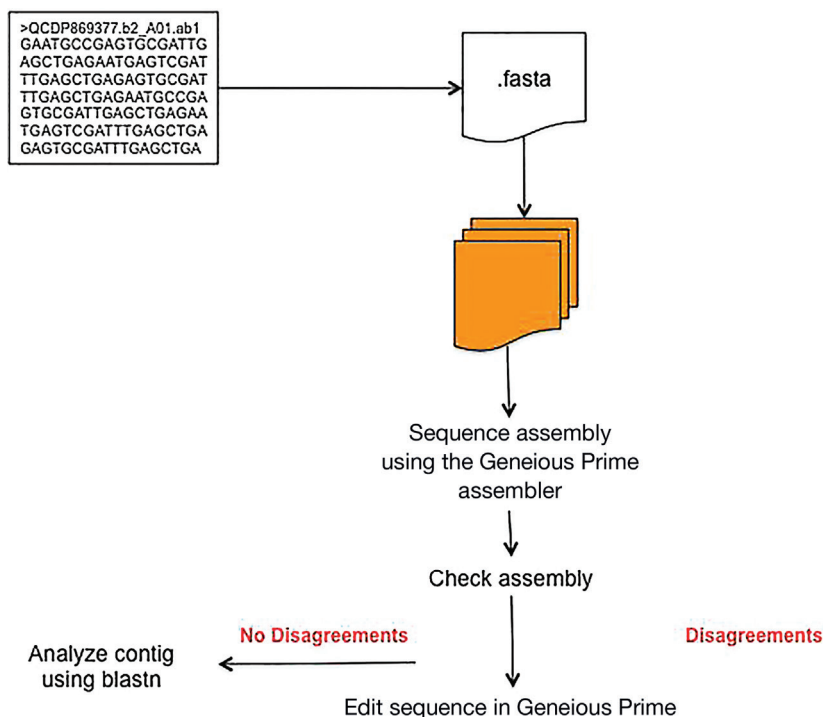
Generation of a contiguous sequence. Individual shorter sequences are compared, aligned, and assembled by programs such as Geneious Prime to generate longer contiguous consensus sequences. This methodology is used to generate continuous sequences that are longer than the current sequencing instruments are capable of generating in a single sequencing reaction.

In genome sequencing, the next step that occurs is called **finishing**. Finishing is a process in which researchers examine the contigs to look for misassemblies or regions that require additional coverage. That information may be used to identify errors, to edit sequences and reassemble them, or to synthesize new primers and generate additional sequences to cover gaps and put contigs together. In this lab, you will carry out the finishing step after you have assembled contigs.

Sequence assembly workflow

In this portion of the project, you will assemble your trimmed and edited sequences into a contiguous sequence called a contig. This assembly will be performed by the Geneious Prime aligner. The software looks for regions where the order of nucleotides is the same in the sequences (or are reverse complements of the sequences) and applies rules to determine whether the alignment is a valid one. If after alignment different sequences have different base calls for the same location, the Geneious Prime aligner uses quality values to generate the best contig. You will then need to review the assembly and determine whether there are any discrepancies, or **disagreements**, between the reads. If you see disagreements, you will need to review the chromatogram trace files, edit the traces if necessary, and reassemble the edited reads. The workflow for this portion of the project is outlined below.

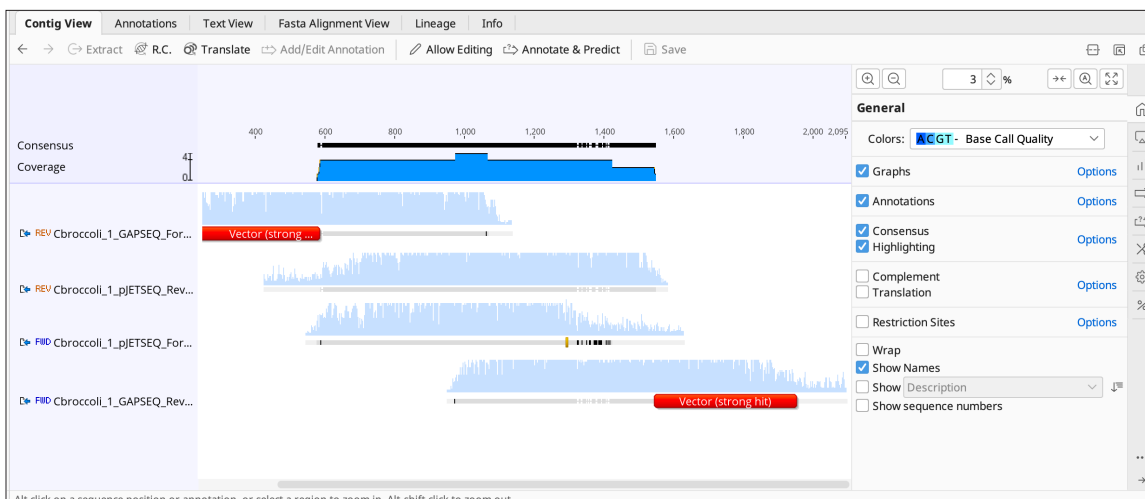
Sequence assembly workflow



Workflow for assembling sequences.

What is a contig and how is sequence assembly useful?

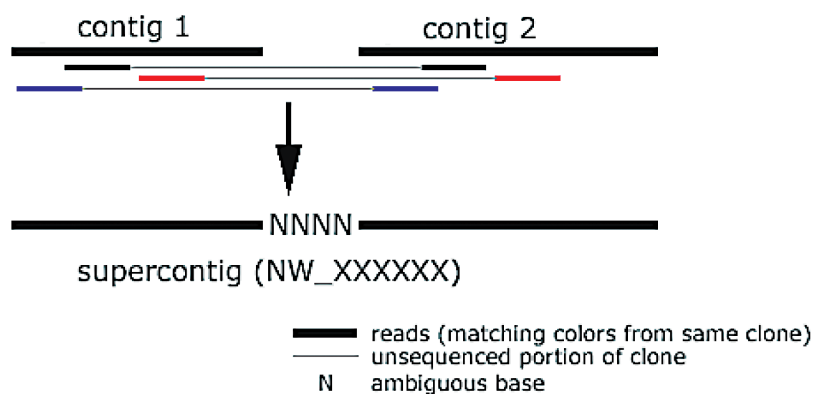
A **contig** (short for contiguous) refers to either a DNA segment or the reassembly of overlapping segments that form a continuous, extensive, and uninterrupted DNA sequence. By analyzing these segments, a researcher is able to discover the order of segments that make up various sequences. Contigs can be added, removed, or rearranged to form new sequences. Genomic contigs are connected to one another by the overlaps of matching sets of sequences.



Example of a contig assembled from four individual cbroccoli sequences.

Using this process of assembling DNA segments, contigs may be replicated or cloned into sequences, and segments may be added or eliminated.

Contigs can be assembled to form a **scaffold**. A scaffold is one contiguous length of genomic sequence in which the order of bases is known to a high confidence level. It is not uncommon to find



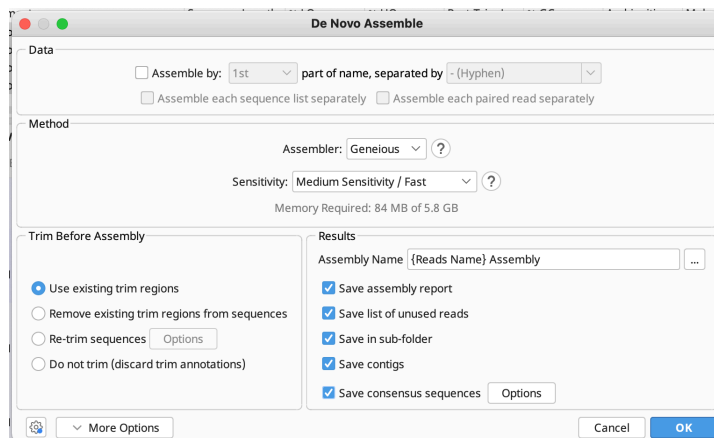
gaps within scaffolds. Gaps occur where reads from the two sequenced ends of at least one fragment overlap with other reads in two different contigs.

Assembling contigs. Since the lengths of the fragments are roughly known, the number of bases between contigs can be estimated.

2.1 Assembling your sequences in Geneious Prime.

2.1.1 Select all of the documents in your folder that you would like to assemble.

Click the Align/Assemble icon in the menu bar. Choose **De Novo Assemble** from the dropdown list. A new dialog box will appear:



- Leave the “Assemble by” box in the Data section unchecked
- Select Geneious as the Assembler (should be the default selection)
- Select **Medium Sensitivity / Fast** in the Sensitivity dropdown menu of the Method section
- Select **Use existing trim regions** for Trim Sequences since you have already trimmed your sequences

Note: If you have already permanently deleted the trimmed regions from section 1.6.2.4, the “Use existing trim regions” option will be grayed out and unavailable. In that case, select the **Do not trim** option instead, since you have already trimmed your sequences.

- Select all of the boxes in the Results section, which will generate a number of useful documents that you can use for troubleshooting purposes if required

Save assembly report — saves the document that records the fate of each sequence used for the assembly

Save list of unused reads — saves the document that lists all the sequences that failed to assemble into the contig

Save in sub-folder — saves all the results of the assembly to a new subfolder inside the folder containing the fragments. This folder will always contain only the assembly results from the most recent assembly; it creates a new folder each time it is run

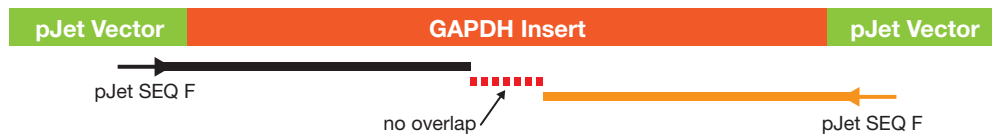
Save contigs — saves the assembly results as a contig

Save consensus sequences — saves the assembly results as a consensus sequence

Tip: It is good practice to save your results in separate subfolders to keep your data organized.

- Click **OK**. A new subfolder will be created with the term Assembly appended to the end of the name of your samples. For example, Cbroccoli_1_Assembly

- 2.1.2 Depending on the gene that you cloned, it is possible that there are sequences that could not be used in the assembly. For example, if you cloned a long gene and the GAP SEQ F and GAP SEQ R primers did not anneal well to your gene, then the sequence generated by the pJET SEQ F and pJET SEQ R primers might not overlap enough to enable their assembly. In these cases, you will see all sequences that cannot be assembled saved in a separate list in the Unused Reads document.



Potential causes of single sequences and no contigs. If the gene insert was too large or either the GAP SEQ F or GAP SEQ R primer did not anneal well to the insert, it is possible that the sequences cannot be assembled.

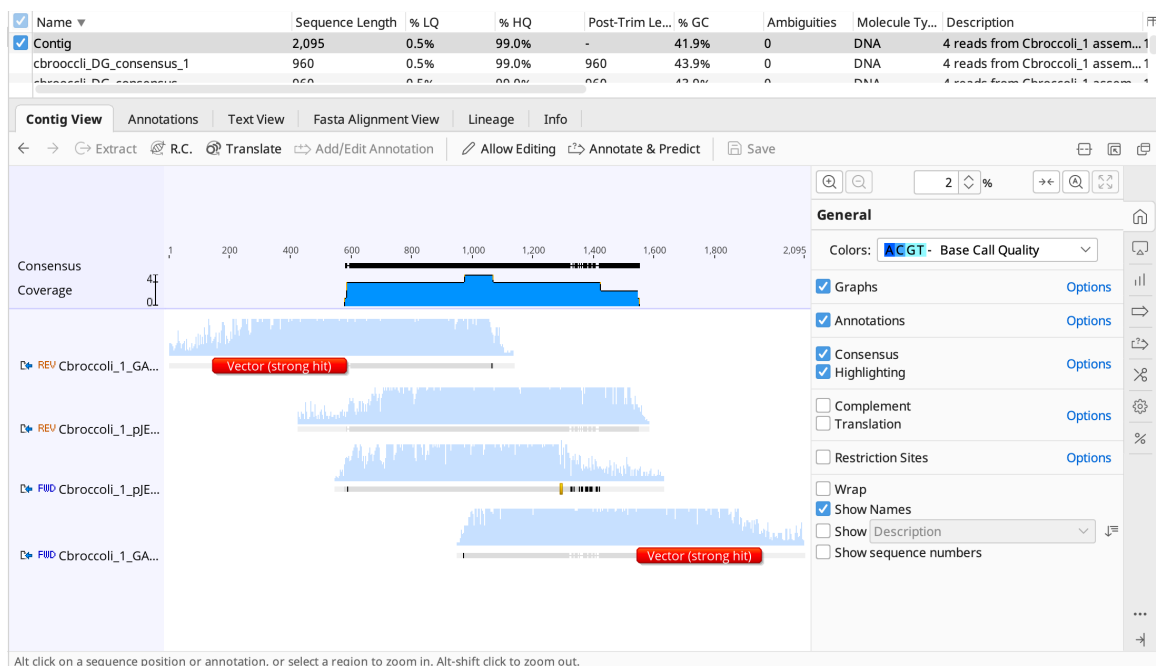
- If you obtained good sequence from all four sequencing primers, then you should be able to assemble all your read sequences into a single contig
- If you obtain multiple contigs, it is possible that the sequences you assembled did not really belong together or there wasn't enough overlap to assemble them. Lack of overlap may have occurred if the PCR product was very long, if the reads were relatively short, or if some of the sequencing primers did not yield data. Other explanations could be poor-quality data or mistakes by the assembly program. If you do have multiple contigs, pick one for further analysis


2.2 Viewing and understanding your contig.

The assembly program lines up the different reads of the sequence. Sequences overlap because you used different sequencing primers. Overlapping sequences means some reads will provide higher quality information for a portion of the sequence that has lower quality data in other reads. These overlaps must be analyzed to resolve any discrepancies in sequence between different reads. Sequencing errors may have occurred during the sequencing reaction, either by the detection instrumentation or due to incorrect base calls. The quality scores help determine where reads are more error-prone. Having multiple reads provides a consensus sequence that makes it more likely the actual sequence has been generated correctly. The term used for multiple sequences covering the same region is depth of coverage. A high depth of coverage would have multiple reads of the same sequence using different sequencing primers and, if possible, data from separately isolated clones.

2.2.1 To view your contig:

- 2.2.1.1 Click the subfolder for your assembly, and then click Contig. In the Contig View tab of the document viewer, you will see a zoomed out view of your contig:



Viewing the cbroccoli contig in Geneious Prime. Contig View allows you to see how your individual reads are assembled to form your contig. In this example, the cbroccoli contig, represented by the consensus sequence (black rectangle at the top), is generated from the four individual cbroccoli sequencing reads. Note that the quality-trimmed (pink bars) and vector-trimmed (red bars) are hidden and thus do not contribute to the contig. This view is zoomed all the way out (you can tell because the 'zoom out' icon  in the options panel is grayed out).

2.2.1.2 When you zoom in to your sequence you will be able to view:

- Consensus sequence at the top (black bar). By convention, the consensus sequence is shown in a 5' to 3' orientation
- Depth of coverage chart (blue) beneath the consensus
- The individual sequences that were assembled to form the contig

2.2.1.3 If a read is in the same orientation as the consensus sequence, Geneious Prime automatically denotes the read name with "FWD" (on the left-hand side of the sequence name). If the sequence of a read came from the other strand, Geneious Prime displays the read in the reverse direction, also called the reverse complement, and denotes the read name with "REV".



Example of a zoomed in view of the cbroccoli contig. The cbroccoli contig is zoomed in 71% in the middle of the sequence. Note that the direction of each read relative to the consensus sequence is indicated with a red REV or a blue FWD just before the name of each read.

- 2.2.1.4 If you obtained good sequence from all four sequencing primers, then all your read sequences should be assembled into a single contig.
- 2.2.1.5 If you obtained multiple contigs, it is possible that the sequences you assembled did not really belong together or there wasn't significant enough overlap to assemble them. Lack of overlap may have occurred if the PCR product was very long, the reads were relatively short, or if some of the sequencing primers did not yield data. Other explanations could be poor-quality data or mistakes by the assembly program. If you obtained multiple contigs, you can choose to perform one of the following:
 - Pick the longest contig for further analysis
 - OR
 - Try putting the multiple contigs back into the Geneious Prime assembler. They may now form a single assembly. Alternatively, use untrimmed rather than quality-trimmed sequences. However, this method will introduce many more discrepancies into your assembly

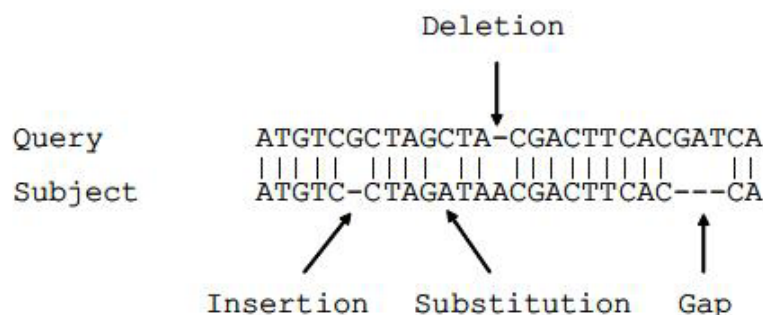
2.2.1.6 If you do not see any contigs, these are sequences that could not be assembled, most likely for reasons described in section 2.1.2 and its accompanying figure. In this case, assemble your single sequences with the sequences from other student teams who worked on the same plant.

- Create a new subfolder in the Local folder in Geneious Prime (see section 1.4), and move your single sequence files and files of the same plant from other student teams to the folder. Perform an assembly as described in 2.1. Multiple contig assemblies may result, since the orientation of the cloned insert could be forward or reverse, or the specific *GAPC* gene cloned may be different between student teams and individual plasmid clones. If this is the case, each team can work on a different contig. While a complete sequence of the PCR product may not be obtained, the benefit of this approach is that the sequence data can be checked for errors and put through a finishing process just like an assembled contig, providing more confidence in the data, which can still be submitted to GenBank

2.3 Identify and correct mistakes in the base calls.




The power of sequencing in both directions and using multiple primers to generate a contig is the ability to generate a consensus sequence with the best possible probability of it being the correct sequence. There may still be some ambiguities or differences in the various sequences, so it is important to look at these differences and study the traces from each sequence that contributed to the contig in those regions and determine what the best sequence should be. Ideally, with good clipping of low-quality bases, there should be few differences that need to be examined.

When an individual sequence conflicts with the consensus, this is referred to as a **discrepancy**. In Geneious Prime, a discrepancy is called a **disagreement**. Disagreements can come in many forms. They might be a substitution, for example when one read identifies a base as an A and another read identifies it as a C. Another kind of disagreement is called an indel, which means either an insertion or a deletion. Indels are important because they can change the reading frame and make it more difficult to predict the correct protein sequence.



To determine which read is correct, you will need to review the chromatogram traces. In this part of the project, you will use Geneious Prime to review the assembly results and resolve disagreements between reads. If you find a mistake in the consensus sequence, you will need to edit the reads and generate a new assembly and a new contig. This must be done manually and is often an iterative process.

2.3.1 Navigate through the disagreements.

- 2.3.1.1 Click to select your contig from the document table and view it using the Contig View tab of the document viewer. To make sure you are starting at the beginning of your sequence, click the “zoom out to full view” icon  in the options panel to see your entire chromatogram. Place your cursor at the beginning of your sequence and zoom in, using the magnifying glass icon .
- 2.3.1.2 In the Display tab , make sure that the box for Highlighting is selected. Click options and then click the arrowhead pointing to the right. This will quickly move you from disagreement to disagreement so that you can examine the base calls and decide what you consider to be the appropriate edits based on the other sequencing reads in the contig. You may have to delete bases and adjust gaps.



Example of a disagreement found in a contig. In this example, the pJETSEQF read for cbroccoli calls an ambiguous base for position 1,292 of the consensus sequence, whereas the other two reads call the base as a T. The base call of N in the pJETSEQF read is called a disagreement relative to the consensus sequence.

Looking at the trace above for the cbroccoli contig, it is pretty clear that one of the reads contained an error and the base that has an N is read as a T in the other three chromatograms, all of which have higher quality scores for base calling. Since the consensus sequence contained the base T at position 1292 and this was correct, you will not need to edit this read.

Tip: As with individual alignments, it is possible to edit the trims in the contig. If you choose to edit the trimmed regions, the consensus will change accordingly. You may decide to edit the consensus sequence directly and the changes will then be applied to all bases in the column below. You may also have Geneious Prime apply your changes back to the source sequence documents.

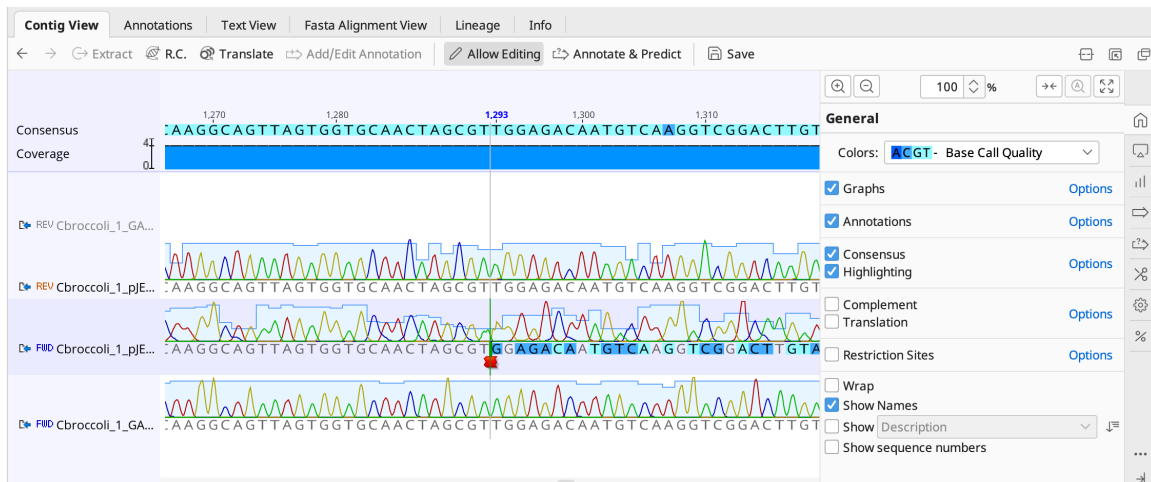
2.3.1.3 To edit a base call:

- Identify the problematic base. In the example above, there appears to be an N in position 1,293 for the middle read whereas the base call is a T for the other two reads. Since the N base call does have a green peak in the chromatogram, and the other two reads have higher quality scores for base-calling, it is most likely that this N should be a T
- In the Contig View toolbar, click the Allow Editing button. Place your cursor to the right of the target base



Editing a disagreement. To edit the N to a T in the middle read, click the **Allow Editing** button in the Contig View toolbar and place your cursor just to the right of the base.

- Click backspace or delete to remove the base. A small red line will appear to mark where you have deleted the base.



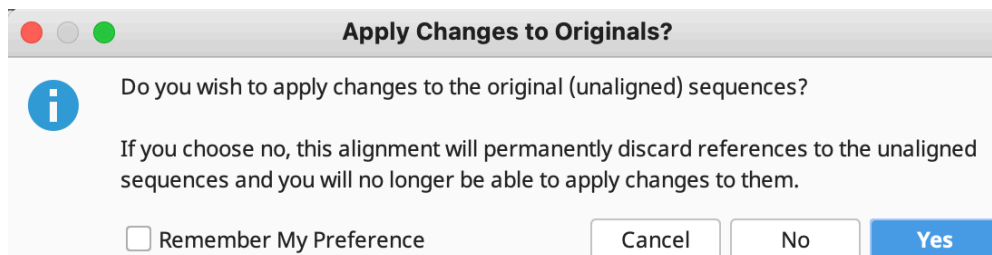
Editing a base call. The N base has been deleted. This change is now marked with a small red line where the base used to be.

- Type in the new base. A yellow line will appear under the new base call to indicate that the base call has been edited

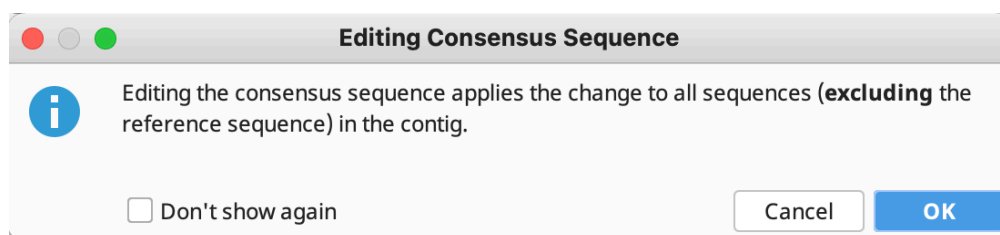


The edited base call. The N base has been changed to a T in the middle read. This edit is now marked with a yellow line beneath the T.

- Click **Save** in the Contig View toolbar. A dialog box will appear. Click **Yes** to save and apply changes



- The change will also be added to the Annotations and Tracks tab ⇨ in the options panel as a Replacement
- If you edit a base in the consensus rather than the individual sequence, a dialog box will appear. You can click **OK** to accept the changes



2.3.2 Reassemble edited reads. If you edited bases in single sequences, you will need to carry out another assembly to get a corrected consensus sequence.

- Repeat the following steps from Section 2.1 above. Your results will be contained in a new Assembly folder that will have the same name, but with a number appended to the end. For example, cbroccoli_1_DG_Assembly 2. Your new contig will be named Contig within the new folder
- Check your new assembly and determine whether there are any disagreements. If there are, repeat the steps for editing a base call (step 2.3.1.3), and continue to iterate until there are no more disagreements. This could take several rounds.

2.3.2.1 Record the name of your finished (fully corrected), final contig and its folder name below

Record the name of your final Assembly folder:

Record the name of your final contig file: _____

Plant name: _____

Clone number: _____

2.4 Results Analysis.

1. What was the length of each of your single reads (that is, once poor-quality data and vector sequences were trimmed away)?

2. What is the length of your contig sequence after editing?

3. Which of your sequences were assembled in the forward direction and which were assembled in the reverse direction? For each of these sequences, which sequencing primer was used to generate the sequence?

- ## CHAPTER 9 PROTOCOL

2.5 Section 3 Focus Questions.

- Bioinformatics

3. Conduct a BLAST search on the contig sequence to verify identity of the cloned gene.

Now that you have assembled all of your sequences into one contiguous sequence and made corrections to get the best consensus sequence, you will be determining which sequence in the **GenBank genomic DNA database** most closely resembles the consensus sequence you generated.

Biological sequences have evolved over time from common ancestors. Comparing a sequence with other known sequences, using an inexact alignment method to find potential relatives, will help you identify the function of an unknown or new sequence. BLAST (Basic Local Alignment Search Tool) programs are designed to find short (local) regions where pairs of sequences match. A blastn search compares a query sequence in turn to each sequence in a nucleotide sequence database. The result of a blastn search will be a set of matching and potentially related sequences ranked according to similarity.

Here, blastn will be used to compare your contig consensus sequence to the **GenBank database of all genomic nucleotide sequences**. Once the search is complete, blastn counts all the nucleotides in the matching regions and awards two points for every pair of bases that match. If one sequence has an insertion, a deletion, or a gap (more than one base missing) and the other does not, BLAST deducts points from this score. The net result is that a blastn score is more or less twice the length of the matching region, depending on how many points were deducted.

The completed search will return a bit-score and an E value for each match of your query sequence to a sequence in the GenBank database. The results also include an alignment of your sequence to each match in the database so that you can compare them. The meaning of the scoring will be explained in more detail in the section titled Interpreting your blastn results.

IMPORTANT NOTE regarding BLAST searches using Geneious Prime: In general, the amount of time it takes to retrieve BLAST results will vary depending on the how many searches NCBI BLAST is asked to run at that moment from researchers around the world. In some cases, searches performed through Geneious Prime are not as fast as performing the BLAST searches directly from NCBI.

If you have short class periods (50 min or less) and need the BLAST results as soon as possible to enable data analysis, you have the option of performing the BLAST searches directly from NCBI's website for this section.

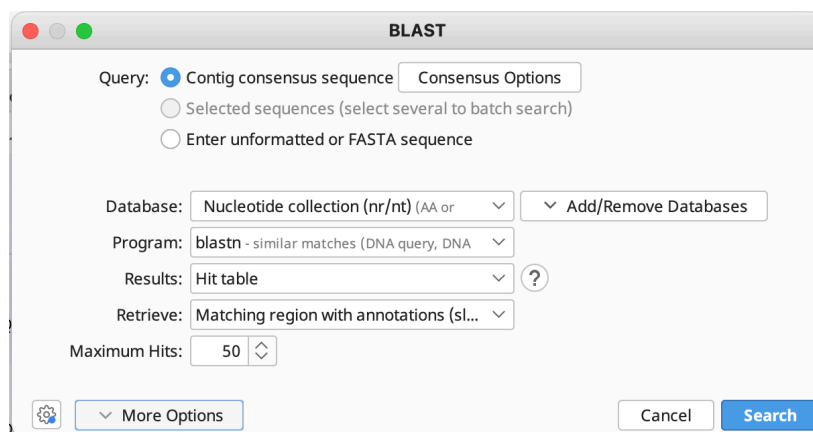
Please consult your instructor on whether you will be performing Section 3.1 using Geneious Prime or using NCBI's BLAST website. Use Appendix G for protocol steps to export sequences as FASTA files for BLAST searching directly on the NCBI website.

3.1 Use Geneious Prime to perform a BLAST search on your contig sequence against the NCBI sequence database.

In this part of the procedure, you will use blastn to compare your contig sequence to the sequences in a selected database.

3.1.1 In your Assembly folder, click to select your final, corrected contig.

3.1.2 At the top of the menu bar, select the BLAST icon. A new dialog box will appear.



The software will automatically select the Contig Consensus Sequence radio button for Query. Keep this selection.

- Select the **Nucleotide collection (nr/nt)** for Database
- Select **blastn** for Program
- Select **Hit table** for Results
- Select **Matching region with annotations (slow)** for Retrieve
- Set Maximum Hits to **50**
- Click **Search**. A new subfolder, named “Contig – Nucleotide collection (nr_nt) blastn,” will be created to contain your search results
- If a new window opens with the title Search Failed, read the description to learn the cause. In this case, the connection to NCBI was not reliable. In some cases, restarting Geneious Prime and rerunning the BLAST will work. See Appendix G for instructions for how to run a BLAST search directly from the NCBI website. If repeated tries do not work, contact Bio-Rad Technical Support for help.
- If your Contig Consensus Sequence contains ambiguities such as uncalled bases or undetermined bases, a notification may appear. Proceed with the BLAST search. These ambiguities will be considered as mismatches by the BLAST search algorithm and will be noted in the results.

3.2 Analyzing your blastn results.

In this section, you will learn about what the blastn results mean and examine the genes that your contig query sequence matches best. You will then calculate a total homology score for the best matches with four *Arabidopsis* *GAPC* genes (*GAPC*, *GAPC-2*, *GAPCP-1*, and *GAPCP-2*). An example of calculated homology scores using data for Chinese broccoli from the cbroccoli folder is shown in the following sections.

3.2.1 Understanding the blastn results using the Hit Table.

The results from a blastn search include many different kinds of information and statistics. These bits of information include the size of the database, length of each query sequence, statistics that describe the number and percent of matching bases, a BLAST score, and the E value.

The sequences in the example shown below come from a *GAPDH* cloning experiment with a plant from the genus *Brassica* (cabbage). Your results may differ from those shown in this manual.

On the Hit Table tab of the document table, you will find summary statistics for the search results. The **E value** indicates the expected frequency of an alignment's occurrence by chance. The smaller the E value, the better the match.

In addition to the E value, there is also a column labeled **% Pairwise Identity**. (You may need to scroll sideways to find this column.) Drag this column over next to the E value. It is also useful as it will indicate how similar the sequence found in the database is to the one you used as a query. Note that sorting by E value and % Pairwise Identity can produce a different ordering because statistical significance is related to alignment length as well as identity, but identity relates only to the aligned region. For example, consider the alignment in the figure below. In this alignment, the % Pairwise Identity is 96% with *Oryza sativa* (rice). However, when you examine the matching regions in more detail, you find that the region where 96% of the bases match is only 28 nucleotides long. This is a good example of how short sequences can give a good match that is not meaningful.

```

Features in this part of subject sequence:
  hypothetical protein

Score = 46.4 bits (50), Expect = 0.007
Identities = 27/28 (96%), Gaps = 0/28 (0%)
Strand=Plus/Minus

Query  6          AGCCTTGGCATCAAAGATGCTCGACCTG  33
      |||||
Sbjct  23549447  AGCCTTGGCATCAAAGATGCTGGACCTG  23549420
  
```

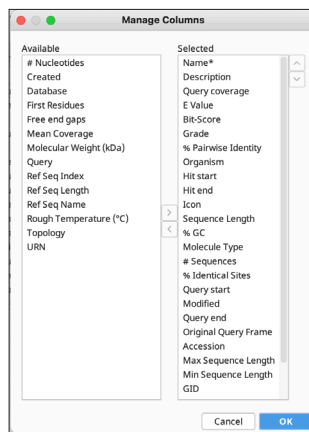
Alignment of rice sequence with query sequence. This is a sequence that has a high % Pairwise Identity score (96%). However, it is a very short sequence so the match is not meaningful.

Tip: It is useful to look at alignments in isolation, but looking at alignments together reveals more information. The Geneious Prime view called **Query Centric View** provides a multiple alignment–style visualization of the BLAST hits mapped against the original query sequence. This isn't a true multiple sequence alignment, but instead a mapping of the individual BLAST hits against the query sequence. It is a useful way to see where the conserved regions in the BLAST search are lining up against the query. Keep in mind that BLAST alignments are local alignments.

Name	Organism	Description	E Value	Bit-Score	Original Query Frame	Grade	% Pairwise L.	Quid
LR699747	Arabidopsis thaliana	Arabidopsis thaliana genome assembly, chromosome: 3	0	719.467	1	90.0%	80.6%	99.
M64119	Arabidopsis thaliana	Arabidopsis thaliana cytosolic glyceraldehyde-3-phosphate deh...	0	702.847	1	89.9%	80.3%	99.
LT669788	Arabidopsis thaliana	Arabidopsis thaliana genome assembly, chromosome: 1	0	708.387	1	88.8%	80.9%	96.
LT669790	Arabidopsis thaliana	Arabidopsis thaliana genome assembly, chromosome: 3	2.38e-154	558.808	1	71.7%	84.8%	58.
HG994355	Brassica napus	Brassica napus genome assembly, chromosome: A01	0	878.279	1	87.3%	85.5%	89.
HG994355	Brassica napus	Brassica napus genome assembly, chromosome: A01	0	830.266	1	86.5%	84.7%	88.
HG994357	Brassica napus	Brassica napus genome assembly, chromosome: A03	0	1,099.88	1	89.6%	90.0%	89.
HG994359	Brassica napus	Brassica napus genome assembly, chromosome: A05	0	952.145	1	87.9%	87.0%	88.
HG994360	Brassica napus	Brassica napus genome assembly, chromosome: A06	0	656.681	1	73.9%	87.1%	60.
HG994362	Brassica napus	Brassica napus genome assembly, chromosome: A08	0	689.921	1	82.9%	82.6%	83.
HG994363	Brassica napus	Brassica napus genome assembly, chromosome: A09	0	798.873	1	91.0%	82.5%	99.
HG994367	Brassica napus	Brassica napus genome assembly, chromosome: C03	0	1,461.82	1	96.9%	94.2%	99.
HG994369	Brassica napus	Brassica napus genome assembly, chromosome: C05	0	981.691	1	88.2%	87.5%	88.
HG994369	Brassica napus	Brassica napus genome assembly, chromosome: C05	0	665.914	1	69.7%	94.5%	45.
HG994372	Brassica napus	Brassica napus genome assembly, chromosome: C08	0	680.687	1	82.8%	82.4%	83.
HG994372	Brassica napus	Brassica napus genome assembly, chromosome: C08	2.38e-154	558.808	1	72.6%	84.5%	60.
JN571725	Brassica oleracea	Brassica oleracea glyceraldehyde 3-phosphate dehydrogenase...	6.61e-155	560.655	1	72.6%	84.5%	60.
JQ659190	Brassica oleracea	Brassica oleracea glyceraldehyde 3-phosphate dehydrogenase...	0	715.774	1	70.4%	98.1%	42.

Example of BLAST results in a Hit Table

3.2.2 Summary of the major categories on the Hit Table. Remember that you may have to scroll to the right to find these columns, or they may not be automatically displayed at all. There are many columns that can be displayed or hidden. To display/hide additional column headers, choose the icon that looks like a small data table just above the scroll bar on the right (the pop up bubble will tell you this icon is called “Change the visible columns”). This will reveal all the column options that are available.



Manage Columns lets you choose to view the data that will be most helpful to you. Click the small data table icon, then select Manage Columns from the list. A dialog box will open in which you can select your options.

Name: A sequence's name is its accession number, which is the unique identifier of your sequence within a database. The main public database that is used for storing and distributing sequence data is NCBI's GenBank. Accession numbers are also used to report sequences in scientific papers and journals.

Description: A brief description of hit matches, including the scientific name of the organism and the chromosomal location if known.

Query coverage: The length of your sequence covered by the one found with the BLAST search.

E Value: The expect (E) value is the number of hits one can expect to see by chance when searching a database. The size of the database being searched will affect E value calculations. For example, an E value of 1 means that in a database of the current size one might expect to see one match with a similar score purely by chance. The E value describes the random background noise in a match, so it decreases exponentially as the score (S) (see Bit-Score), the assessment of an alignment's overall quality, of the match increases.

Generally, the closer the E value is to zero, the more significant the match. An exception is that virtually identical short alignments have relatively high E values because shorter sequences have a higher probability of occurring in a database purely by chance.

Tip: The E value can be a convenient way to create a significance threshold for reporting results. You can change the E value threshold within Geneious Prime easily. Raising the E value threshold will produce a longer list, but more of the hits will have low scores.

Bit-Score: The score (S) describes the overall quality of an alignment; higher values correspond to greater similarity. The bit-score takes the statistical properties of the scoring system into account to normalize an alignment's score (S). Every search is unique, but a bit-score allows alignment scores (S) from different searches, which may have been conducted with different algorithms using different values, to be compared.

Grade: A percentage calculated by Geneious Prime by weighting the query coverage, E value, and identity value (0.5, 0.25, and 0.25 respectively) for each hit. This allows you to sort hits so that the longest, highest identity hits are at the top.

% Pairwise Identity: This is the value, expressed as a percentage, of how similar two sequences (nucleotide or amino acid) are.

% Identical Sites (PID): The percentage identity for two sequences can be variable and depends on several factors. The alignment method and parameters used to compare the sequences will affect the sequence alignment. PID is strongly length-dependent, which means that the shorter a pair of sequences is, the higher the PID you might expect by chance.

3.2.3 Looking at the alignments in the document viewer panel

Now that you have a set of search results, you will need to look at some alignments. You can click any sequence in the Hit Table list and Geneious Prime will display the pairwise alignment for that hit in the document viewer window.


3.2.3.1 Click on one BLAST result from the Hit Table (let's say one with a description of *Arabidopsis thaliana*) and look in the Alignment View tab in the document viewer. You will see a zoomed-out view of the query aligned to the BLAST hit. From top to bottom in the document viewer you will see:

- The consensus sequence, displayed on top
- The depth of coverage chart, displayed in blue. If the depth of coverage chart is not visible, go to the Graphs tab in the options panel and check the box labeled Coverage
- A graphical representation of % pairwise identity, displayed in green
- Your query sequence, then the sequence for the hit


The screenshot displays the Geneious Prime interface. At the top, the 'Hit Table' is visible, showing search results for an Arabidopsis thaliana query. The table includes columns for Name, Organism, Description, E Value, Bit-Score, % Pairwise Identity, Query coverage, Grade, and Original Query. The first hit, M64119, is selected. Below the table, the 'Alignment View' tab is active, showing a detailed view of the alignment between the query and the hit. The alignment view includes a consensus sequence, a depth of coverage chart (blue), a % pairwise identity chart (green), and the query and hit sequences. On the right side, the 'Annotations and Tracks' panel is visible, showing various genomic features such as CDS (1), Exon (5), Gene (1), Intron (4), mRNA (1), and Source (1). The interface also includes a search bar, a zoom slider, and a 'Download Full Sequence(s)' button.

Hit Table	Query Centric View	Annotations	Distances	Fasta Alignment View	Info				
<input checked="" type="checkbox"/>	Name	Organism	Description ▲	E Value	Bit-Score	% Pairwise I...	Query cover...	Grade	Original QueF
<input checked="" type="checkbox"/>	M64119	Arabidopsis thaliana	Arabidopsis thaliana cytosol...0	702.847	80.3%	99.48%	89.9%	1	
	LR699747	Arabidopsis thaliana	Arabidopsis thaliana genom...0	719.467	80.6%	99.48%	90.0%	1	
	LT669788	Arabis alpina	Arabis alpina genome asse... 0	708.387	80.9%	96.77%	88.8%	1	
	LT669790	Arabis alpina	Arabis alpina genome asse... 2.38e-154	558.808	84.8%	58.65%	71.7%	1	
	HG994355	Brassica napus	Brassica napus genome ass...0	878.279	85.5%	89.17%	87.3%	1	
	HG994355	Brassica napus	Brassica napus genome ass...0	830.266	84.7%	88.33%	86.5%	1	
	HG994357	Brassica napus	Brassica napus genome ass...0	1,099.88	90.0%	89.27%	89.6%	1	


Alignment view for one BLAST hit. An *Arabidopsis thaliana* query result is chosen for comparison with the contig generated from cbroccoli. In Alignment View, you can see information such as the consensus sequence, depth of coverage, identity chart, and known annotations from *Arabidopsis thaliana*.

- 3.2.3.2 In the Annotations and Tracks tab  in the options panel, check to make sure that the box for Show Annotations is checked. This way, you will see whether there are any *GAPC* genes within this *Arabidopsis* chromosome.

Tip: Hovering over the annotations will display a pop up with more information, including the gene name and gene product, if known. You can select text from within this popup and copy the text into another file or an electronic lab notebook.

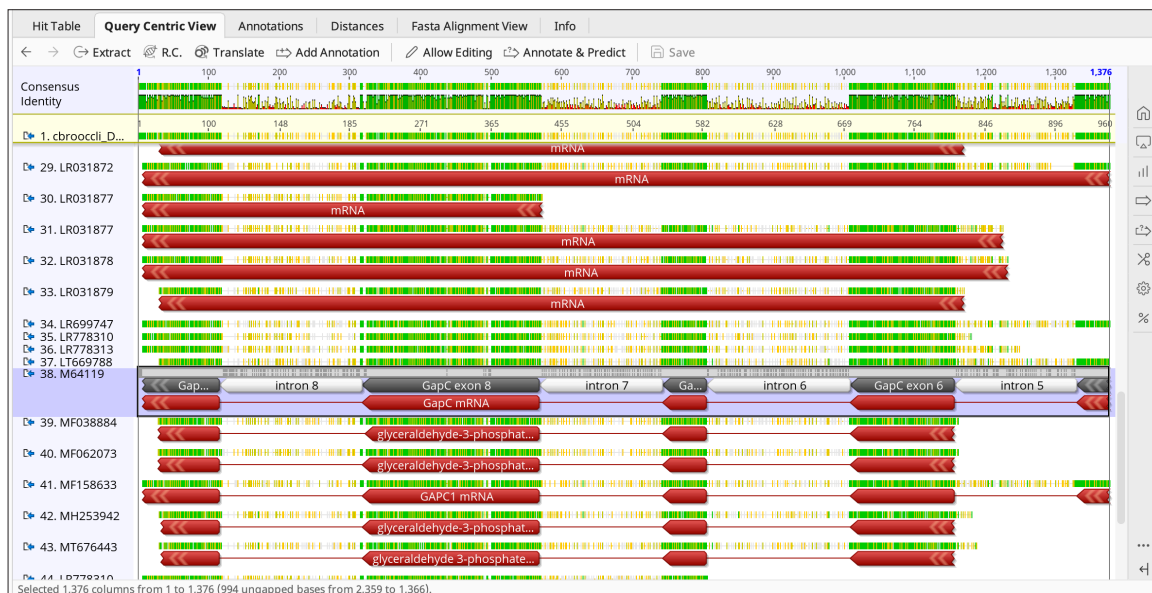
- 3.2.3.3 A consensus sequence is an alignment that occurs, with minor variations, across many genetic locations or organisms. It is constructed from the order of the nucleotides appearing most frequently at each position of a sequence alignment. The consensus is the same length as the contig (which includes only untrimmed bases), and shows which nucleotides are conserved and which are variable. For a nucleotide to be selected for the consensus, it must reach a minimum threshold of occurrence in that position in a variety of sequences. The consensus sequence is available when viewing alignments or contig documents, and is displayed when the box for Consensus is checked in the General tab  in the options panel.

Tip: Ambiguity codes, such as an R designation for a nucleotide that could be either an A or a G, are counted as fractional support for each nucleotide in the ambiguity set (A and G, in this case). Thus, two rows with Rs are counted the same as one row with an A and one row with a G.

Tip: When “Ignore gaps” is checked (in the Display tab  of the options panel), the consensus is calculated as if each alignment column consisted only of the non-gap characters. Otherwise, the gap character is treated like a normal nucleotide. However, mixing a gap with any other nucleotide in the consensus always produces the total ambiguity symbol (N for nucleotides and X for amino acids).

- 3.2.3.4 Depth of coverage represents the number (often an average) of nucleotides contributing to a portion of assembled sequences. On a whole-genome basis it means that each base has been sequenced, on average, a particular number of times (for example, 10x, 20x, etc.). For a specific nucleotide, it represents the number of reads that contributed information about that nucleotide. The depth can vary depending on the genomic region being sequenced. In the figure above (the single-hit alignment to *Arabidopsis*) the depth of coverage across the contig is 2.

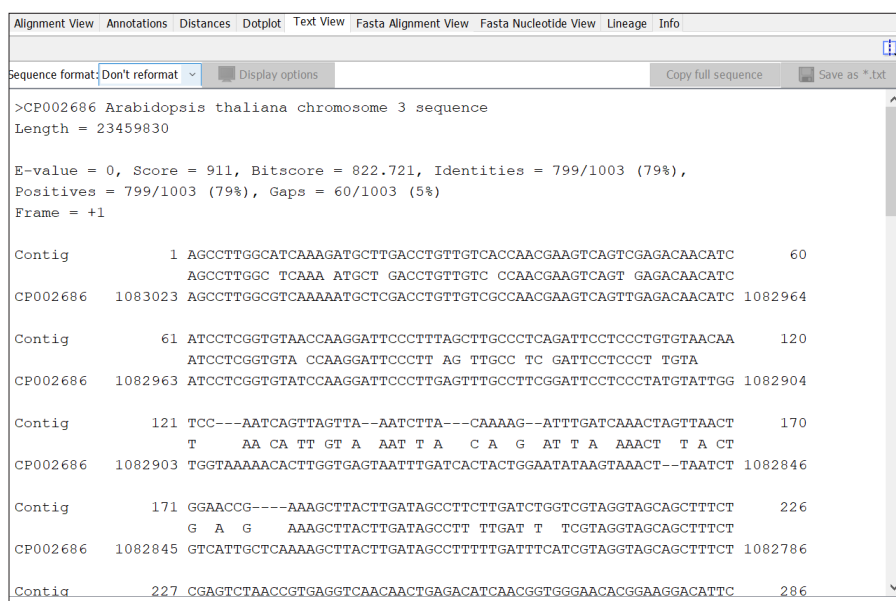
- 3.2.4 Click the Query Centric View tab in the document table. You will see all the hits with annotations aligned to your query sequence displayed. This gives you a quick survey of how many hits have annotations for the *GAPC* family of genes:



Viewing the BLAST query results in Query Centric View.

3.2.5 Select a result from the Hit Table and navigate to Text View in the document viewer. Select Don't reformat from the Format drop-down field. You can use the Text View tab to take a quick look at the information from the BLAST hit.

- At the top of the page is the accession number and the description of the BLAST hit
- Beneath that is a summary of the statistics for the hit, such as length, bit-score, E value, etc.
- Next is the alignment in text format. Your query is located at the top, followed by the consensus sequence in the middle and the BLAST hit on the bottom
- The numbers at the ends of the sequence refer to the gene's location on the chromosome
- No letter means there is no match between the query and the BLAST hit
- A dash means a gap



```

>CP002686 Arabidopsis thaliana chromosome 3 sequence
Length = 23459830

E-value = 0, Score = 911, Bitscore = 822.721, Identities = 799/1003 (79%),
Positives = 799/1003 (79%), Gaps = 60/1003 (5%)
Frame = +1

Contig          1 AGCCTTGGCATCAAAGATGCTTGACCTGTTGTACCAACGAAGTCAGTCGAGACAACATC          60
                  AGCCTTGGC TCAAA ATGCT GACCTGTTGTC CCAACGAAGTCAGT GAGACAACATC
CP002686    1083023 AGCCTTGGCGTCAAAAATGCTCGACCTGTTGTGCCAACGAAGTCAGTTGAGACAACATC 1082964

Contig          61 ATCCTCGGTGTAACCAAGGATTCCTTTAGCTTGCCCTCAGATTCCCTCCCTGTGTAACAA        120
                  ATCCTCGGTGTA CCAAGGATTCCTT AG TTGCC TC GATTCCTCCCT TGTA
CP002686    1082963 ATCCTCGGTGATCCAAGGATTCCTTTAGCTTGCCCTCGGATTCCTCCCTATGTATTGG 1082904

Contig          121 TCC---AATCAGTTAGTTA--AATCTTA---CAAAG--ATTGATCAAAGTAACT          170
                  T   AA CA TT GT A AAT T A   C A G AT T A AACT T A CT
CP002686    1082903 TGGTAAAAACACTTGGTGAGTAATTGATCACTACTGGAAATATAAGTAACT--TAATCT 1082846

Contig          171 GGAACCG---AAAGCTTACTTGATAGCCTTCTTGATCTGGTCGTAGGTAGCAGCTTTCT        226
                  G A G   AAAGCTTACTTGATAGCCTT TTGAT T TCGTAGGTAGCAGCTTTCT
CP002686    1082845 GTCATTGCTCAAAAGCTTACTTGATAGCCTTTTGTATTCATCGTAGGTAGCAGCTTTCT 1082786

Contig          227 CGAGTCTAACCGTGAGGTCAACAACAGACATCAACGGTGGAACACGGAAGGACATTC        286
  
```

Text view of the Cbroccoli contig sequence compared with the *Arabidopsis thaliana* BLAST query result. The alignment between the Cbroccoli contig and the *Arabidopsis* sequence is displayed in text view. The Text View was expanded by clicking the Expand Viewer button  in the options panel, which hides the Sources panel on the left and the document table at the top of the Geneious Prime window.

3.3 BLAST searching the contig and recording matches.

For your contig, sort according to % Pairwise Identity and record the top three matches and their statistics in the charts below.

Contig Sequence

Description	E Value	Bit-Score	% Pairwise Identity	Query Coverage	Grade

3.4 Verification of sequence.

It is possible that PCR products have been generated as a result of contamination by the control PCR reactions. To verify that your sequence is not an *Arabidopsis* gene, it is necessary to compare it against the BLAST results.

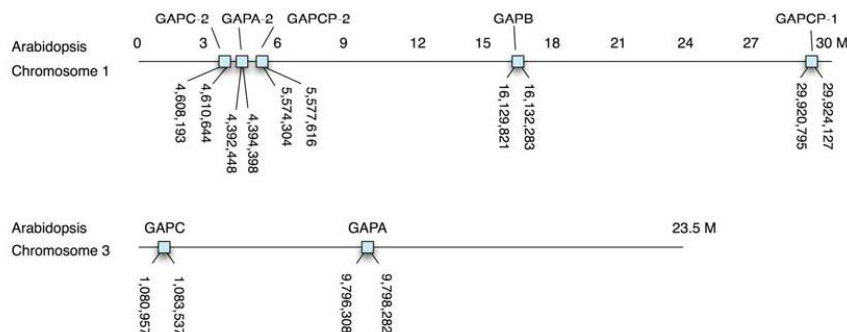
Note: If your goal was to clone an *Arabidopsis thaliana* gene, then this is not necessary. If no *Arabidopsis* genes were returned in the BLAST results this step is not necessary. You can proceed to step 3.6.

3.4.1 In the Hit Table, sort the results by clicking on the column headers for Bit-Score, E value, and Grade.

3.4.2 View your BLAST results in the sorted Hit Table. Using the Accession and Description columns, look to see whether any of your top hits come from *Arabidopsis thaliana*.

A number of scenarios can occur, depending on how your cloning reaction took place:

- If the top match to your sequence IS NOT an *Arabidopsis* sequence, it is unlikely that you have cloned an *Arabidopsis* gene
- If your top match IS an *Arabidopsis* sequence, then look at the sequence alignment of your novel sequence with the *Arabidopsis* sequence
- If the aligned sequence is broken up into two or more sections, this suggests there is a region that does not match the subject sequence, indicating your sequence IS NOT from an *Arabidopsis* gene.
- If the entire query sequence aligns in a single block with *Arabidopsis thaliana*, then look at the green Identity graph at the top of the sequence alignment, just beneath the blue depth of coverage chart. This graph indicates the homology of the query sequence with the subject sequence. Hovering over the chart will display the value as a percentage; clicking a base will display the value as a fraction in the bottom left corner of the window
- If the Identity value is below 90%, then it is unlikely to be an *Arabidopsis* GAPC gene
- If your sequence has low homology or has gaps that have no homology, it is unlikely to be an *Arabidopsis* gene
- If your sequence has high homology with the same *Arabidopsis* gene, it is likely that the gene is from *Arabidopsis* and may have been cloned accidentally. However some plant species closely related to *Arabidopsis* may have high homology. You will need to determine how closely your plant is related to *Arabidopsis* and whether to continue analysis with this contig



Chromosomal locations of *Arabidopsis* GAPDH genes. Note that GAPC is also known as GAPC-1.

3.5 Calculate a total homology score for the best matches with four *Arabidopsis* GAPC genes (GAPC, GAPC-2, GAPCP-1, and GAPCP-2).

In the Hit Table and Query Centric View, examine the sequence alignments between your query sequence and the subject sequences to see which of the four *Arabidopsis* GAPC genes have the greatest **homology**.

Note: The number of GAPC genes you identify will depend on how similar your plant sample is to the four *Arabidopsis* GAPC genes. For example, for the cbroccoli contig, only two of the four *Arabidopsis* GAPC genes (GAPC and GAPC-2) were identified as best matches when retrieving 50 results.

- **Homolog** — a gene related to a second gene by descent from a common ancestral DNA sequence. The term homology may apply to the relationship between genes separated by speciation (ortholog), OR to the relationship between genes originating via genetic duplication (see paralog)
- **Ortholog** — orthologs are genes in different species that have evolved from a common ancestral gene via speciation. Orthologs often (but certainly not always) retain the same function(s) in the course of evolution. Thus, functions may be retained, lost, or gained when comparing a pair of orthologs
- **Paralog** — paralogs are genes produced via gene duplication within a genome. Paralogs typically evolve new functions or else eventually become pseudogenes

To increase the certainty of your identification, you will need to determine the total score for the match between the highest overall scoring *Arabidopsis* GAPC subject sequence and your query sequence. To make this easier, you can use the prepared tables at the end of this section to analyze your results.

The first row in each of the prepared tables shows the four GAPC genes (GAPC, GAPC-2, GAPCP-1, and GAPCP-2), their chromosomal locations, and the coordinates in the *Arabidopsis* genome.

Use the prepared tables to record the beginning and ending chromosomal positions where the subject sequence (one of the four *Arabidopsis* GAPC genes) matches your query sequence for each fragment for which a match is found. Record the beginning and ending positions of each match of the query (contig) sequence for each alignment. And finally, record the score for that alignment. When you are through entering the information for each region that aligns within that gene, calculate a total score for that gene by adding the scores for each alignment.

3.5.1 Use the total score calculated from the table to identify the gene that your sequence matches best.


3.5.2 The result from the best match gives you the identity of your gene. To help you with this, below are examples using the cbroccoli contig.

- The BLAST results from the contig (cbroccoli, in this example) are viewed using the Hit Table, sorted by E value and Grade.

Hit Table	Query Centric View	Annotations	Distances	Fasta Alignment View	Info							
<input checked="" type="checkbox"/>	Name	Organism	Description	E Value	Bit-Score	Original Query Frame	Grade ▾	% Pairwise I...	Query cover...	% Identical ...	Sequence L...	Hit st...
<input checked="" type="checkbox"/>	HG994367	Brassica napus	Brassica napus genome ass...0	1,461.82	1		96.9%	94.2%	99.48%	94.2%	972	24,07
	LR031872	Brassica oleracea	Brassica oleracea HDEM ge...0	1,378.72	1		96.2%	93.0%	99.48%	93.0%	968	23,78
	LR031572	Brassica rapa	Brassica rapa genome, scaf...0	1,160.82	1		94.1%	88.8%	99.48%	88.8%	979	15,18
	MF158633	Brassica oleracea	Brassica oleracea glycerald...0	815.493	1		91.1%	82.8%	99.48%	82.8%	992	954
	KF030135	Brassica rapa	Brassica rapa subsp. nippos...0	804.413	1		91.0%	82.5%	99.48%	82.5%	991	954
	LR031568	Brassica rapa	Brassica rapa genome, scaf...0	798.873	1		91.0%	82.5%	99.48%	82.5%	992	54,58
	HG994363	Brassica napus	Brassica napus genome ass...0	798.873	1		91.0%	82.5%	99.48%	82.5%	992	50,40
	LR699747	Arabidopsis thaliana	Arabidopsis thaliana genom...0	719.467	1		90.0%	80.6%	99.48%	80.6%	1,003	1,092
	JQ248954	Halesia sp.	Halesia sp. LK-2012 isolate ... 0	708.387	1		89.9%	80.4%	99.48%	80.4%	1,004	1,009
	M64119	Arabidopsis thaliana	Arabidopsis thaliana cytosol...0	702.847	1		89.9%	80.3%	99.48%	80.3%	1,004	2,359
	JN083805	Eruca vesicaria	Eruca vesicaria subsp. sativ...0	684.381	1		89.8%	80.1%	99.48%	80.1%	992	968
	HG994357	Brassica napus	Brassica napus genome ass...0	1,099.88	1		89.6%	90.0%	89.27%	90.0%	872	15,41
	LT669788	Arabis alpina	Arabis alpina genome asse... 0	708.387	1		88.8%	80.9%	96.77%	80.9%	988	10,95
	LR031877	Brassica oleracea	Brassica oleracea HDEM ge...0	981.691	1		88.2%	87.5%	88.96%	87.5%	870	53,47
	HG994369	Brassica napus	Brassica napus genome ass...0	981.691	1		88.2%	87.5%	88.96%	87.5%	870	55,61
	LR031878	Brassica oleracea	Brassica oleracea HDEM ge...0	957.685	1		88.1%	86.9%	89.27%	86.9%	884	50,37
	HG994359	Brassica napus	Brassica napus genome ass...0	952.145	1		87.9%	87.0%	88.85%	87.0%	872	40,31
	LR031570	Brassica rapa	Brassica rapa genome, scaf...0	979.845	1		87.9%	87.8%	88.02%	87.8%	866	46,26
	LR778313	Raphanus sativus	Raphanus sativus genome a...0	891.205	1		87.7%	85.3%	90.00%	85.3%	901	42,85
	HG994355	Brassica napus	Brassica napus genome ass...0	878.279	1		87.3%	85.5%	89.17%	85.5%	890	30,78
	LR031571	Brassica rapa	Brassica rapa genome, scaf...0	902.285	1		87.3%	86.1%	88.44%	86.1%	876	36,73
	HG994355	Brassica napus	Brassica napus genome ass...0	830.266	1		86.5%	84.7%	88.33%	84.7%	871	32,19

BLAST results using cbroccoli contig as the query. The Hit Table is sorted by E value.

- Using the Organism column, the BLAST hits can be re-sorted to find the *Arabidopsis* chromosome hits from the contig

Tip: Alternatively, you can use Description to sort your Hit Table. The description will include the scientific name of the organism as well as the chromosome where the query result is found, if known. Use the small data table icon  at the upper right of the document table to enable the description column in the Hit Table.

- Click to select *Arabidopsis thaliana* chromosome 3. In Alignment View (document viewer), you can see that there are annotations for *GAPC-1*

The screenshot displays the NCBI BLAST results for a query sequence. The top section shows the 'Hit Table' with columns for Name, Organism, Description, E Value, Bit-Score, Original Query Frame, Grade, % Pairwise Identity, Query coverage, % Identical, Sequence Length, and Hit score. The bottom section shows the 'Alignment View' for the selected sequence, M64119, which is an Arabidopsis thaliana cytosol sequence. The alignment view includes a consensus sequence, identity, and coverage. The right panel shows 'Annotations and Tracks' for the alignment, including CDS (1), Exon (5), Gene (1), Intron (4), mRNA (1), and Source (1). The alignment view also shows the sequence alignment with gaps and the corresponding gene model for GAPC1.

Hit Table	Query Centric View	Annotations	Distances	Fasta Alignment View	Info						
Name	Organism	Description	E Value	Bit-Score	Original Query Frame	Grade	% Pairwise I...	Query cover...	% Identical ...	Sequence L...	Hit str...
HG994367	Brassica napus	Brassica napus genome ass...0	1,461.82	1	96.9%	94.2%	99.48%	94.2%	972	24,07	
LR031872	Brassica oleracea	Brassica oleracea HDEM ge...0	1,378.72	1	96.2%	93.0%	99.48%	93.0%	968	23,78	
LR031572	Brassica rapa	Brassica rapa genome, scaf...0	1,160.82	1	94.1%	88.8%	99.48%	88.8%	979	15,18	
MF158633	Brassica oleracea	Brassica oleracea glycerald...0	815.493	1	91.1%	82.8%	99.48%	82.8%	992	954	
KF030135	Brassica rapa	Brassica rapa subsp. nippos...0	804.413	1	91.0%	82.5%	99.48%	82.5%	991	954	
LR031568	Brassica rapa	Brassica rapa genome, scaf...0	798.873	1	91.0%	82.5%	99.48%	82.5%	992	54,58	
HG994363	Brassica napus	Brassica napus genome ass...0	798.873	1	91.0%	82.5%	99.48%	82.5%	992	50,40	
LR699747	Arabidopsis thaliana	Arabidopsis thaliana genom...0	719.467	1	90.0%	80.6%	99.48%	80.6%	1,003	1,092	
JQ248954	Halesia sp.	Halesia sp. LK-2012 isolate ... 0	708.387	1	89.9%	80.4%	99.48%	80.4%	1,004	1,009	
M64119	Arabidopsis thaliana	Arabidopsis thaliana cytosol...0	702.847	1	89.9%	80.3%	99.48%	80.3%	1,004	2,359	

Selected sequences are only summaries [Download Full Sequence\(s\)](#)

Alignment View Annotations Distances Dotplot Text View Download Fasta Alignment View Fasta Nucleotide View Lineage Info

← → ↺ Extract R.C. Translate Add/Edit Annotation Allow Editing Annotate & Predict Save

Consensus Identity Coverage

1. cbrooccl_D... 2. M64119

Annotations and Tracks (12 of 13)

Show Annotations

Types

- CDS (1)
- Exon (5)
- Gene (1)
- Intron (4)
- mRNA (1)
- Source (1)

FF Columns Track Pop out

Click in the document table to select *Arabidopsis thaliana* chromosome 3 and click Alignment View in the document viewer. Here, you can see that there are annotations for *GAPC1*.

- Switch to Text View in the document viewer. Use the information provided to fill out the gene tables. You should see one BLAST hit for each fragment

Information such as gene and query locations on chromosomes is displayed in **Text View**. Use this information to fill out the gene tables for your contig, such as chromosome location, bit-score, and the beginnings and ends of the sequence locations for both the query and the gene. Only the top part of the Text View page is shown above; scroll down for query end and end of gene location.

Alignment score of cbroccoli contig with *Arabidopsis* GAPC genes

Gene	<i>Arabidopsis</i> Chromosome	Beginning of <i>Arabidopsis</i> Gene Location	End of <i>Arabidopsis</i> Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPC-1</i>	3	1,083,023	1,082,032			
				1	954	822.721
Total score						822.721

Note: If your top match does not have an *Arabidopsis* sequence, or if your last results do not contain *Arabidopsis*, you may complete the tables using the gene with the best homology to your sequence.

In the example shown in the table above, one fragment of our cbroccoli query sequence has homology within the *GAPC-1* gene on chromosome 3. Our analysis shows that the identity of the cbroccoli contig query sequence is most likely to be *GAPC*. This identification is supported by the total blastn bit-score of 822.721.

Arabidopsis GAPC gene tables

Use these tables to calculate the homology scores of your contig with the *Arabidopsis GAPC* genes. Note that the number of *GAPC* genes that appear in your BLAST results will vary and will depend on the plant source you choose and how many BLAST results you choose to retrieve.

Gene	Arabidopsis Chromosome	Beginning of Arabidopsis Gene Location	End of Arabidopsis Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPC-2</i>	1	4,608,193	4,610,644			
Total score						

Gene	Arabidopsis Chromosome	Beginning of Arabidopsis Gene Location	End of Arabidopsis Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPCP-2</i>	1	5,574,304	5,577,616			
Total score						

Gene	Arabidopsis Chromosome	Beginning of Arabidopsis Gene Location	End of Arabidopsis Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPCP-1</i>	1	29,920,795	5,577,616			
Total score						

Gene	Arabidopsis Chromosome	Beginning of Arabidopsis Gene Location	End of Arabidopsis Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPC</i>	3	1,080,957	1,083,357			
Total score						

3.6 Results Analysis.

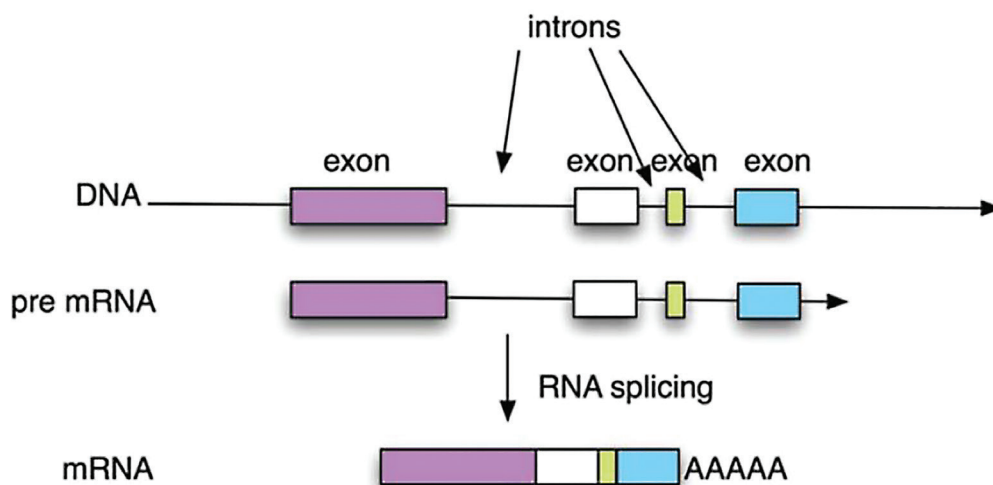
Based on your total scores, which *GAPC* gene does your contig show the highest homology to? Why?

3.7 Section 3 Focus Questions.

1. Does an E value of zero mean that your sequence matched the subject sequence well or poorly? Explain your answer.
2. What would it mean if you found a subject sequence with an E value of 3?
3. Describe how the % pairwise identity score could help determine how closely two species are related.

4. Determine gene structure (intron/exon boundaries) Using BLAST — build a gene model.

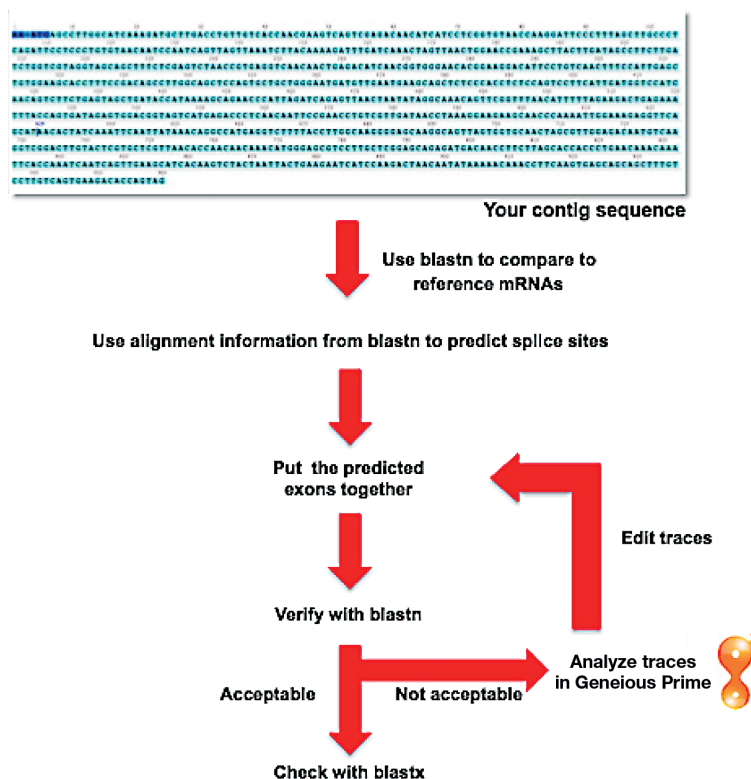
Researchers are often faced with having to build gene models to show where such features as exons and introns exist within the gene. Our goal at this stage of the project is to construct a gene model that shows where the exons are likely to be located within our contig. The process of identifying the protein coding sequences and adding that information to the sequence is called annotating the sequence. The extra bits of information are called annotations.



A gene is composed of introns and exons. Introns are spliced out of pre-mRNA to make mRNA.

There are algorithms that attempt to predict gene splice sites; they are not yet sophisticated enough to work every time, for every gene, in every organism. This is because we do not yet understand splicing signals well enough to accurately predict splice sites. If a gene has had its complete mRNA cloned and sequenced, then the splice sites can be predicted by aligning the mRNA sequence (which represents the coding region or exons of the gene) with the genomic contig sequence to help build a gene model. However, this requires a large level of experimental work. So to predict the mRNA of a well-conserved gene like *GAPDH*, we can compare the genomic contig sequence to reference mRNA sequences. These are mRNA sequences that have been reviewed and characterized by NCBI staff.

Using Geneious Prime, you will first compare your contig sequence to the nonredundant (nr/nt) database to get an approximation of where the intron/exon boundaries are located so you can generate an initial gene model. Then you will refine the gene model and make corrections as needed. These steps will be repeated multiple times until the model is correct. The workflow for this step of the bioinformatics lab is displayed below.

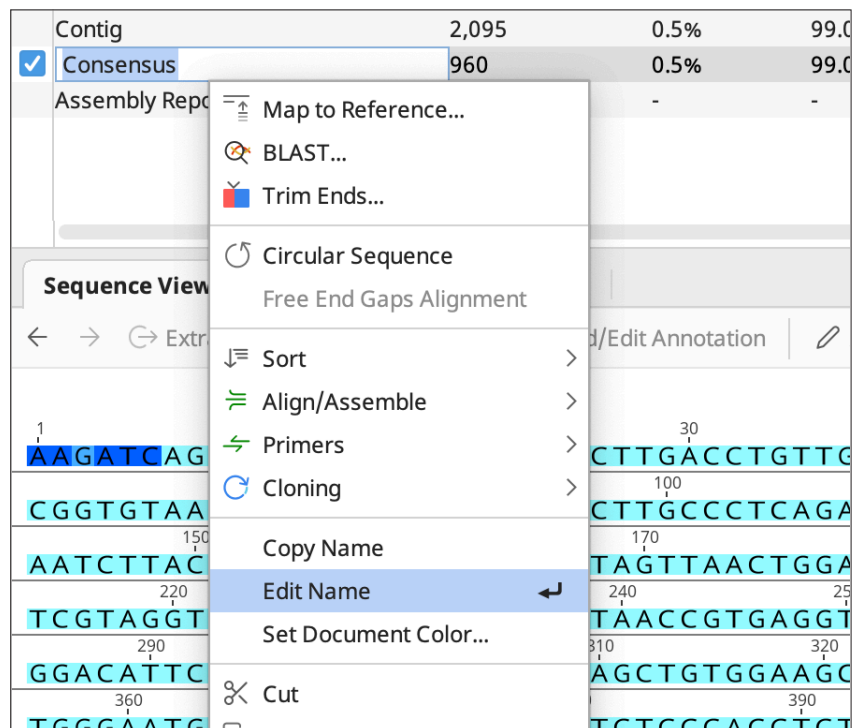


Workflow to determine intron/exon structure and the putative mRNA sequence.

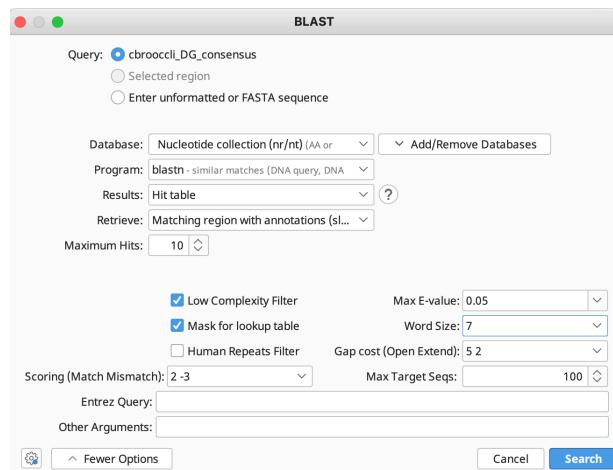
4.1 Using blastn to align the contig to sequences in the nr database.

For this section, retrieving results from a BLAST search using the Geneious Prime platform and from the NCBI BLAST website takes about the same amount of time. If you have been using the NCBI BLAST website for your BLAST searches in previous sections, try doing this BLAST within the Geneious Prime program.

- 4.1.1 To avoid confusion with file names in the subsequent workflow steps, rename your consensus sequence. In your assembly folder, select your Consensus file and right click to open a menu. Select **Edit Name**. Rename your file to something easy to remember, such as the name of your sample, your initials, and then consensus. In the example below, the file will be renamed to **cbroccoli_DG_consensus**.



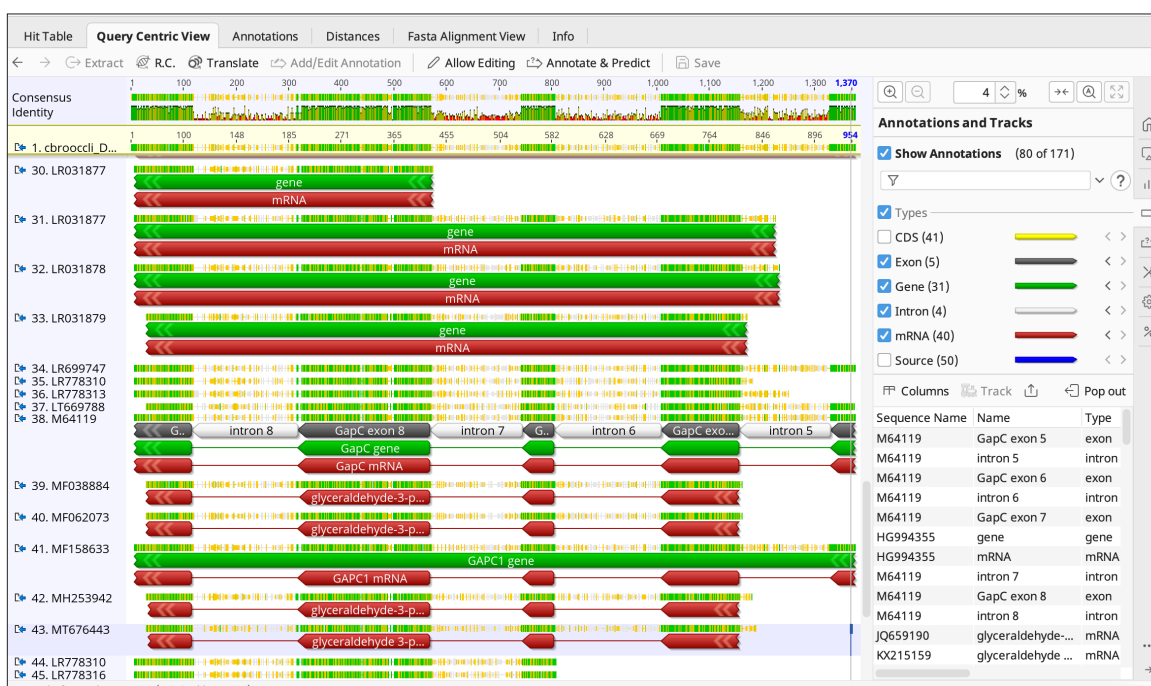
- 4.1.2 To begin your BLAST search, select your newly renamed consensus file. Select the BLAST icon from the menu bar. A new dialog box will appear:



- Select **Nucleotide collection (nr/nt)** as your database
 - The default selection for Query should be your renamed consensus file
 - Select **blastn** for Program
 - Select **Hit table** for Results
 - Select **Matching region with annotations (slow)** for Retrieve
 - Set Maximum hits to **10**. This change will make the results easier to interpret because fewer sequences will be shown
- Click **More Options**
 - Change word size to **7**. This will increase the sensitivity of the blastn search and allow you to detect more distantly related sequences and short exons
 - Click **Search**. A new folder will be created within your folder with the name of your consensus file — Nucleotide collection (nr_nt) blastn (10)

4.2 Interpreting the results and predicting the exon positions.

The Query Centric View will show the consensus sequence and nucleotide coordinates near the top of the page. Below, you will see the BLAST results alignment showing where portions of the sequences from the nr database align to your contig. If you do not see the annotations colored in red and green like the example below, navigate to the Annotations and Tracks tab ⇨ and check the boxes for Show Annotations, then Gene and mRNA. The mRNA annotations (in red) correspond to known exons.



- 4.2.1 Identify the BLAST result with the longest and most extensive match to your contig. You can do this by looking at the statistics in the Hit Table and the corresponding alignments in Query Centric View. In the example below, the best match is the sequence named HG994367 (*Brassica napus*) because it has the lowest E value and highest Bit-score, Grade and % Pairwise values

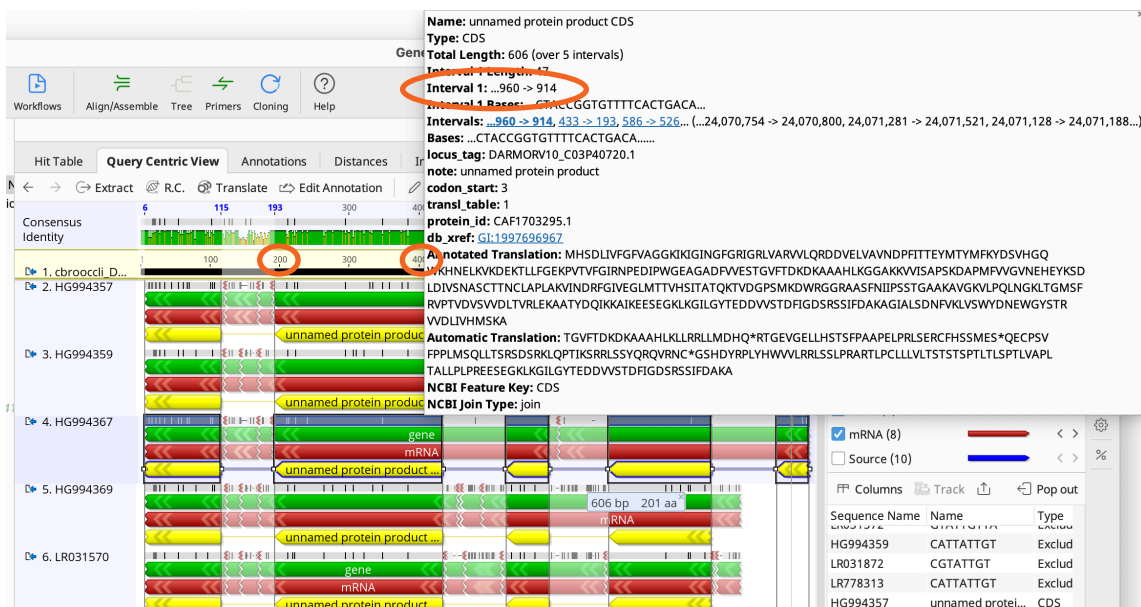
Tip: How should the statistics be prioritized to identify the best mRNA match? Recall from Section 3 that the smaller the E value, the lower the chances of this hit being identified by chance, a higher bit-score represents higher similarity, and that the grade represents a calculation of E value, query coverage, and % identity to help sort for the longest, strongest identity hits from the list.



The best mRNA match for cbroccoli_1 was selected based on the ranking of bit-score, E value and grade. In this case, the best match comes from the organism, *Brassica napus*.

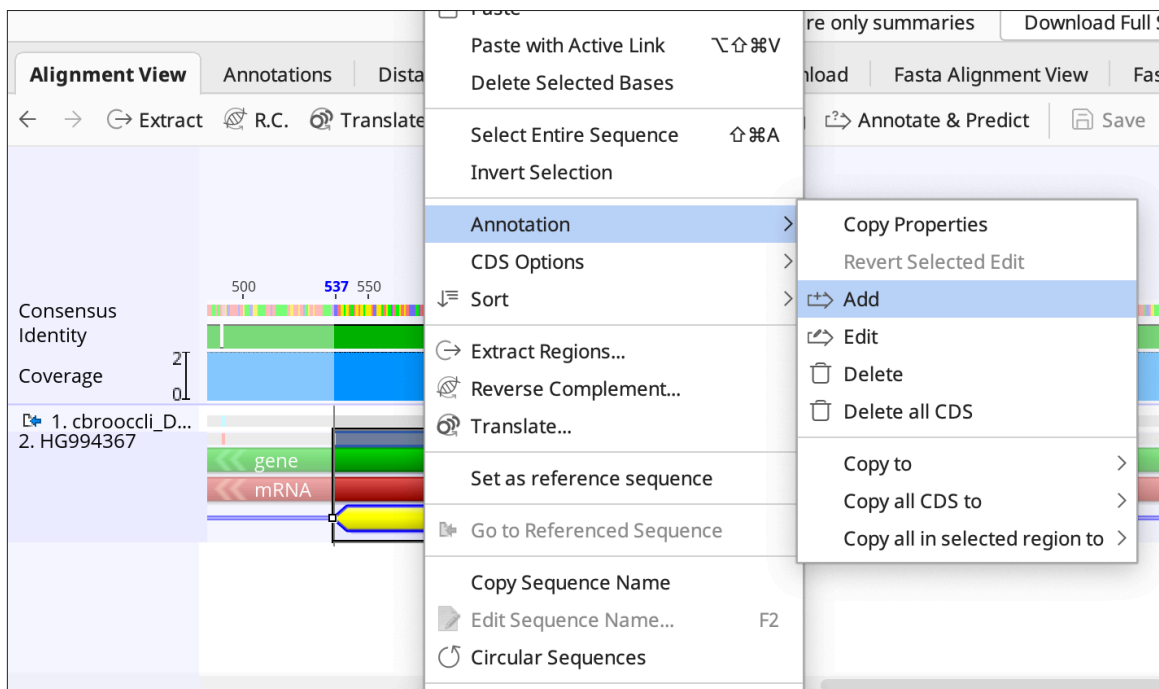
4.2.2 From the BLAST result with the best match to your contig, record the nucleotide coordinates of the mRNA intervals. These nucleotide coordinates will help you predict intron/exon boundaries in your consensus sequence.

- Go to Query Centric View
- For your best match from the BLAST results, click on one of the mRNA annotations (red). All of the mRNA annotations will be selected
- There are two places where you can look for the nucleotide coordinates:
 - o The coordinates for each interval appear in blue above the consensus sequence. See examples encircled in orange in the figure below. You now have all the nucleotide coordinates for all of your intervals
 - o The coordinates for the interval you are hovering over can be found in the pop up window. In the example below, interval 1 is highlighted, which spans bases 1,174–1,128 . You will note that this range is decreasing; this is because this interval is in the reverse direction (see arrowhead at the end of the annotation). Repeat this for all the intervals
- Record these coordinates and note whether these intervals are in the forward or reverse direction, based on which direction the interval arrowhead is pointing. In the example below, the orange arrow points to the arrowhead for interval 1, which is pointing in the reverse direction
- Also note whether any of the intervals are truncated, or cut off at the end. In the example below, interval 1 is truncated at the right side because it appears to continue beyond our sequence. Interval 5 is also truncated but on the left side, because it appears to continue to the left beyond our sequence



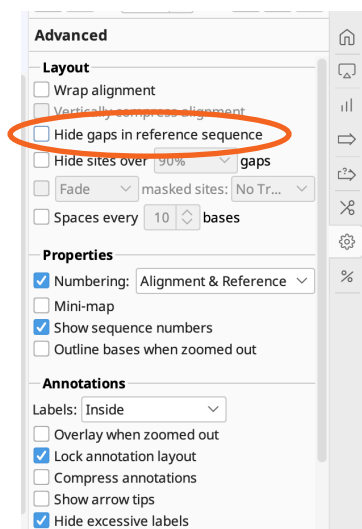
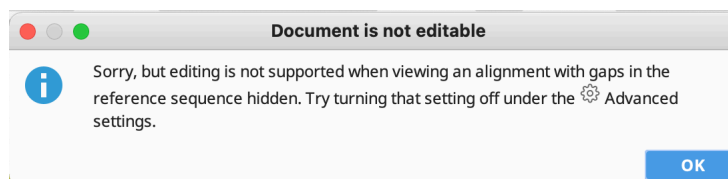
There are two places to find nucleotide coordinates to map intron/exon boundaries: 1, click the mRNA annotation (red) for your best match and look for the blue numbers (see orange circles). 2, hover over an interval and the pop up window will list the nucleotide range (see orange oval). Also note the directionality of the interval by looking for the arrowhead (see orange arrow). An arrowhead on the left side indicates that the interval is in the reverse direction.

- 4.2.3 Find the intron/exon boundaries in your renamed consensus sequence and mark them. In Query Centric View, right click on your consensus sequence. In the new menu, navigate to **Annotation**, then **Add**.



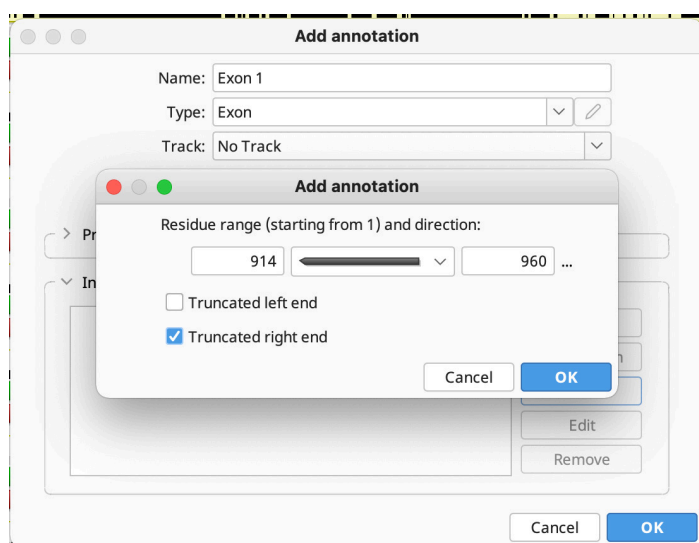
Right click on your renamed consensus file to open the menu to add an annotation.

If you see a warning stating that your Document is not editable, follow the instructions to uncheck the box for **Hide gaps in reference sequence**.



In the new window that appears, you will add all the nucleotide intervals that you noted down from section 4.2.2. Begin with the interval nearest the 5' end.

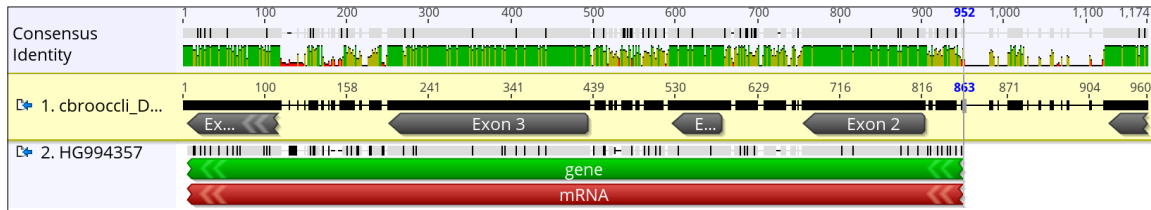
- For Name, the first interval will be named Exon 1
- For Type, select **Exon** from the dropdown menu
- For Track, keep the default selection
- Select the direction for that interval (that is, which way is the arrowhead pointing?)
- In the **Intervals** section, click to highlight the existing interval, and select **Remove** (this represents your entire consensus; you do not need to annotate this). Now click **Add**, and enter the nucleotide coordinates that you previously noted down. Make sure the arrow is pointing in the correct direction. If your interval was cut off or truncated at either end, indicate that by checking the correct box. Otherwise, leave them unchecked



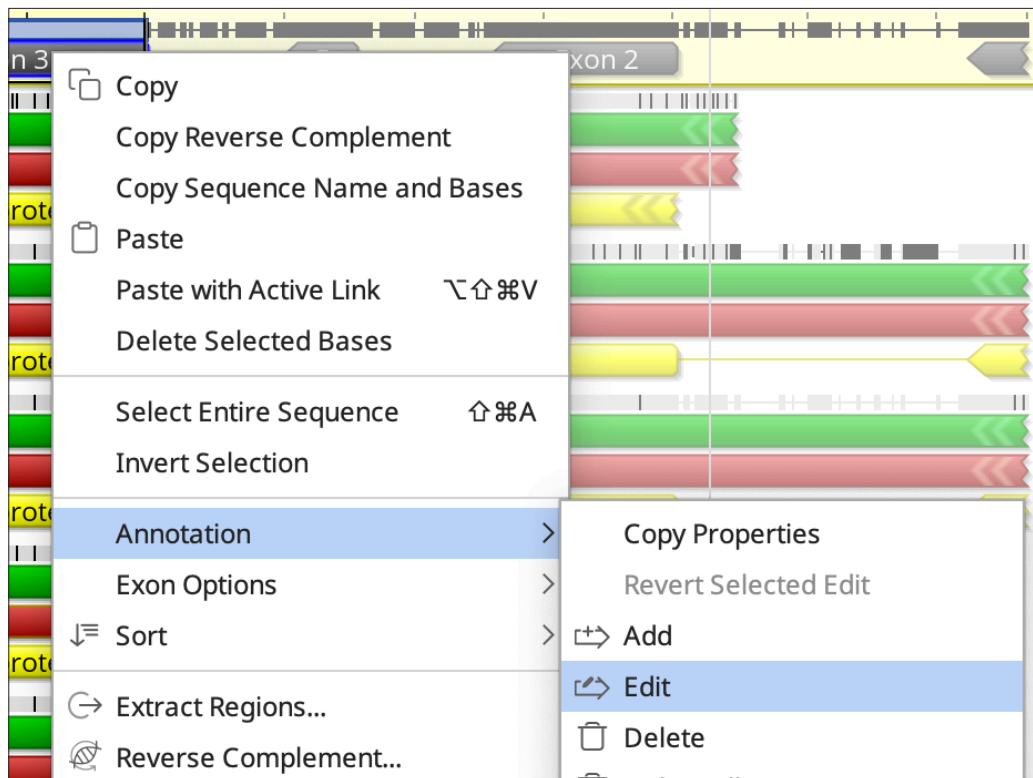
- Click **OK** to exit both windows. Your new interval called Exon 1 has been added as an annotation
- For your next interval, repeat from the beginning of step **4.2.3** and label it Exon 2
- Go through all of your intervals and add them all as annotations on your consensus sequence

Note: If you want to submit your sequence to GenBank, the exons must be numbered sequentially in this fashion. Otherwise, you can use the Add button to add all your intervals at the same time without having to open new windows each time, but all the exons will have the same name.

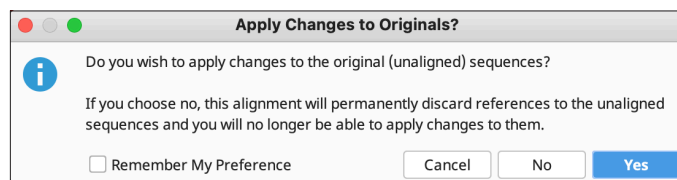
- When you are finished, your mapped exons will appear on your consensus document as gray annotations. (You may also see other gray annotations in the BLAST results, because adding your exons automatically checked the box to show all exon annotations)



- If you made a mistake while annotating an exon(s) and want to delete it:
 - Click to highlight the exon. Right click and navigate to **Annotation**, then **Delete** (or **Edit**)



- In the Query Centric View toolbar, be sure to click **Save** to save your work. If you navigate away from the view, you will be prompted to save your changes
- You will also be asked to apply changes to the originals. Click **Yes**



4.3 Check initial gene model (putative mRNA) with blastn and further refine model.

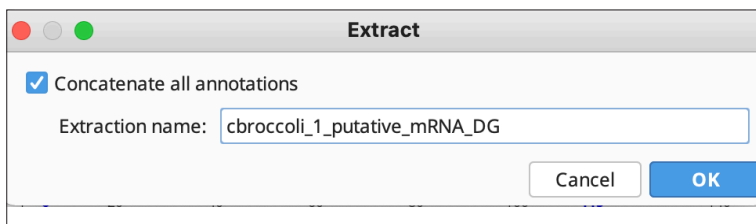
Now it is time to check your work by doing a blastn search with your proposed sequence for the mRNA. You will create a new document with only the exon sequences as your putative mRNA, run a new blastn search, and see how well the top results match your putative mRNA.

4.3.1 To create a document with only your exon sequences as your putative mRNA, go to your Assembly folder and select your annotated consensus sequence. Click on the Annotations and Tracks tab ⇨. Select all the exons by holding the Shift button and highlighting the first and last exon on the list. In Sequence View, you will see that the sequences of these exons are highlighted.

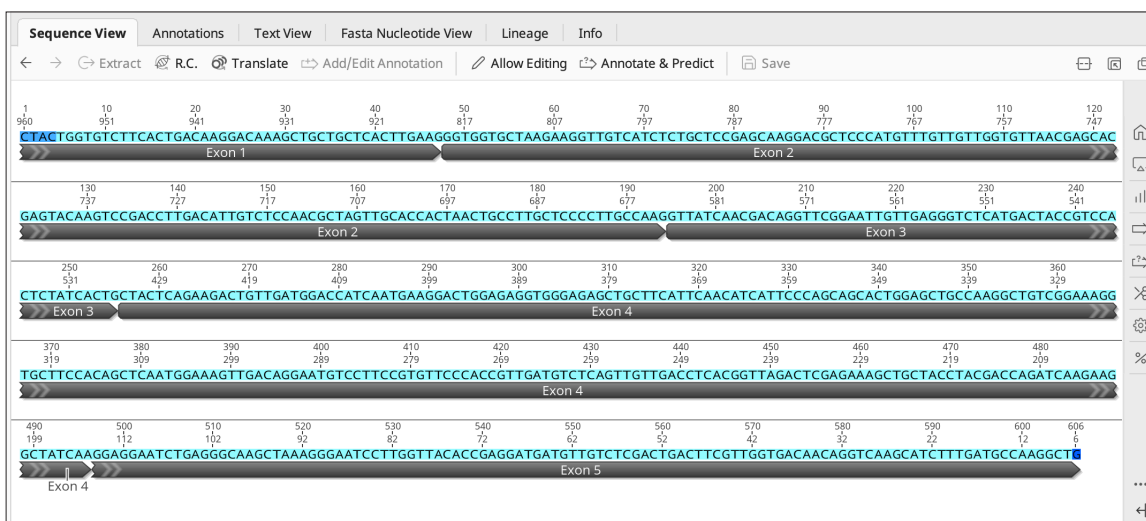
- Hover your mouse over the selected exons in the Annotations and Tracks tab. Right click and navigate to **Extract Regions**

Name	Minimum	Maximum
Exon 1	914	>960
Exon 2	673	819
Exon 3	526	586
Exon 3	193	433
Exon 5	6	>115

- In the new window that appears, check the box for **Concatenate all annotations**. For the extraction name, be sure to add “putative mRNA” and select a naming convention similar to your previous consensus document. In the example below, the new file will be named **cbroccoli_1_putative_mRNA_DG**



- Click **OK**. You will see your new putative mRNA document in the Assembly folder. In Sequence View, you'll see that all the intronic sequences are removed and only the exonic regions are kept



Intronic sequences are removed, with only exon sequences remaining in the putative mRNA sequence.

4.3.2 Run a blastn search on your putative mRNA sequence. Select your putative mRNA sequence file. Click the BLAST icon in the menu bar. Keep the defaults from the previous search, namely:

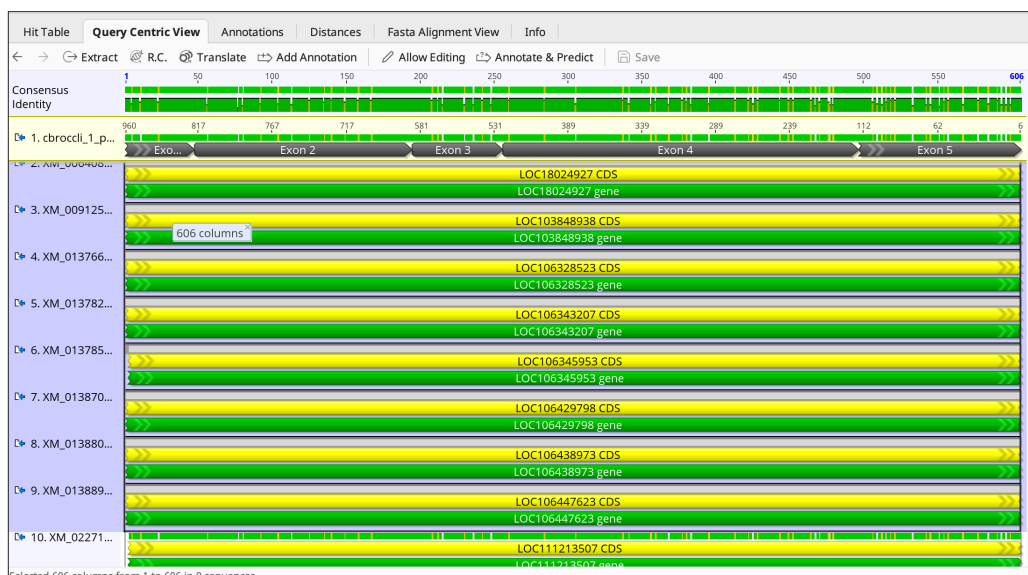
- Database should be **Nucleotide collection (nr/nt)**
- Program should be **blastn**
- Results should be **Hit table**
- Retrieve should be set to **Matching region with annotations (slow)**
- Maximum Hits should be set to **10**
- Click **More Options** and make sure Word Size is set to **7**
- Click **Search**. A new folder containing your results will appear. It will use your document name with – Nucleotide collection (nr_nt) blastn (10) appended to the end as the name of the folder. For example, cbroccoli_1_putative_mRNA_DG – Nucleotide collection (nr_nt) blastn (10)

The screenshot shows the BLAST search interface with the following settings:

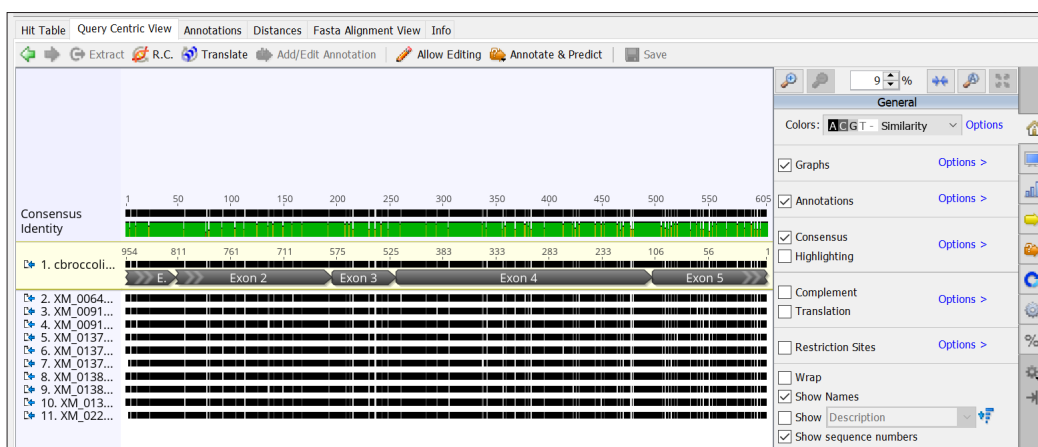
- Query:** cbroccoli_1_putative_mRNA_DG (Selected region)
- Database:** Nucleotide collection (nr/nt) (AA or)
- Program:** blastn - similar matches (DNA query, DNA)
- Results:** Hit table
- Retrieve:** Matching region with annotations (sl...)
- Maximum Hits:** 10
- Filters:**
 - ☒ Low Complexity Filter
 - ☒ Mask for lookup table
 - ☐ Human Repeats Filter
- Max E-value:** 0.05
- Word Size:** 7
- Gap cost (Open Extend):** 5 2
- Scoring (Match Mismatch):** 2 -3
- Max Target Seqs:** 100
- Entrez Query:** (empty)
- Other Arguments:** (empty)
- Buttons:** Fewer Options, Cancel, Search

4.4 Interpreting the results and finding and resolving errors.

In your putative mRNA blastn folder, you should now see your sequence and the ten best fits from the blastn search aligned in Query Centric View. These matches most likely will be different than the ones returned from the general nucleotide database when you searched for alignments for your single sequences and contig. This is because sequences in the reference database are scrutinized at a higher level than ones in the general database by NCBI scientists. In the Hit Table, there should be a high level of query coverage and a low E value; the % pairwise identity will vary. There is much more variation in the intron regions of genes than in the exons (coding regions), so the level of homology should be high. However, there could still be a reasonable amount of variation between plants in the coding regions.

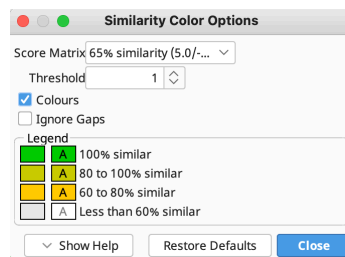


To facilitate seeing the % similarity between your putative mRNA and the blast results, uncheck the boxes for CDS (coding sequence) and Gene in the Annotations and Tracks tab. Navigate to the General tab and for Colors, select Similarity from the dropdown menu (see encircled in orange in the screenshot below). Your bases are now color coded in shades of gray according to % similarity. The closer the similarity the darker the color.

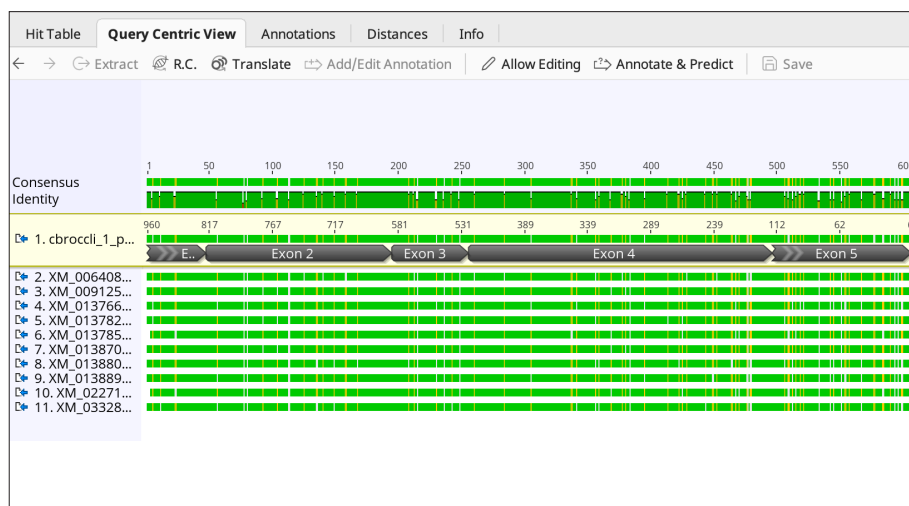


Annotations are not shown, and the color of the bases are set for similarity according to gray scale.

To make the similarity even easier to see, click on the Options button next to the Colors dropdown menu. In the new dialog box that appears, select Colours, then Close. The % similarity is now shaded in from highest (green) to lowest (gray) similarity.




When you did a blastn search in Section 2 using your contig sequence, your query sequence contained segments that would not be found in mRNA (introns). Since you used a putative mRNA segment as a query this time, your BLAST results should not show gaps. In the example below, the putative mRNA matches several subject sequences along the entire length with almost no gaps.



% similarity between the bases of the blast hits are now in a gradient of green, yellow, and gray.

Your results may be similar, or you may find breaks where a portion of sequence is missing or differs between plant species. You will now need to examine your results in further detail, since there may be regions where two joined exons have errors. There may also be errors that were missed on the first run-through.

- 4.4.1 Look through your sequence. Are there any indels relative to all the matched sequences? If so, there may be an issue with your assigned splice locations. View your putative mRNA and blastn results in the Hit Table and Query Centric View to see if there are any gaps.

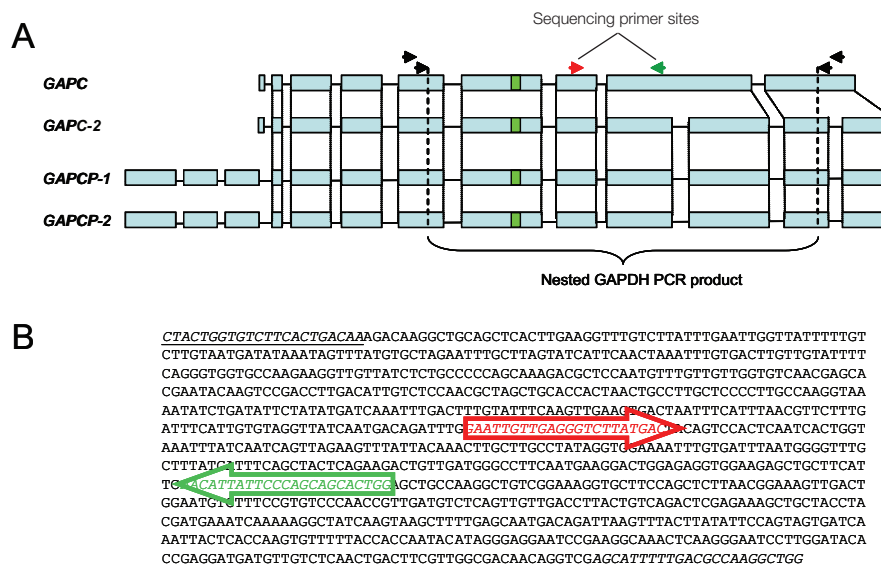
Tip: You can also navigate to the Advanced Settings tab  and toggle the box for **Hide gaps in reference sequence** (as in step 4.2.3) to see if the sequences look any different. If the sequences look the same, there are no gaps.

- 4.4.2 If gaps are observed, you will need to examine your results in further detail. Return to your consensus document and double check your intron/exon mapping. Errors in coordinates could contribute to affect your BLAST results.
- 4.4.3 If there are no gaps, you are ready to proceed to the next section, Predict an amino acid sequence from the cloned gene.

4.5 Comparing your plant's intron/exon structure to *Arabidopsis GAPC* at the DNA level.

Your corrected contig now contains the annotations that denote where intron and exon boundaries are located. At the DNA level, you have probably noticed that there are many differences between your sequence and that of *Arabidopsis*, especially in the intronic regions. Intronic DNA has little selective pressure, and therefore introns can vary in many ways, such as in sequence, length, or even whether or not they are present. Exons are parts of DNA that are converted into mature messenger RNA (mRNA), therefore they are more conserved than introns, which do not get incorporated into mRNA.

The following diagram will allow you to compare the number of introns and exons in your plant to that of *Arabidopsis GAPC* genes. The arrows denote the location where the initial PCR reaction and the subsequent nested PCR reactions occurred and the locations of the internal *GAPC* sequencing primers GAP SEQ F and GAP SEQ R.



Intron/exon structure of *Arabidopsis GAPC* gene family. Top panel **A**, the black arrows show positions of initial and nested PCR primers. The red arrow shows the position where the GAP SEQ F sequencing primer would anneal and the green arrow shows where the GAP SEQ R sequencing primer would anneal. The green bars show the sequence coding for the active site of the GAPDH enzyme. Bottom panel **B**, the base pair sequence shown is that of the *GAPC* gene cloned in the control pGAP plasmid. The location of the nested PCR primers are underlined, and the GAP SEQ F and GAP SEQ R sequencing primers are depicted by the red and green arrows, respectively.

4.6 Results analysis.

1. How many exons and how many introns did your gene fragment have?
2. Does your plant's *GAPC* gene have more, fewer, or the same number of exons as its most homologous *Arabidopsis* *GAPC* gene?
3. From your BLAST analysis, what reference mRNA was your mRNA most similar to? What was its E value?

4.7 Section 5 Focus Questions.

1. What kinds of sequences will be found in a genomic sequence — exons, introns, or both?
2. Which kinds of sequences will we find in mRNA — exons, introns, or both? Explain your answer.
3. Why might the number of introns in a gene be different between plant species?

5. Predict an amino acid sequence from the cloned gene (blastx).

mRNA is translated into proteins by ribosomes and tRNA. Each amino acid is encoded by a group of three RNA bases called codons. A DNA sequence encodes six potential protein sequences, three for each strand, which are referred to as reading frames. The first codon in a sequence can start at base position 1, 2, or 3, and a gene can be transcribed in two directions, forward (+) and reverse (–), for a total of six ways to read the sequence.

From the data generated thus far, we have not determined which reading frame is valid for this GAPDH protein. To determine the protein sequence encoded by the putative *GAPDH* mRNA, a different BLAST program, blastx, will be used. The blastx program translates a nucleotide sequence in all six reading frames and compares the resulting six amino acid sequences to a database of protein sequences. Usually only one frame has any significant matches.

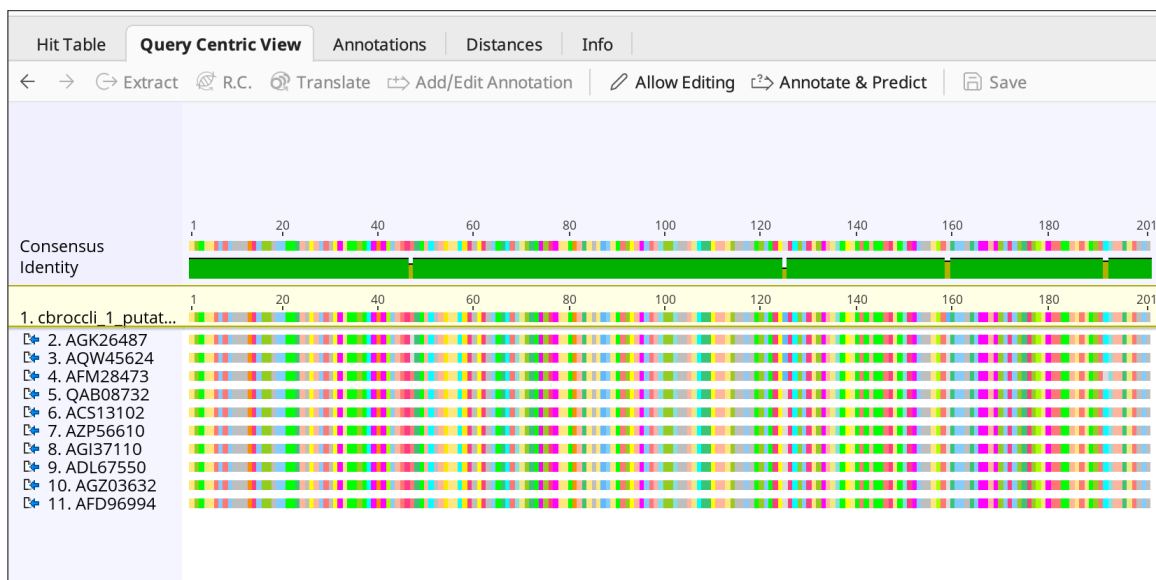
5.1 Checking the mRNA prediction with a blastx search.

- 5.1.1 In your Assembly folder, select your putative mRNA sequence from step 4.4.3.
- 5.1.2 Click the BLAST icon in the menu bar. A new dialog box will appear.


- Select **Nucleotide collection (nr/nt)** as the Database
- Select **blastx** as the Program
- Select **Hit table** for Results
- Select **Matching region with annotations (slow)** for Retrieve
- Set Maximum Hits to **10**
- Click **Search**. A new folder containing your results will appear. The name of the results folder will be the name of your document with “– Nucleotide collection (nr_nt) blastx (10)” appended to the end. For example, the cbroccoli contig would be: cbroccoli_1_putative mRNA_DG – Nucleotide collection (nr_nt) blastx (10)

5.1.3 If your mRNA model is correct, you should see the following results:

- The alignment should span the entire length of your query sequence



Blastx search results using a putative mRNA sequence as the query. The putative mRNA sequence from the cbroccoli contig was used as the query for a blastx search. The results are viewed in Query Centric View.

- The reading frame defines which base is assumed to be the first base that codes for the first amino acid when the mRNA sequence is translated. In the case of the cbroccoli putative mRNA, the reading frame is 3. This information can be found in the Hit Table in a column called Original Query Frame; you may have to scroll to the right in the Hit Table to find it. If you don't find it at all, use the small data table icon  on the upper right to select this column from the list and make it visible

Hit Table	Query Centric View	Annotations	Distances	Info					
Name	Organism	Description	E Value	Bit-Score	Original Query Frame	Grade	% Pairwise I...	Query	
AGK26487	Lobularia maritima	glyceraldehyde-3-phosphate dehydrogenase, partial [Lobularia ...	1.17e-128	370.548	3	99.3%	99.0%	99.50%	
AQW45624	Brassica oleracea	glyceraldehyde-3-phosphate dehydrogenase, partial [Brassica ...	1.69e-128	370.163	3	99.3%	99.0%	99.50%	
AFM28473	Braya pilosa	glyceraldehyde-3-phosphate dehydrogenase, partial [Braya pil...	1.85e-128	370.163	3	99.3%	99.0%	99.50%	
QAB08732	Schefflera arboricola	glyceraldehyde-3-phosphate dehydrogenase, partial [Scheffler...	2.99e-128	369.777	3	99.3%	99.0%	99.50%	
AZP56610	Tarenaya hassleriana	cytosolic glyceraldehyde-3-phosphate dehydrogenase, partial [...	3.50e-128	369.392	3	99.3%	99.0%	99.50%	
AGI37110	Schefflera actinophylla	glyceraldehyde-3-phosphate dehydrogenase, partial [Scheffler...	3.72e-128	369.392	3	99.3%	99.0%	99.50%	
ACS13102	Pilea cadierei	GapC, partial [Pilea cadierei]	3.13e-128	369.392	3	99.3%	99.0%	99.50%	
AGZ03632	
AFD96994	

Identifying the reading frame of the putative mRNA sequence using the blastx search results. The Original Query Frame column in the Hit Table identifies the reading frame of the putative mRNA to facilitate matching the results from the blastx search.

The 3 stands for “frame 3.” This means that amino acid coding begins on the third base of the forward strand. In our cbroccoli putative mRNA example, the first codon would be ACT (or ACU in RNA), which codes for threonine. This is the first amino acid in the translated query sequence as well as in the blastx hits.

5.1.4 To check the amino acid sequence of your putative mRNA, go to your Assembly folder and select your putative mRNA


- Select the **Display** tab in the Options panel, and click **Translation**
- In Translation Options, select the Frame from the dropdown menu. For the cbroccoli example, this would be Frame 3
- Keep the Genetic Code as **Standard**
- If you wish, you can change the Colors or check the box for **Three letter amino acids**

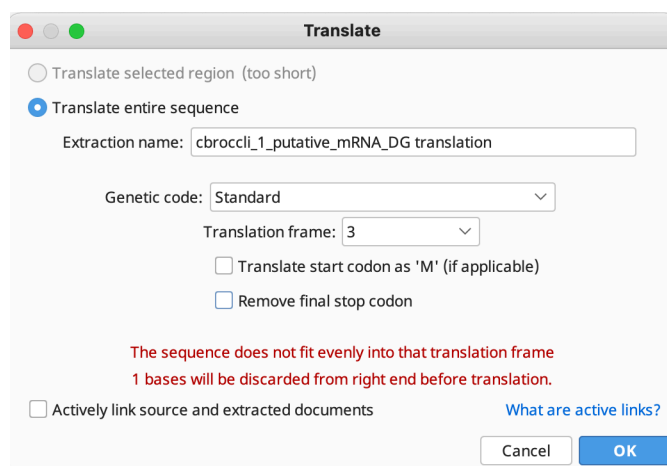
Identifying the first codon of the reading frame of the putative mRNA sequence. Since the reading frame for the cbroccoli putative mRNA is 3, also known as frame 3, the first codon in the sequence is predicted to be ACU, which codes for threonine (see orange circle).

Query Centric View of the cbroccoli putative mRNA sequence with blastx query results. The first codon for the cbroccoli putative mRNA sequence is threonine, which matches the rest of the blastx query results.

- 5.1.5 Record the reading frame of the best GAPDH protein sequence match for your clone: _____
- 5.1.6 If your translated sequence matches with sequences in the database but requires different reading frames for the entire sequence to match, it is possible that there is still an error in the mRNA sequence. You can go back to section 4.3 to check your mRNA sequence. Any insertions or deletions can affect the reading frame as well as which amino acid a codon codes for (because intron/exon boundaries can occur in the middle of a codon).

5.2 Translate your putative mRNA sequence to create a document of the predicted sequence of the protein.

- 5.2.1 In your Assembly folder, select your putative mRNA document file.
- 5.2.2 In the Sequence View toolbar, click the Translate button . A new dialog box will appear:
 - Select **Translate entire sequence**
 - Keep the default Extraction name (or rename it to something you prefer)



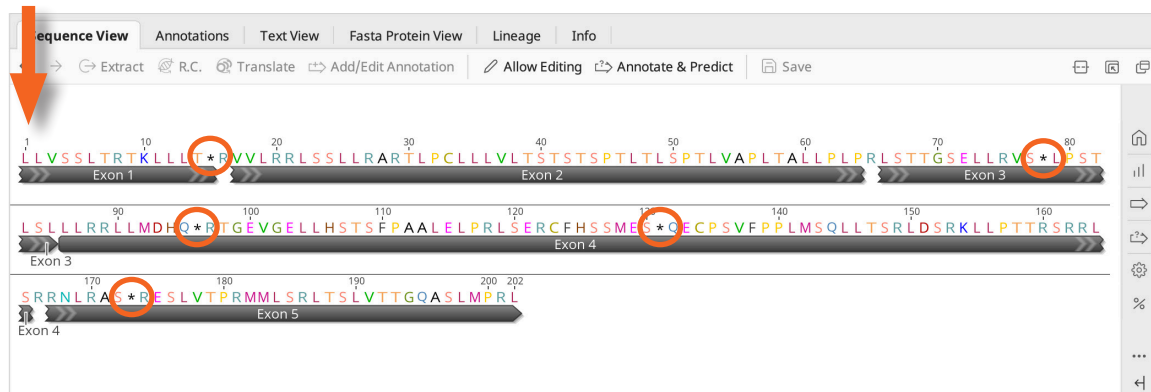
- Select **Standard** as the Genetic code
- Select the predicted frame from your sequence. If your predicted frame is not 1, a notice will alert you that the sequence does not evenly fit in the translation frame and that bases will be trimmed.
- Uncheck the box for **Treat first codon as start of coding region**. This is because you don't know for sure where the actual start of the coding region is relative to your contig
- Click **OK**. A document will appear in your Assembly folder, which represents a translation of your putative mRNA sequence

- 5.2.3 Look at your translated document in Sequence View. Make sure that the first amino acid is what you expected it to be, and that you see no stop codons within your sequence.



Translation of the cbroccoli putative mRNA sequence. Using the Translate button in Sequence View allows you to create a separate document of the predicted protein sequence.

Tip: A good prediction of protein sequence should result in a sequence of amino acids that read all the way through with no stops (indicated by an asterisk).



An example of a poorly predicted sequence. The first codon is not threonine, and multiple stop codons appear within the sequence (see orange circles for example).

- 5.2.4 Record the name of your putative mRNA document containing the correct reading frame here:
-

5.3 Run a blastp search on your putative protein sequence.

You now have a putative protein sequence with a reading frame that matches proteins found in a protein database. As a final verification step, you can use your putative protein sequence as a query sequence for blastp, another type of BLAST program that uses a protein sequence to search a protein database. If you have the correct sequence, it will match the correct protein.

5.3.1 Select the document in your Assembly folder for your translated putative protein.

5.3.2 Click the **BLAST** icon in the menu bar. A new dialog box will appear.

The screenshot shows a BLAST dialog box with the following settings:

- Query:** ☒ cbroccli_1_putative_mRNA_DG translation
 - ☐ Selected region
 - ☐ Enter unformatted or FASTA sequence
- Database:** Nucleotide collection (nr/nt) (AA or ...) [Add/Remove Databases]
- Program:** blastp - (AA query, AA database)
- Results:** Hit table [?]
- Retrieve:** Matching region with annotations (sl...) [?]
- Maximum Hits:** 10
- Buttons:** More Options, Cancel, Search

- Select **Nucleotide collection (nr/nt)** as the Database
- Select **blastp** as the Program
- Select **Hit table** for Results
- Select **Matching region with annotations (slow)** for Retrieve
- Set Maximum Hits to **10**
- Click **Search**. A new folder that contains your results will appear, named "putative mRNA from consensus – Nucleotide collection (nr_nt) blastp (10)"
- Your results can be examined under both the Hit Table and Query Centric View tabs.

5.4 If you wish to submit your sequence to GenBank, be sure you have determined the positions of the intron/exon boundaries of your putative mRNA sequence. See Appendix C for instructions on how to use the GenBank Submission plugin in Geneious Prime, or submit your sequence directly into GenBank using BankIt.

5.5 Section 5 Focus Questions.

1. blastx translates a nucleotide sequence in six reading frames and then uses each one to query a protein database. Why are there six possible reading frames?
2. In the blastx results the letters are no longer limited to A, G, C, or T. What do the letters in the blastx results represent?
3. The sequence SNASCTTNCLAP in exon 6 of *Arabidopsis GAPC* and *GAPC-2* is the active site of the GAPDH enzyme. Do you have an exon similar to exon 6 of *Arabidopsis GAPC* and *GAPC-2*? Is the active site sequence mentioned above exactly the same for your gene?
4. Do you have more, fewer, or the same number of introns and exons as the *Arabidopsis GAPC* genes? Does a difference in number of introns affect the final protein sequence?

Assemble Contig Sequences from the Entire Class

Once all students or groups of students have identified and corrected errors in their contig sequences, the class can assemble contigs from clones that represent the same gene from the same plant species to obtain greater depth of coverage for the gene. For submission to GenBank, it is good practice to have sequence depth of coverage of at least 6 to 8.

1. To assemble contigs, make a new folder in Geneious Prime.
2. Export the contig sequences from each student group as Geneious Prime files. Be sure to rename each contig to keep track of which contig came from which student group.
3. Import the contig files into the new folder in Geneious Prime.
4. Highlight all the contig files and run a sequence assembly as in Section 2.1.
5. Examine the assembly, check for any errors, and make corrections to the original reads and contigs using the techniques described in Section 2.3.
6. Review the locations of introns and exons and make corrections to the original reads using the techniques described in Section 4.

Congratulations!

You have completed the Cloning and Sequencing Explorer series. You have successfully cloned and sequenced a portion of a *GAPDH* gene from a plant of your choice and determined its potential intron/exon structure and protein sequence. If the class is satisfied with the depth of coverage and accuracy of the data, see Appendix C for instructions on how to prepare a gene sequence for GenBank submission. The data will then be available for other researchers to access for their own experiments.

Remember

If you wish to keep your sequence files and data long term, back up your files to your hard drive or a storage account. When your Geneious Prime license expires, you will still have access to the software, but some functions will be restricted.

APPENDIX A1: AGAROSE GEL ELECTROPHORESIS METHODS

Preparation of Electrophoresis Running Buffer and Agarose Gels

The recommended agarose concentration for gels in this classroom application is 1% (1 g agarose per 100 ml of electrophoresis buffer), which provides good resolution and minimizes the run time required for electrophoretic separation of DNA fragments. The recommended thickness of the gels is 0.75–1.0 cm for easy sample loading and gel handling.

Visualization of DNA

UView 6x Loading Dye and Stain is included with this series. The loading dye and stain is added directly to samples, the gel is run, and gels can be directly imaged on a UV imaging system. An alternative means of visualizing DNA is Bio-Rad's Fast Blast DNA staining solution (catalog #1660420), which is a biologically safe stain that does not require a documentation system. After electrophoresis, gels are stained with Fast Blast stain either using a quick 15-minute or an overnight protocol. Fast Blast stain is approximately five times less sensitive than ethidium bromide, which means that some faint DNA bands may be visible with ethidium bromide but not with Fast Blast stain. Another alternative method of DNA visualization is staining with SYBR® Green I, which also requires a visualization and documentation system. The protocol described below uses ethidium bromide for visualizing DNA. Finally, ethidium bromide could be used to visualize DNA using a UV transilluminator and documentation system. However, treat ethidium bromide as toxic. If you choose to use ethidium bromide, handle with care and follow standard laboratory practices, including wearing eye protection, gloves, and a laboratory coat to avoid contact with eyes, skin, and clothing.

Preparation of Reagents

Note: Be sure to use electrophoresis buffer, not water, to prepare agarose gels.

- 1. Prepare electrophoresis buffer for casting gels.** Tris-acetate-EDTA (TAE) electrophoresis buffer comes as a 50x concentrated solution that must be diluted to 1x for preparing agarose gels. To prepare 12 agarose gels, 750 ml of 1x TAE will be adequate. To make 750 ml of 1x TAE, add 15 ml of 50x TAE to 735 ml of distilled water. Additional 1x TAE will be needed to fill the electrophoresis chamber with electrophoresis running buffer. See step 3 for details.
- 2. Prepare and cast agarose gels.*** Gels may be prepared up to 2 weeks ahead of time by the instructor or during class by the individual student teams.
 - 2.1** To make a 1% agarose solution, use 1 g of agarose for each 100 ml of 1x TAE electrophoresis buffer. If the number of electrophoresis chambers is limited, you can use a 7 x 10 cm tray and two gel combs to pour a gel that can be used to run two sets of student samples. Depending on the number of samples to be run per gel, 8-well or 15-well combs can be used.
 - 2.2** Use the following table as a guide for gel volume requirements when casting single or multiple gels:

Number of Gels	Volume of 1% Agarose Needed for a 7 x 7 cm Tray, ml	Volume of 1% Agarose Needed for a 7 x 10 cm Tray, ml
1	40	50
4	160	200
12	480	600

- 2.3** Add the appropriate amount of agarose powder to a suitable container with plenty of room to allow liquid to boil and be swirled (for example, use a 500 ml Erlenmeyer flask for preparing 200 ml or less of agarose solution). Add the appropriate amount of 1x TAE electrophoresis buffer and swirl to suspend the agarose powder in the buffer. If using an Erlenmeyer flask, invert a smaller flask into the open end of the 500 ml flask containing the agarose. The small flask acts as a reflux chamber, thus allowing long or vigorous boiling without much evaporation.

Caution: Always wear protective gloves, goggles, and laboratory coat while preparing and casting agarose gels. Boiling agarose solution and the flasks containing hot agarose can cause severe burns if allowed to contact skin.

The agarose solution can be prepared for gel casting by boiling until the agarose has melted completely on a hot plate or in a microwave oven. Use of a microwave oven is the fastest and safest way to dissolve agarose. To prepare the agarose solution in a microwave oven, place the flask or bottle containing the 1x TAE buffer and agarose powder into the microwave (always loosen the cap if you are using a bottle). Use a medium setting and set the timer to 3 min. Stop the microwave oven every 30 sec and swirl the flask or bottle to suspend any undissolved agarose. Boil and swirl the solution until all of the small transparent agarose particles are dissolved.

- 2.4** Cool agarose solution to 55–60°C (a water bath is useful for this step).
- 2.5** Prepare the gel casting apparatus and pour the molten agarose into the gel casting tray containing the comb(s). Allow the agarose to solidify at room temperature for 15–20 minutes.
- 2.6** Carefully remove the comb(s) from the solidified gel. Agarose gels can be stored wrapped in plastic wrap and sealed in plastic bags for up to 2 weeks at 4°C.

* Convenient precast agarose gels (catalog #1613015EDU (8 well) and #1613057EDU (2 x 8 well)) are available from Bio-Rad. These are 1% TAE gels that fit into Bio-Rad's Mini-Sub Cell GT cell or any horizontal gel electrophoresis system that fits 7 x 10 cm gels.

3. Prepare electrophoresis running buffer for filling the electrophoresis chamber.

Concentrated TAE buffer must also be diluted to 1x for use as an electrophoresis running buffer. There are two ways to run and stain gels:

Conventional electrophoresis — this method uses running buffer at **1x concentration** with the gels run at **100 V for 30 min**. Each Bio-Rad Mini-Sub Cell electrophoresis chamber requires ~275 ml of buffer. To prepare enough 1x TAE for 12 chambers, add 70 ml of 50x TAE to 3.43 L of distilled water and mix. For one chamber, add 5.5 ml of 50x TAE to 270 ml of distilled water. The 1x electrophoresis buffer can be saved after use and reused 3–4 times.

The fast gel protocol — this method uses a reduced concentration of running buffer (**0.25x TAE**), so gels can be run at **200 V in less than 20 min**. In this protocol, the agarose gels are still prepared using 1x TAE. (Preparing gels with 0.25x TAE will result in loss of DNA resolution. See Bio-Rad Bulletin 5396 for more information.) To prepare enough 0.25x TAE for 12 electrophoresis chambers, add 17.5 ml of 50x TAE to 3.48 L of distilled water and mix. Monitor migration of the loading dye from the samples during electrophoresis to prevent running the gel for too long.

4. Prepare the electrophoresis chamber.

4.1 Place the gel tray into the electrophoresis chamber and make sure the sample wells are at the black cathode end.

4.2 Add running buffer to the chamber to cover the gel with 1–2 mm of buffer (~275 ml of running buffer).

5. Load samples. Use a separate pipet tip for each sample.

5.1 Load samples containing loading dye into the wells of the gel in the order specified in your electrophoresis plan. The loading dye makes samples heavier than the buffer, so the samples should drop easily into the wells of the gel.

5.2 Once all samples are loaded, secure the lid on the gel box, ensuring the red and black electrodes are properly orientated and connect the electrical leads to the power supply.

5.3 Turn on the power supply. See step 3 for running conditions dependent on the formulation of the electrophoresis running buffer.

6. When electrophoresis is complete, turn off the power supply and remove the lid from the electrophoresis chamber. Visualize DNA according to directions from your instructor.

APPENDIX A2: MICROBIAL CULTURING METHODS

It is essential that proper sterile technique is used for all microbiological manipulations in this laboratory. This includes but is not limited to decontaminating work surfaces, keeping sterile solutions covered, preventing contact between nonsterile surfaces and sterile loops, pipet tips, or tubes, and disposing of all contaminated materials properly according to local EHS regulations. With respect to media preparation and sterilization, an autoclave is much more effective at sterilization than a microwave oven and media sterilized using a microwave oven is more prone to contamination.

Microbial Culturing Reagent Preparation Checklist

Components from Cloning and Sequencing Explorer Series	Where Provided	(✓)
Ampicillin, lyophilized	Microbial Culturing Module	<input type="checkbox"/>
LB broth capsules	Microbial Culturing Module	<input type="checkbox"/>
LB nutrient agar powder	Microbial Culturing Module	<input type="checkbox"/>
IPTG, 1 M, 0.1 ml	Ligation and Transformation Module	<input type="checkbox"/>
Petri dishes, 60 mm, sterile	Microbial Culturing Module	<input type="checkbox"/>
Cell culture tubes, 15 ml, sterile	Microbial Culturing Module	<input type="checkbox"/>
Inoculation loops, sterile	Microbial Culturing Module	<input type="checkbox"/>
<i>E. coli</i> strain HB101 K-12, lyophilized	Microbial Culturing Module	<input type="checkbox"/>
Disposable plastic transfer pipets	Microbial Culturing Module	<input type="checkbox"/>

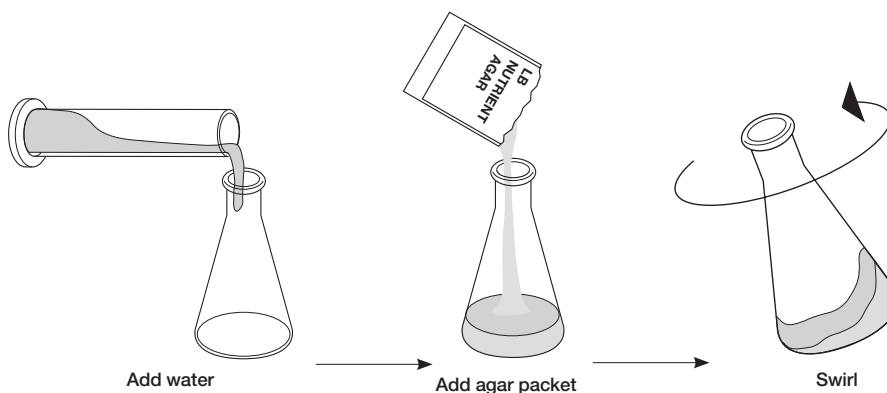
Required Accessories (Not Provided)	Quantity	(✓)
Autoclave, microwave, or hot plate, and stir bar	1	<input type="checkbox"/>
Flasks and bottles	1–12	<input type="checkbox"/>
200 µl adjustable-volume micropipet and sterile tips	1–12	<input type="checkbox"/>

Preparation of Ampicillin

Ampicillin is an antibiotic used to prevent growth of bacteria that have not been transformed with plasmids containing an ampicillin resistance gene. Ampicillin is shipped freeze-dried in a small vial containing 30 mg ampicillin. With a sterile pipet, add 3 ml of sterile water directly to the vial to rehydrate the antibiotic to make a 10 mg/ml of 200x solution. Ampicillin is used at a final concentration of 50 µg/ml in both LB Amp broth and LB Amp IPTG agar. Ampicillin should be stored at –20°C and is good for 1 year.

Preparation of Agar Plates

This protocol is used to prepare solid LB agar media for the growth of bacteria. Prepare plates at least 3 days prior to the transformation laboratory. Each student team requires one LB agar plate and two LB Amp IPTG agar plates. Agar plates should be prepared at least 2 days before required, left out at room temperature for 2 days and then refrigerated until they are used. Two days on the benchtop allows the agar to dry out and more readily take up the liquid transformation solution.



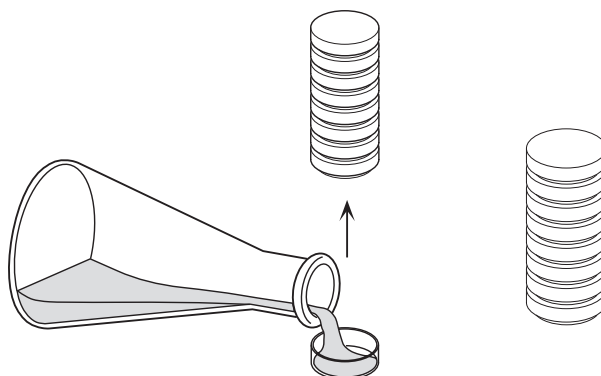
1. **To prepare 500 ml LB agar, add the entire contents of one LB agar packet to 500 ml of distilled water in a 1 L or larger flask and cover the flask opening.** Swirl to dissolve the agar, or add a magnetic stir bar to the flask and stir on a stir plate. A stir bar will also aid in mixing the solution completely once sterilization is complete. Autoclave LB agar on wet cycle for 30 min. Once the autoclave cycle is complete, check the solution to ensure that the agar is all dissolved.

Note: Be sure to wear appropriate protective equipment and be careful to allow the flask to cool slightly before swirling so that the hot medium does not boil over onto your hand.

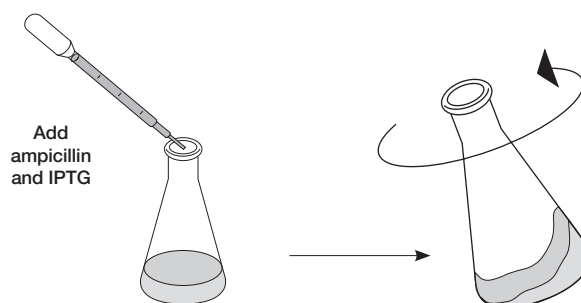
If no autoclave is available, heat the agar solution to boiling in a microwave or on a hot plate. Repeat heating and swirling about three times until all the agar is dissolved (that is, no more clear specks swirl around). Use a reduced power setting on the microwave to reduce evaporation and boiling over.

Allow the LB agar to cool so that the outside of the flask is just comfortable to hold (~55°C). A water bath set at 55°C is useful for this step. Be careful not to let the agar cool so much that it begins to solidify. Solidified agar can be reheated to melt it if antibiotics have not been added.

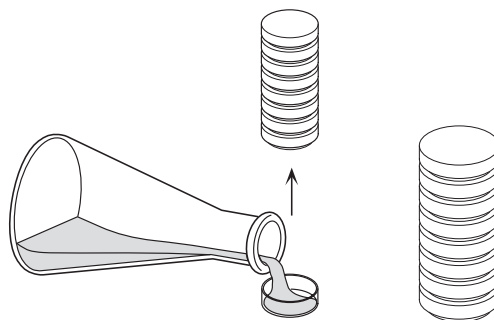
2. **While the agar is cooling, label the plates.** Label the outside of the lower plates rather than the lids to avoid confusion when the plates are opened.
3. **Pour the LB agar plates first.** One LB agar plate for each team is required — 12 in total. Stack empty plates 4–8 high and starting with the bottom plate lift the lid and the upper plates straight up and to the side with one hand and pour the LB agar with the other. Fill the plate about one third to one half (~10 ml) with agar, replace the lid, and continue up the stack. Let the plates cool and solidify in this stacked configuration. Do not disturb them until the agar has solidified. Avoid getting bubbles in the plates if possible — bubbles can be removed while the agar is still liquid by carefully flaming the plates with a Bunsen burner.



4. **Add ampicillin and isopropyl b-D-1-thiogalactopyranoside (IPTG) to LB agar that has cooled 5 to at least 55°C.** IPTG is used to increase protein expression induced by the lac operon. The final concentration of ampicillin in LB Amp IPTG agar should be 50 µg/ml and the final concentration of IPTG should be 0.2 mM. Add 2.5 ml of 10 mg/ml ampicillin stock and 100 µl of 1 M IPTG to the remaining molten LB agar. It is okay for the ampicillin and IPTG to be at a slightly higher concentration than stated. Ensure LB agar is cooled to 55°C before adding ampicillin and IPTG, as excessive heat will degrade the reagents. Swirl or use the stir bar and a stir plate to mix the ampicillin and IPTG into the agar, taking care not to introduce bubbles.



5. **Pour the LB Amp IPTG agar plates.** Two LB Amp IPTG plates are required for each team — so 24 in total.



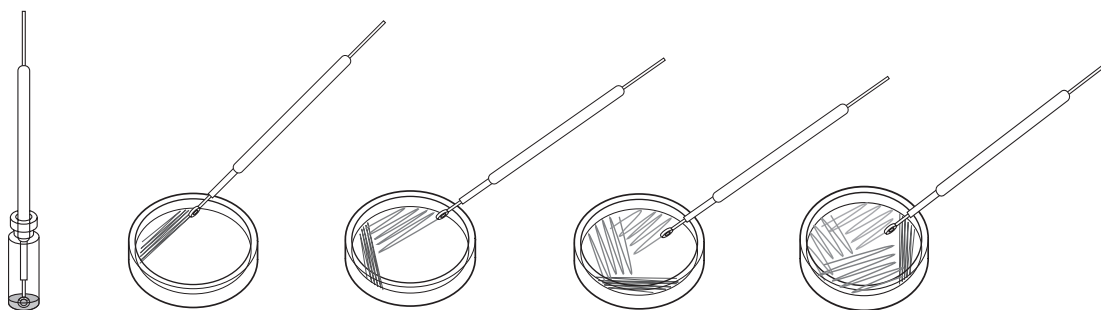
6. **After the plates have dried at room temperature for two days they should be stored at 4°C enclosed in plastic bags or plastic wrap to prevent the plates drying out.** Plates are good for 2 months. Pour excess agar in the garbage, not the sink. Wipe any agar drips off of the sides of the plates.

***E. coli* Starter Plates**

Agar plates are streaked with bacterial cultures to form single colonies. Each colony grows from one bacterium and thus a colony is a “clone,” or group of genetically identical individuals. A tiny drop of the original bacterial suspension contains millions or billions of individual bacteria and must be diluted multiple times to isolate single bacteria. Under favorable conditions *E. coli* can double every 20 minutes and thus a single bacterium will multiply to become billions of genetically identical cells in less than 24 hours.

At least 2 days prior to the transformation laboratory, each student team should streak one LB agar plate with *E. coli* to form single colonies.

- 1. Rehydrate the stock vial of HB101 *E. coli* bacteria with 250 μ l of sterile water, recap, and shake to mix, then let stand at room temperature for 5 min.**
- 2. Insert a sterile inoculation loop into the vial of bacteria.** Remove the loop and check that there is a liquid film across the loop — this is a volume of 10 μ l. Gently rub the loop back and forth over the agar in the top left hand corner as shown below. The first streak dilutes the cells. Go back and forth with the loop about a dozen times in the first quadrant. Do not break the surface of the agar.



- 3. For subsequent streaks, the goal is to use as much of the surface area of the plate as possible.** Rotate the plate about 45 degrees (so that the streaking motion is comfortable for your hand) and start the second streak. **Do not dip the loop into the vial of bacteria again.** Go into the previous streak one or two times and then back and forth as shown about a dozen times.

In subsequent quadrants the cells become more and more dilute, increasing the likelihood of producing single colonies. Remember a single colony arose from one cell and all the cells in the colony are genetically identical.

- 4. Rotate the plate again and repeat streaking into the third quadrant.**
- 5. Rotate the plate again and make the final streak. Do not touch the first quadrant.**
- 6. Place the plates upside down inside the incubator for 16–24 hr at 37°C.**
- 7. Once bacteria have grown, wrap edges of plates with Parafilm to make an airtight seal and store at 4°C. Colonies will remain viable for up to 2 weeks.**

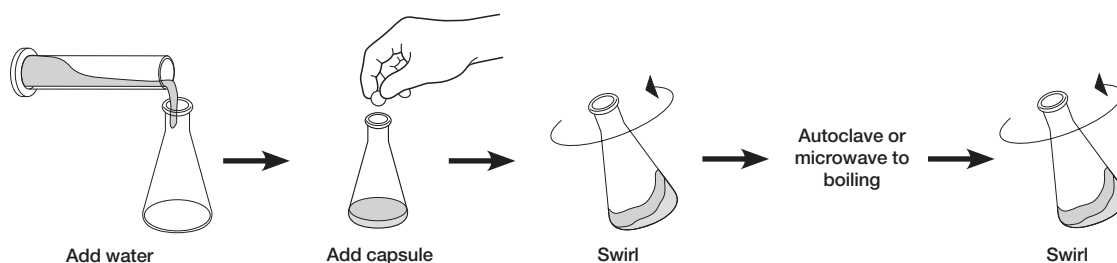
Preparation of LB Broth

This protocol is used to prepare liquid LB broth for growth of bacteria. Each student team requires at least 25 ml of LB broth, of which 5 ml will be used for a starter culture and 20 ml will be made into LB Amp broth. LB broth has been supplied in a convenient capsule form; each capsule makes 50 ml of LB broth when reconstituted with water. To avoid contamination of a common stock of LB broth, it is recommended that each capsule be reconstituted with 50 ml of water in a separate container. To make 50 ml of sterile LB broth, label a clean container and add one capsule of LB and 50 ml of distilled water and cover appropriately.

Note: never completely seal a container that is to be autoclaved.

Autoclave the container on the wet cycle for 30 min. Allow the broth to cool to room temperature before use. If an autoclave is not available, LB can be sterilized in the microwave by heating to boiling at least three times or after heating to dissolve the LB capsules, the solution can be filter sterilized through a 0.2 μ m filter.

Note: evaporation can be reduced by using a low power setting on the microwave so the solution simmers rather than boils.



Storage of LB broth at 4°C is recommended and LB can be stored for up to 1 year. LB broth can also be stored at room temperature if desired — however this is not recommended if the microwave method has been used for sterilization, or if the bottle has been opened after sterilization. LB broth containing antibiotics should be stored at 4°C and is good for two months.

Preparation of LB Amp Broth

This protocol is used to prepare liquid LB broth with ampicillin which will select for the growth of bacteria transformed with plasmids containing an ampicillin resistance gene. Each student team requires at least 15 ml of LB Amp broth. The final concentration of ampicillin in LB Amp broth should be 50 μ g/ml. To make 20 mls of LB Amp broth, using sterile technique add 100 μ l of 10 mg/ml ampicillin stock. Ensure LB broth is cooled to room temperature before adding ampicillin, as excessive heat will degrade the antibiotic. LB Amp broth should be stored at 4°C and is good for two months.

Growing Bacterial Cultures for Plasmid Minipreps

This protocol is used to grow small cultures of bacteria in LB Amp broth, which will select for the growth of bacteria transformed with plasmids containing an ampicillin resistance gene. The small cultures are often processed to isolate plasmid DNA for experiments and activities in molecular biology.

1. **Prepare LB Amp broth as described above.**
2. **Label an appropriate number of 15 ml culture tubes.**
3. **Using sterile technique, pipet 3 ml of LB Amp broth into each 15 ml culture tube.**
4. **One day prior to the plasmid purification laboratory session, use a sterile pipet tip or sterile loop to pick a single colony and inoculate an LB Amp culture tube by swirling the loop containing the single colony in the broth.** If the colonies are very small or tightly packed, a sterile pipet tip may be better than a sterile loop to pick single colonies. Repeat for each miniprep culture, each time using a different colony.
5. **Cap the culture tubes loosely.** Culture tubes have two lid positions — loose to allow air into the culture that is necessary for bacterial growth, and tight to store bacteria once they have grown.

Note: Occasionally satellite colonies may grow surrounding the real colonies so it is important to pick the large individual colonies instead of the tiny colonies surrounding larger colonies. Be sure that a single colony is picked, otherwise you may isolate multiple plasmids from your miniprep and these cannot be sequenced.

6. **Place the miniprep cultures to grow overnight (20–28 hr) at 37°C in a shaking water bath or incubator shaking with a speed of 275 rpm.** The liquid cell culture should be shaken vigorously to provide a sufficient amount of oxygen to the dividing cells. Cells may be grown at room temperature with shaking but will require more time to grow.

Note: Shaking at the indicated speed is important to properly aerate the culture. Bacteria will not grow to a high enough density if they are not properly aerated. Growing cultures with additional surface-to-air ratios, such as in wider tubes, or using less culture medium and splitting the miniprep can increase yield if shaking speeds are too low.

7. **Plasmid DNA can be isolated from minipreps using a plasmid purification protocol.** It is best to use freshly grown bacterial cultures for plasmid isolation. However, if storage is necessary, cultures may be stored at 4°C for up to 1 week — a reduction in DNA yield may be experienced.

Note: Larger bacterial cultures, sometimes called maxipreps, can be grown from miniprep cultures. To grow a maxiprep culture, add 1/100 of the final maxiprep volume of the miniprep culture to the required volume of LB Amp broth in a flask that is five times the culture volume. For example, add 1 ml of miniprep to 100 ml of LB Amp broth in a 500 ml flask for a 100 ml maxiprep. Grow cultures in a shaking water bath or incubator at 37°C at a speed of 275 rpm for at least 6 hr.

APPENDIX A3: STERILE TECHNIQUE FOR PCR

PCR is a powerful and sensitive technique that enables researchers to produce large quantities of specific DNA from very small amounts of starting material. Because of this sensitivity, contamination of PCR reactions with unwanted DNA is always a possible problem. Therefore, utmost care must be taken to prevent cross-contamination of samples. Steps to prevent contamination and failed experiments include:

- 1. Use of aerosol barrier pipet tips.** The end of the barrels of micropipets can easily become contaminated with aerosolized DNA molecules. Pipet tips that contain a filter at the end can prevent aerosol contamination from micropipets. DNA molecules that are found within the micropipet cannot pass through the filter and contaminate PCR reactions. Xcluda aerosol barrier pipet tips (catalog #2112006EDU and 2112016EDU) are ideal pipet tips to use in PCR reactions. For this laboratory, aerosol barrier tips should be used until the second nested PCR step has been completed, including electrophoresis of the initial PCR products if this is performed prior to the nested PCR step.
- 2. Aliquoted reagents.** Sharing of reagents and multiple pipetting into the same reagent tube can easily introduce contaminants into your PCR reactions. When at all possible, divide reagents into small aliquots for each team or, if possible, for each student. If only one aliquot of a reagent does become contaminated, then only a minimal number of PCR reactions will become contaminated and fail.
- 3. Changing pipet tips.** Always use a new pipet tip when entering a reagent tube for the first time. If a pipet tip is used repeatedly, contaminating DNA molecules on the outside of the tip will be transferred to other solutions, resulting in contaminated PCR reactions. If you are at all unsure if your pipet tip is clean, err on the safe side and discard the tip and get a new one. The price of a few extra tips is a lot smaller than the price of failed reactions.
- 4. Use good sterile technique.** When opening tubes or pipetting reagents, leave the tubes open for as little time as possible. Tubes that are open and exposed to the air can easily become contaminated by aerosolized DNA molecules. Go into reagent tubes efficiently, and close them as soon as you are finished pipetting. Also, try not to pick tubes up by the rim or cap, as you can easily introduce contaminants from your fingertips.
- 5. Bleach at a concentration of 10% destroys DNA, so wiping down surfaces and rinsing plastic pipet barrels, mortars, and pestles with 10% bleach can get rid of any surface DNA contamination that may arise.**

APPENDIX B: ADDITIONAL BACKGROUND ON GAPDH

The Role of GAPDH in Carbon Metabolism

Members of the GAPDH family of enzymes are found in all cells. GAPDH carries out the sixth step of the ten enzymatic steps in the universal process of glycolysis (the breakdown of glucose to pyruvate to produce energy and reducing power for cells). As noted in the General Background, recent research has found that GAPDH plays many roles outside of glycolysis as well, suggesting many avenues for future research. The focus in this laboratory activity, however, will be on the central process of glycolysis, the mechanism of the specific reaction carried out by the GAPDH enzyme, and an overview of some of the known diversity within this family of enzymes.

Plants are considered to be the primary producers of energy on the planet because they carry out the most important biochemical process on earth: photosynthesis. During photosynthesis, sunlight provides plants with the energy to forge a myriad of important chemical bonds between atoms of carbon, oxygen, and hydrogen, ultimately synthesizing energy-rich molecules called carbohydrates.

Carbohydrate molecules are used by plants in two ways: first as building blocks for making more complex macromolecules (like DNA, enzymes, cellulose, cellular membranes, and the cytoskeleton), and second as the energy source for synthesizing macromolecules and dynamic processes like growth, development, and responses to the environment. Since animals, fungi, and most bacteria are not capable of photosynthesis, they are totally dependent on plants for their carbohydrates.

Carbohydrates and the energy from carbohydrates are stored in many different forms. The monosaccharides glucose and fructose provide sweetness to most fruits. Plants transport carbohydrates through their phloem in the form of sucrose, a disaccharide molecule resulting from a bond forming between a glucose and a fructose molecule. A few plants, such as sugarcane, store their excess glucose as sucrose, however most plants store their excess carbohydrates in the form of the polysaccharide starch; this is most evident in the seeds used for human nutrition, such as wheat and beans. In addition, a few plants, such as soy and sesame, store energy in their seeds as lipids (oils).

Animals store excess carbohydrate in the form of glycogen, a glucose polymer that is similar to starch. A large amount of glycogen gets stored in the liver and is readily broken down to glucose and released to the bloodstream between meals to keep up blood sugar levels. Animals have an alternative source of energy as well. The fatty acids that make up fats and oils in the body can be broken down to provide acetyl coenzyme A molecules for feeding into the citric acid cycle.

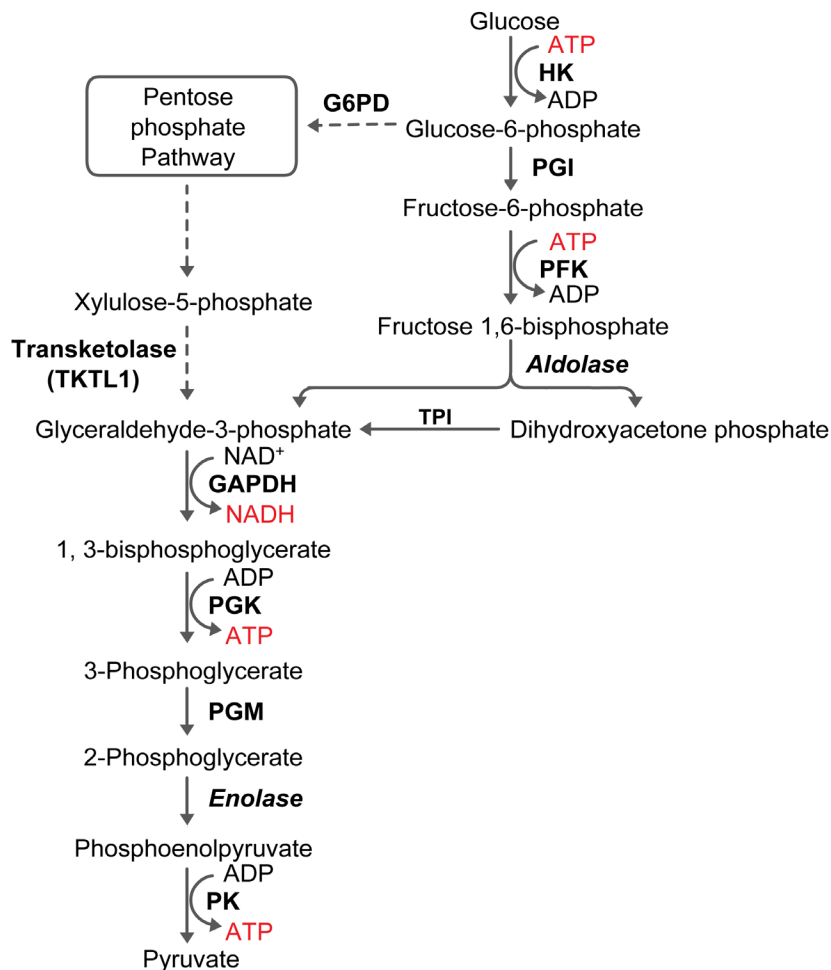
In plant cells, glucose can be synthesized in many ways: directly from CO₂ in the Calvin cycle of photosynthesis, converted from fructose by modifying chemical bonds in the six-carbon sugar, derived by breaking down sucrose or starch into their component monosaccharides, or generated from lipids through the process of gluconeogenesis (the reverse process to glycolysis).

Whenever biochemical energy is needed by a cell, whether it is in a plant or in an animal that has just eaten a plant, the process by which the cell produces that energy is nearly the same: respiration. Although the process of respiration is incredibly complex, it is also highly standardized among the different forms of life that compose the biosphere. Plants, animals, fungi, and bacteria are surprisingly uniform in how they convert the carbohydrate energy into adenosine triphosphate (ATP), the real chemical currency of the cell.



Chemical reaction of respiration. Through the process of aerobic respiration, one glucose and six oxygens make 36 ATP molecules, as well as six molecules each of CO₂ and water.

Aerobic respiration is composed of four main stages: glycolysis, formation of acetyl coenzyme A, the citric acid cycle, and the combined processes of electron transport and chemiosmosis. Glycolysis is the most conserved stage of respiration in the biological world. Cells undergoing anaerobic respiration due to lack of oxygen, like fermenting yeast or vigorously exercised muscles, carry out only glycolysis. Prokaryotic organisms (bacteria) that do not contain mitochondria (where the latter three stages occur in eukaryotes) still carry out glycolysis. Glycolysis generates energy by catabolizing glucose into ATP, reducing agents (NADH or NADPH), and the three-carbon organic acid pyruvate. Intermediates of glycolysis also form the building blocks for numerous anabolic reactions. For instance, glucose-6-phosphate is a precursor for the synthesis of ADP, NAD⁺, and coenzyme A; phosphoenolpyruvate is a precursor for the synthesis of the three aromatic amino acids (tyrosine, phenylalanine, and tryptophan), which are in turn used to synthesize various flavonoids, alkaloids, and hormones. For these reasons, glycolysis is considered to be the central metabolic pathway in plants, microbes, and animals. This explains why glycolysis was the first major metabolic pathway to be thoroughly elucidated by biochemists (Kresge et al. 2005).



Glycolysis pathway. Abbreviations: HK (hexokinase), PGI (phosphoglucose isomerase), PFK (phosphofructokinase), TPI (triose phosphate isomerase), GAPDH (glyceraldehyde-3-phosphate dehydrogenase), PGK (phosphoglycerate kinase), PGM (phosphoglyceromutase), and PK (pyruvate kinase).

Process of Glycolysis

The process of glycolysis is generally the same in plants and animals and occurs as ten simple enzymatic steps catalyzing the breakdown of the six-carbon sugar glucose into the three-carbon pyruvate, also yielding ATP and the reducing agents NADH or NADPH. The first three enzymatic steps convert glucose into fructose-1,6-bisphosphate. The first enzyme (hexokinase) catalyzes the formation of glucose 6-phosphate by shifting the phosphate bond from ATP to the glucose. The second enzyme (phosphoglucose isomerase) causes a rearrangement in the molecular structure to form fructose-6-phosphate so that it can accept another phosphate bond, a reaction catalyzed by the third enzyme, phosphofructokinase. The phosphate for this third reaction also comes from ATP, the second ATP consumed by glycolysis. This expenditure of ATP in the early steps of respiration doesn't immediately make sense for a process whose overall purpose is to synthesize that molecule. The resulting product, fructose-1,6-bisphosphate, however, is such a high-energy molecule that when broken down in the later steps of glycolysis, its energy will be used to synthesize four molecules of ATP and two molecules of NADH, an energy-rich coenzyme. So overall, glycolysis is an energy-producing process.

The positioning of the phosphates in the first and last carbon positions of fructose-1,6-bisphosphate sets the stage for the fourth enzyme, aldolase, to pull the sugar apart, yielding two three-carbon molecules, each with one phosphate group. Dihydroxyacetone phosphate and glyceraldehyde-3-phosphate (GAP) have the same molecular formula ($C_3H_7O_6P$) and are, therefore, isomers. In the fifth step of glycolysis, the enzyme triose phosphate isomerase interconverts these two molecules.



The GAPDH enzyme reaction. GAPDH catalyzes the conversion of glyceraldehyde 3-phosphate (GAP), nicotinamide adenine dinucleotide (NAD^+), and inorganic phosphate (P_i) into 1,3-bisphosphoglycerate (BPG), NADH and a proton.

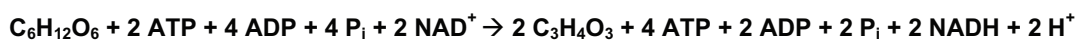
GAP is invaluable to the second part of glycolysis. One of the unique features of GAP is its ability to have another phosphate added to it without having to consume ATP. The phosphorylation of GAP by GAPDH in the sixth step of glycolysis does not consume ATP differently from phosphorylation in the earlier steps of glycolysis. In addition, the GAPDH reaction generates a molecule of NADH (or NADPH), a coenzyme that provides reducing power to hundreds of different enzymes involved in catalyzing oxidation-reduction reactions in the cell.

In the seventh step of glycolysis, the BPG produced by GAPDH loses one of its two phosphates during the formation of ATP from ADP by phosphoglycerate kinase. The eighth enzyme, phosphoglyceromutase, transfers the other phosphate from the third carbon to the second carbon position, allowing the ninth step, which is reversible dehydration by enolase, to form phosphoenolpyruvate. Finally, phosphoenolpyruvate has its phosphate transferred to ADP by the enzyme pyruvate kinase to yield ATP and pyruvate.

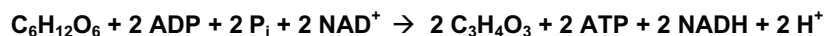
In plants growing where oxygen levels are normal, pyruvate is transported into mitochondria via specific carriers in the inner membrane for aerobic respiration. Once in the matrix of the mitochondria, pyruvate is oxidized and decarboxylated to form acetyl coenzyme A, the initial substrate for the citric acid cycle, as well as CO_2 and ATP. Oxygen deprivation can occur in plants when roots or seeds are flooded by water. Under such conditions, pyruvate can be converted into acetaldehyde and then ethanol, lactate, or alanine. The exact anaerobic product that results depends on the species, the tissue, and level of oxygen deprivation (Bray et al. 2000).

The GAPDH reaction is the only step in glycolysis that produces NADH. NAD^+ , a coenzyme that is synthesized from nicotinic acid (also known as niacin or vitamin B3), is an important carrier of electrons in many oxidation-reduction reactions. More than 500 different enzymes require the electron-carrying capacity of NAD^+ or NADP^+ (Kasimova et al. 2006). Both NADH or NADPH (reduced coenzymes) are important to the later stages of aerobic respiration because they provide the electrons for the electron transport chain located in the mitochondria.

The reaction and the overall net reaction for glycolysis are shown below. For each molecule of glucose, glycolysis produces two pyruvates, two molecules of ATP, two molecules of NADH, and two protons. The energy-rich ATP and NADH can be used in the later respiratory reactions, but can also be used in the cytoplasm for other purposes.



The reaction for glycolysis. Glycolysis uses one glucose, two ATP, four ADP, four inorganic phosphates and 2 NAD⁺ molecules to generate two pyruvates, four ATP, 2 ADP, 2 inorganic phosphates, two NADH and two proton molecules.



The net reaction for glycolysis. For each molecule of glucose, glycolysis produces two pyruvates, two molecules of ATP, two molecules of NADH, and two protons. The energy-rich ATP and NADH can be used in the later respiratory reactions, but can also be used in the cytoplasm for other purposes.

GAPDH Reaction Mechanism

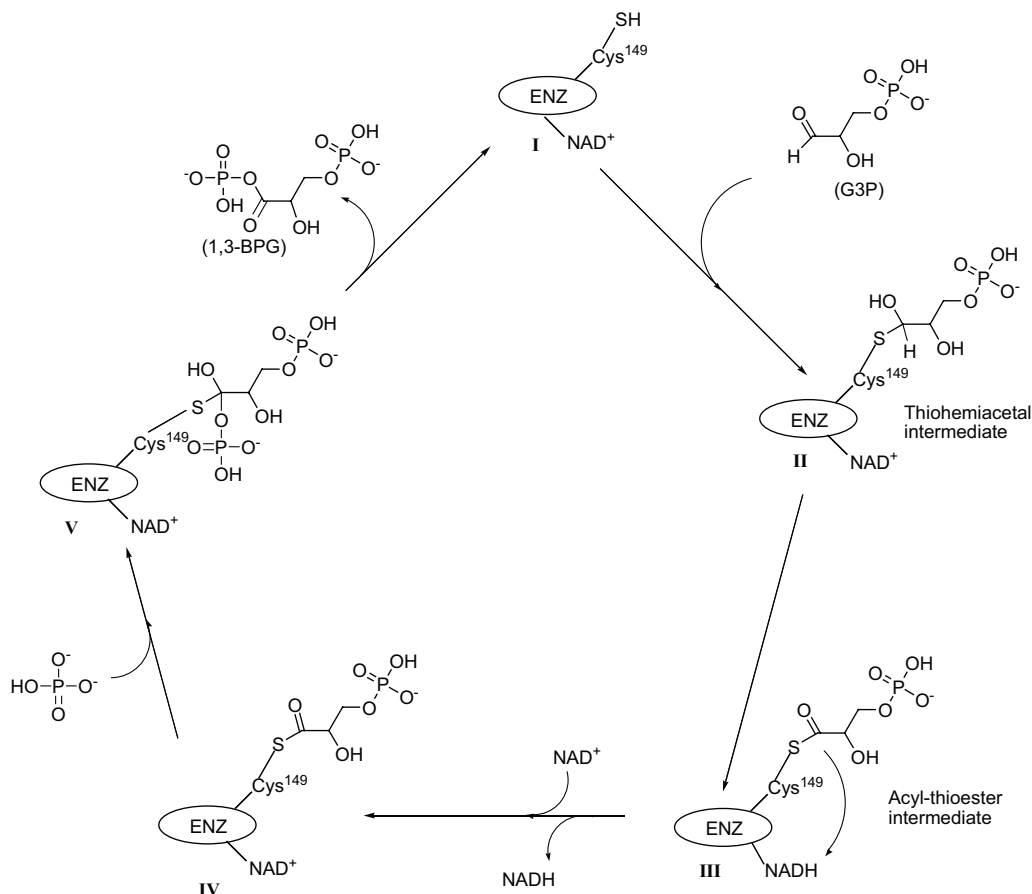
GAPDH makes up around 10% of the soluble protein in the skeletal muscles of animals, which has permitted biochemists to purify the protein for detailed structural and functional studies (Walsh 1979).

Note: GAPDH proteins are highly conserved between plants and animals, and so structural and functional information derived from animal GAPDH is applicable to plant GAPDH; however, some of the amino acid positional information is not precisely the same.

The GAPDH in skeletal muscles is composed of four identical subunits (a homotetramer), with each subunit containing an active site. GAPDH protein can efficiently bind to NAD⁺ and NADH. The NAD⁺-binding domains are dispersed throughout the sequence and position NAD⁺ near the active site of the enzyme, which binds the GAP substrate.

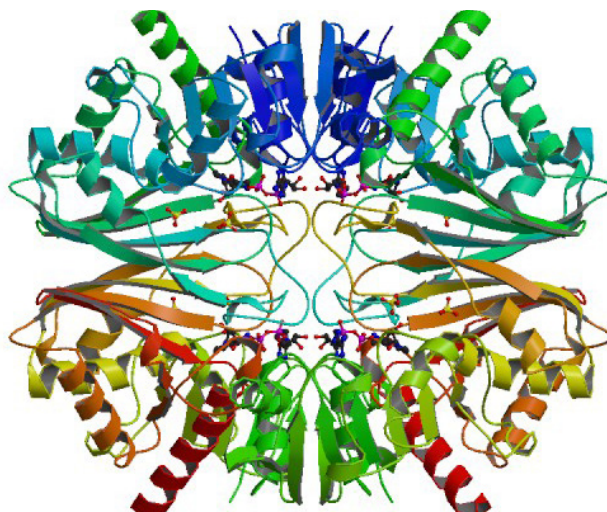
The active site of the animal skeletal muscle GAPDH is centered on the 149th amino acid of the GAPDH polypeptide subunit, which is a cysteine.

The reaction mechanism for GAPDH is shown in the figure below (Biesecker et al. 1977, Walsh 1979).



GAPDH reaction mechanism. In step I of the figure, the GAPDH enzyme with an NAD⁺ bound to it. The catalytic site is indicated by the sulfur group of the amino acid cysteine (Cys¹⁴⁹). Step II occurs in the presence of the substrate, GAP. The carbonyl group (C=O) of GAP is converted to a hydroxyl group (C-OH) and a new bond is made with the sulfur of Cys¹⁴⁹. In step III, the hydrogen that is bound to carbon-1 of GAP is directly transferred to NAD⁺, leading to the reformation of a carbonyl at that first carbon. In this configuration, the affinity of GAPDH for NADH is relatively low, so NADH is released from the enzyme-substrate complex. In step IV, however, NADH is immediately replaced by another molecule of NAD⁺. The enzyme-substrate complex shown in step IV is capable of temporarily binding to inorganic phosphate, the other major substrate of this reaction. The formation of the larger complex shown in step V is transitory and results in the release of the now-phosphorylated BPG, the phosphorylated version of GAP. This final step regenerates the original enzyme seen in step I. These five steps are what biochemists call "transition states" and occur very quickly, perhaps millions of times per second, and are determined using a collection of experimental techniques, including X-ray crystallography, spectrophotometry, and assays using inhibitors and radioactive isotopes, as well as amino acid sequence determination.

The binding of the GAP substrate by the GAPDH enzyme actually occurs at a different stage than the reduction of NAD⁺ to NADH, which is at a different stage than the phosphorylation step leading to the release of BPG in the final stage. Although Cys¹⁴⁹ is critical to the first step in this reaction, other amino acids are important for the other interactions that take place. For instance, amino acids 148, 150, 176, 209, and 231 are important parts of the active site (Skarzynski et al. 1987), since they largely interact with each other and with the GAP and inorganic phosphate substrates via hydrogen bonding (Song et al. 1999). The carbon-3 of GAP is held by hydrogen bonding between its phosphate group and amino acids 179, 181, and 231. NAD⁺ is thought to be held in position by amino acids 8, 10, 11, 32, 96, and 313 (Biesecker et al. 1977, Clermont et al. 1993). The inorganic phosphate is held by amino acids 148, 150, and 208.



Structure of GAPDH bound to NAD⁺ as determined by X-ray crystallography. Structure can be downloaded from the Protein Databank (www.rcsb-org) using the pdb identifier 1szj.

The GAPDH reaction can also go in the reverse direction — dephosphorylating BPG. However, a normally functioning cell undergoing glycolysis at typical rates will be synthesizing BPG because the phosphoglycerate kinase reaction of glycolysis consumes BPG at such a rate that the equilibrium is shifted in the glycolytic direction (Dennis and Blakely 2000). In other words, there is such a drain on the supply of BPG in a respiring cell that there is not enough BPG for GAPDH to use to make GAP. BPG synthesis by GAPDH is also favored because there is a higher concentration of NAD⁺ than NADH in the cytoplasm.

Gluconeogenesis

Glycolysis is critical to the life of an organism; as such it is delicately regulated by complex and internal feedback mechanisms. As with the GAPDH reaction, most of the reactions of glycolysis are reversible and are also involved in synthesizing sugar from small carbon precursors such as pyruvate, lactate, glycerol, and certain amino acids. This reverse glycolytic process is called gluconeogenesis and involves some of the same enzymes as glycolysis, as well as other enzymes that catalyze parallel reactions. Glycolysis and gluconeogenesis are reciprocally regulated, so that only one will predominate in a given cell at any specific time.

The human body has only about a 12–24 hour supply of glycogen stored. With vigorous exercise or fasting, the body consumes the glycogen even faster. Once the glycogen is depleted, the parts of the body that require large amounts of glucose will depend on gluconeogenesis to supply it. In animals, the gluconeogenic pathway commonly occurs in the liver and kidneys, which produce glucose that then circulates to the other parts of the body that demand a lot of it, like the muscles and the brain. Gluconeogenic synthesis of glucose from organic acids and amino acids is an important consideration for anyone who is dieting to lose weight, as it results in not only loss of fat, but also loss of muscle mass. Seven of the ten enzymes involved in gluconeogenesis are the same enzymes that are used in glycolysis except that they catalyze the reverse reaction. In those three situations where a glycolytic reaction is not reversible, such as that catalyzed by hexokinase, a different enzyme (in this case, glucose-6-phosphatase) carries out the gluconeogenic reaction. Gluconeogenesis does occur in plants. For instance, lipids are converted into sugars in seeds that are high in oils (like soybean or sesame) and in the senescing leaves of deciduous plants during autumn.

Different GAPDH Isozymes

Glycolysis can occur in two different places in a plant cell: the cytosol and the plastid. When the source of carbohydrate is a sugar, glycolysis will occur in the cytosol of the cell, as described by most introductory biology textbooks. When the source of carbohydrate is starch, which is composed of long networks of glucose molecules, glycolysis can occur in the plastids (Plaxton 1996). All starch is synthesized in the plastids of plant cells. In green leaf and stem cells, the starch accumulates in the chloroplasts during the day, to store excess glucose. It is thought that the appearance of starch grains in the chloroplasts during the latter part of the day interferes with the ability of the thylakoid membrane to carry out photosynthesis. At night, however, when those same cells are metabolizing, growing, and shipping carbohydrate to other parts of the plant, such as the roots, the starch grains shrink or even disappear, thus allowing for another day of photosynthesis when the sun rises again. Plant cells that are not green also contain plastids. For instance, amyloplasts are plastids that specialize in storage of starch grains. Amyloplasts do not contain thylakoid membranes and can be found in the roots, the inner regions of the stem, and other non-green parts of the plant. Starchy foods, like potatoes or seeds, are tissues very rich in amyloplasts.

The breakdown of starch via glycolysis in the plastids is carried out by isozymes, enzymes that catalyze essentially the same biochemical reaction but are encoded by separate genes. Isozymes are very common in plants and animals, and typically result from a gene duplication event. The GAPDH isozymes found in chloroplasts are nuclear-encoded and are thought to have resulted from an endosymbiotic transfer of an ancient chloroplast gene to the nucleus (Meyer-Gauen et al. 1994, Liaud et al. 1990, Figge et al. 1999). GAPDH isozymes in plants are encoded by a small family of genes (Russell and Sach 1991, Pérusse and Schoen 2004). Some of the genes are very similar to one another, while others are more divergent. These differences show up not only in the DNA sequences of the genes, but also in levels of gene expression under different environmental conditions (for example, anaerobic conditions or high temperatures). Glycolysis is particularly sensitive to variables like oxygen level, nutrient level, temperature, and osmotic stress.

There are four different GAPDH enzyme isozymes in plants, all of which are encoded by nuclear genes:

- The NAD⁺-dependent enzyme (encoded by the *GAPC* gene) is found in the cytosol and is intimately involved in glycolysis
- Another NAD⁺-dependent enzyme (encoded by the *GAPCP* genes) is found in the chloroplasts of many species (Peterson et al. 2003, Meyer-Gauen et al. 1994)
- An NADP⁺-dependent GAPDH (encoded by the *GAPA* and *GAPB* genes) is located in the chloroplasts but is thought to be involved in the Calvin cycle of photosynthesis
- A nonphosphorylating NADP⁺-dependent GAPDH (encoded by the *GAPN* gene) is located in the cytosol

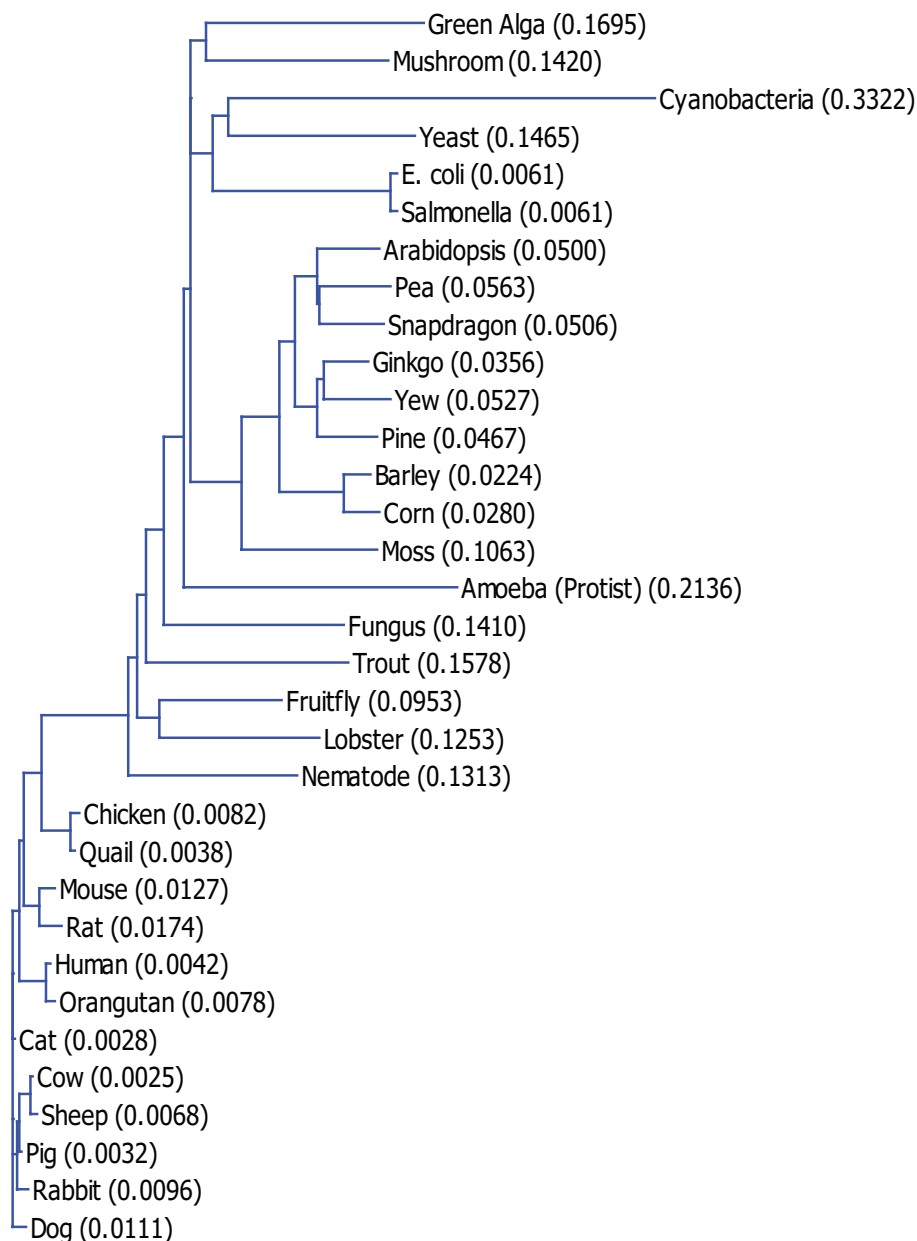
The NAD⁺-dependent GAPDH encoded by the *GAPCP* genes is similar to the one encoded in the cytosol but has a transit peptide sequence at the amino end of the polypeptide that allows its import into plastids. In lower plants (ferns, mosses, and gymnosperms), the NAD⁺-dependent GAPDH enzyme encoded by the *GAPCP* genes is present in photosynthesizing chloroplasts, whereas in higher plants (angiosperms), it is mostly found in nongreen plastids, where the starch is synthesized in roots, or where colorful pigments are synthesized, as in the fruit of red peppers. Although there are doubts about whether glycolysis occurs in the green chloroplasts of angiosperms, it is believed to take place in these nonphotosynthetic chloroplasts, thus requiring an NAD⁺-dependent GAPDH (Petersen et al. 2003). A nonphosphorylating NADP⁺-dependent GAPDH (encoded by the *GAPN* gene) is located in the cytosol.

The amino acid sequence of the nonphosphorylating NADP⁺-dependent GAPDH enzyme is quite different from that of the other enzymes, with only 15% of the amino acid sequence being the same as in the NAD⁺-dependent enzymes. The nonphosphorylating NADP⁺-dependent GAPDH enzyme appears to be important in plants that are deficient in phosphate. In spite of their variable locations and metabolic roles, all of these enzymes are encoded by nuclear genes. The *GAPDH* genes that are being isolated in this laboratory activity are the *GAPC* and *GAPC-2*, which encode the NAD⁺-dependent cytosolic GAPDH proteins. However, due to homology to *GAPC* and *GAPC-2*, *GAPCP* and *GAPCP-2* may also be isolated.

The NAD⁺-dependent GAPDH cytosolic enzyme is a tetrameric protein composed of four identical polypeptide subunits. The GAPDH subunits found in nature are surprisingly uniform in length and molecular mass, ranging from 330 to 340 amino acids in length and from 35,000 to 36,000 kD in molecular mass in species as diverse as bacteria, plants, birds, and mammals (Olsen et al. 1975).

Phylogeny Using GAPDH

The following figure shows a phylogenetic tree comparing the amino acid sequences of the NAD⁺-dependent GAPDH from numerous species. Plants and animals that are believed to be evolutionarily related based on other criteria such as morphology are similarly clustered together based on the amino acid sequence of the GAPDH proteins. Among the animals, for instance, mice and rats (both rodents) are clustered together, as are the two birds (chickens and quail), the two arthropods (fruit flies and lobsters), the two primates (humans and orangutans), and the two even-toed, cud-chewing ungulates (cows and sheep). Pigs (even-toed, non-cud chewers) are clustered near the other ungulates. In flowering plants, the monocots (barley and corn) are clustered together, but separately from the dicots (*Arabidopsis*, pea, and snapdragon). The gymnosperms (ginkgo, yew, and pine) are also clustered together. Two bacteria (*E. coli* and *Salmonella*) are clustered together near photosynthetic bacteria. Two fungi (a yeast and a mushroom) are near one another, but far from a third fungal species.

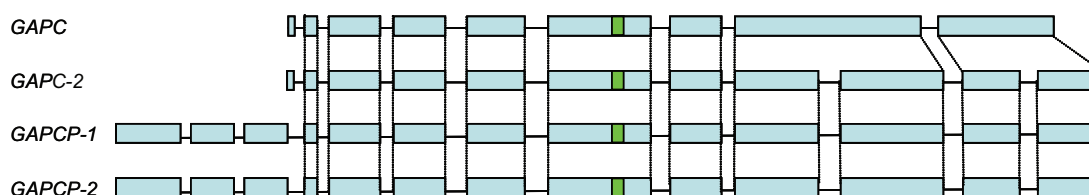


GAPDH phylogenetic tree. The phylogenetic tree is drawn from multiple sequence alignment of the NAD⁺-dependent GAPDH proteins from numerous species. Amino acid sequences that are most similar will be clustered together and connected by a shorter line. Data are available at NiceZyme view of EC 1.2.1.12 (www.expasy.ch/cgi-bin/enzyme-search-ec) on the ExPASy Proteomics server provided by the Swiss Institute of Bioinformatics.

The cytosolic NAD⁺-dependent GAPDH protein (EC 1.2.1.12, GAPC) is 335 amino acids long in humans and 338 amino acids long in *Arabidopsis*. The shortest known length of this protein is 329 amino acids (in common yeast) while the longest length is 363 amino acids (in the Egyptian jerboa, a rodent). The protein sequence of this cytosolic NAD⁺-dependent GAPDH enzyme in most other plants (such as peppers, potatoes, soybeans, ginkgo, and moss) is 70–80% identical to the sequence in *Arabidopsis*.

The chloroplast NAD⁺-dependent GAPDH protein (EC 1.2.1.12, GAPCP) is 420–422 amino acids long in *Arabidopsis*. The chloroplast GAPDH protein is 79 amino acids longer than the cytosolic protein because of the presence of the transit peptide. The remainder of the chloroplast GAPDH amino acid sequence is about 70% identical to the sequence of the cytosolic form in *Arabidopsis*.

The differences among the various GAPDH enzymes are more obvious at the DNA level. The genes of different species that encode GAPDH show variation for several reasons. First of all, the genetic code is degenerate, meaning that a single amino acid can be encoded by up to six different triplet combinations (codons). For instance, the amino acid glycine is encoded by either GGT, GGC, GGA, or GGG. Typically, it is the third nucleotide base of the triplet that varies. A second characteristic that can vary is the number of introns and exons in the *GAPDH* gene even in the same species. Teich et al. (2007) reported that the locations of introns in *GAPC* and *GAPCP* are highly conserved between plants, indicating that the duplication event leading to the two genes occurred relatively late, probably with the emergence of terrestrial plants.



GAPC genes of *Arabidopsis*. The four GAPC genes of *Arabidopsis* have different numbers of exons and introns. *GAPC* has two fewer introns than *GAPC-2* even though the encoded amino acid sequence is highly similar.

Third, the length of the introns can also vary widely between species (Pérusse and Schoen 2004) as you may discover during your own investigation. Finally, the regions of the GAPDH protein that are not essential to carry out the enzymatic reaction may not be as well conserved due to reduced selective pressure, allowing divergent amino acids to evolve in the protein sequence.

APPENDIX C: SEARCHING AND SUBMITTING SEQUENCES TO THE GENBANK

As demonstrated in this project, the process of isolating a region of DNA and determining its exact nucleotide sequence is not an easy accomplishment. It is also not a trivial one. Even though the number of DNA sequences published in the GenBank doubles every 18 months (with 15 million new submissions in 2006 alone), each one of these submissions is invaluable (Benson et al. 2007). Each new sequence that is discovered tells us more about how nature works.

The nucleotide database, GenBank, is maintained by the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine housed on the campus of the National Institutes of Health (NIH) in Bethesda, MD, USA (Benson et al., 2007). DNA sequence information that is published by the NCBI GenBank is immediately shared with the DNA Database of Japan and the European Molecular Biology Laboratory (EMBL), which likewise reciprocate in sharing their new submissions. All of these databases can be accessed by the public, free of charge, via the Internet.

GAPC sequences from plants resulting from this project should be submitted for publication in the NCBI GenBank. This information could be useful to researchers interested in the evolution and functioning of the GAPDH enzyme, as well as the plant carrying the gene. Even *GAPC* sequences from a plant species that has already been published in the GenBank should be submitted again as confirmatory data. The sequence should be submitted especially if the *GAPC* gene is discovered to contain a polymorphism, a difference to the published gene sequence, because changes as minor as a single nucleotide base can have a dramatic affect on the functioning of the enzyme.

Note: Prior to GenBank submission, verify that the sequence and related annotation to be submitted are as good as they can be and that the sequence is not an *Arabidopsis* gene or another contaminating sequence. Perform a blastn search on the genomic sequence vs. the reference genomic database and the putative mRNA sequence vs. the reference mRNA database, and a blastp search on the putative amino acid sequence. From the genomic sequence verify that the intronic sequences are not identical to *Arabidopsis* or another plant that is different from that you have cloned. From the mRNA sequence, verify that the putative coding sequence aligns with known *GAPC* mRNAs in the database and do the same with the amino acid sequence.

Note: The NCBI frequently updates the GenBank. As such, the information presented here may differ. Use the resources on the NCBI website to help with tasks if the information presented here is unclear. Refer to General Submission Information available at ncbi.nlm.nih.gov/BankIt/index for more detailed and up-to-date information than presented here.

Searching the GenBank for Existing GAPDH Sequences

1. To examine the *GAPDH* sequences that have already been published in the GenBank, go to ncbi.nlm.nih.gov
2. Look for the Search bar (near the top of the page), and use the pull-down feature to select **Nucleotide** (instead of the default All Databases).
3. In the query box, enter the words *GAPC* Viridiplantae and click **Search**. In the results page will be hundreds of potential *GAPC* accessions from all types of plants. They are listed in chronological order, with the most recently submitted (or updated) accessions listed first.

4. To narrow the list down more, on the left side bar of the results page, under the section for **Molecule types**, click **genomic DNA/RNA**. This should eliminate the mRNA and cDNA sequences from the list. Some of the sequences will be partial, and there will be numerous cases where the same *GAPC* from the same plant species is listed more than once. These multiple listings are most likely the result of researchers looking at sequence variability within different populations or lineages of the same species, although they may also be different genes for isozymes.
5. To check whether a specific plant species has already been published in the GenBank, type the scientific name of the plant in the search field, along with the word **GAPC** and see what accessions come up. If in doubt of the scientific name for a plant, use the Taxonomy browser shown in the NCBI template along the top of the web page. Type in the common name and click **Search**.

Submitting a Sequence to the GenBank

Before beginning the submission process, be sure you have:

- Your annotated consensus sequence, in FASTA format (see Appendix G for instructions)
- The intron/exon boundary nucleotide numbers in your consensus sequence (as noted in section 4.2.2)
- The taxonomic name of the species your sample is from

Visit ncbi.nlm.nih.gov/WebSub/html/requirements.html for a detailed summary of the requirements for sequence submission to BankIt.

1. Go to ncbi.nlm.nih.gov/WebSub/ to access the BankIt submission web page.
2. In the upper right corner of the BankIt site, click on **Log in** and follow the prompts to create an NCBI account and/or log in.
3. Select **Sequence data not listed above...** and click start.
4. Click on the **Start BankIt Submission** button.
5. Continue following the guided steps on the BankIt submission web portal to submit your sequence.

Record your BankIt number: _____

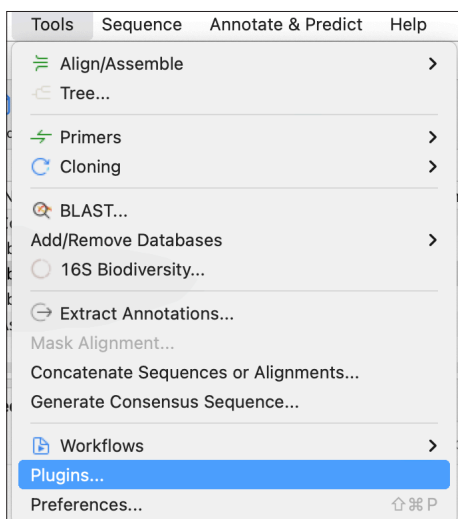
Once you have successfully entered all the necessary fields in BankIt, you will receive confirmation by email almost immediately. This confirmation includes the facsimile of your submission. Within two working days, you will receive another email with more details, including your official accession number.

Using the Geneious Prime Plugin to Submit Sequences to GenBank

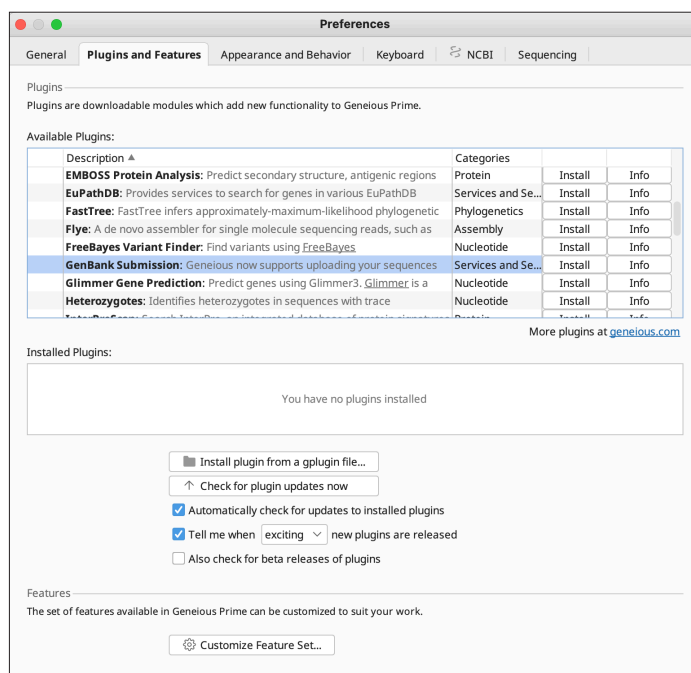
You can also submit your sequence to GenBank using the Geneious Prime GenBank Submission plugin. This plugin allows you to upload your sequences directly from the Geneious Prime program, retaining the annotations and features that you have added to appear on the GenBank record.

Installing the Geneious Prime plugin

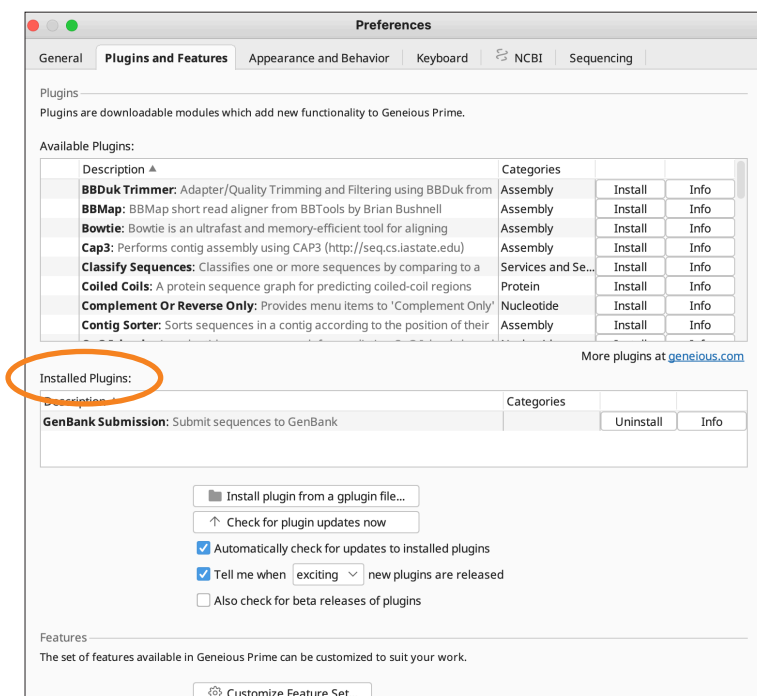
Use the plugin manager in Geneious Prime to install the GenBank plugin. Select Tools in the Geneious Prime toolbar, then select Plugins.



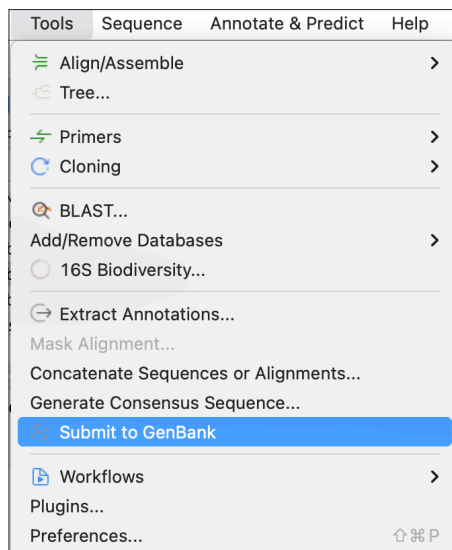
The new window that opens will automatically take you to the Plugins and Features tab. If the GenBank Submissions plugin is already installed, you will see it in the list of Installed Plugins. Otherwise, under Available Plugins, scroll down the list until you see GenBank Submission. Clicking on the Info button will give you a short description on the function of the plugin. To install the plugin, click Install.



A new window will appear to show you the download progress. You will be notified when your installation is complete, and the Preferences window will list the GenBank Submission plugin in the list of Installed Plugins.



Click **OK** to exit the window. To begin the submission process through Geneious Prime, select your annotated consensus sequence. Then click **Tools** in the toolbar. Look for an option called **Submit to GenBank**.



Submitting Sequences to GenBank Using the Geneious Prime Plugin

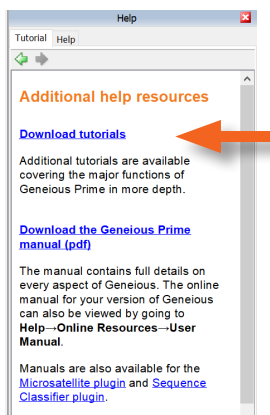
Access the GenBank Submission plugin by clicking Tools from the Geneious Prime toolbar, then Submit to GenBank.

A new window will open titled Submit to GenBank, with a note at the top informing you that you are about to create a submission of a single sequence.

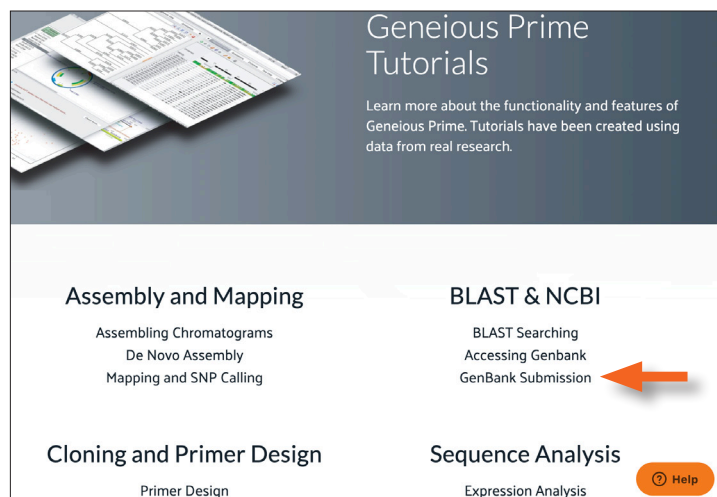
- For Submission Name, you can enter a description that will help you remember this sequence submission, such as the name of your plant. If you were able to identify the scientific name of your sequence, include that.
- For Molecule Type, select Genomic DNA from the drop down menu.
- If you were able to identify exons, select Include Features/Annotations.

For more detailed instructions on using this Plugin, use the Geneious Prime help tool to download the tutorial.

- Click the Help button in the main toolbar to open up the Help panel on the right side of your window.
- Click on the Tutorial tab in the Help panel and click on the final link for Additional help resources.

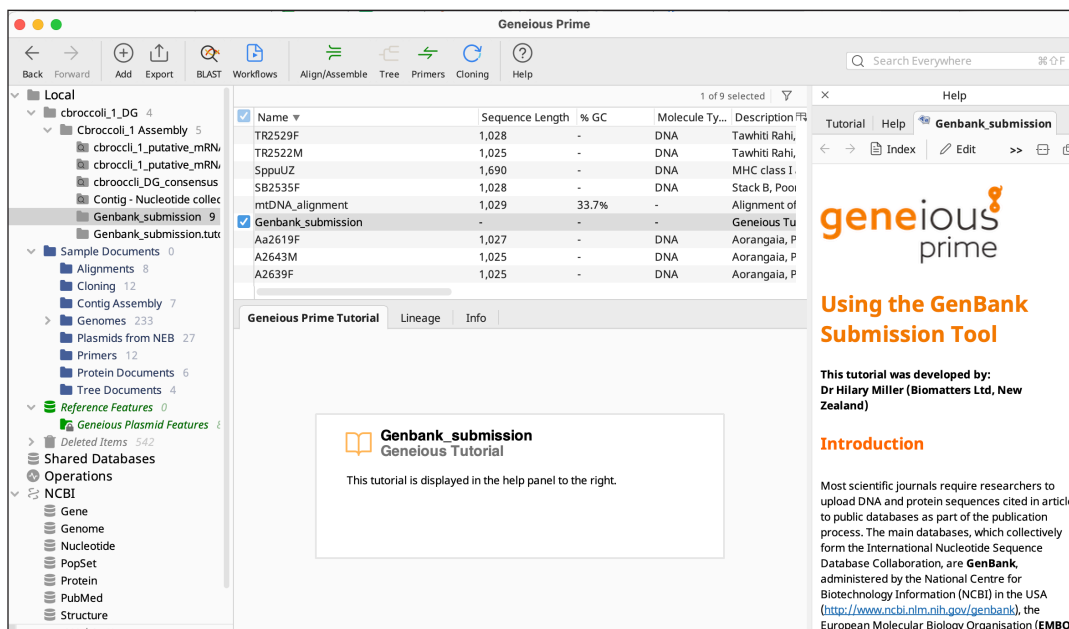


- On this new page, click on the first link for Download tutorials. This will take you to the Geneious Prime website for tutorial downloads. Find the one for Genbank Submission, in the Sequence Searching category. Save the .zip file on your computer and note down the file location.



- Go back to Geneious Prime and select your Local folder.
 - Drag and drop the zip file in the document table to import the files. OR
 - Use the File tab in the main toolbar, select Import, then select From File. Choose the zip file you just downloaded (typically named Genbank_submission.tutorial.zip).

A new folder will appear named Genbank_submission. The tutorial can now be accessed using the Help panel.



APPENDIX D: PREPARATION OF RESEARCH PAPERS AND PRESENTATIONS

Research Papers

Research papers are an important means of communication in science. The cliché that scientific writing is dull and formulaic is not true. The process of writing a scientific paper does allow for self-expression and individuality as long as the writer follows some basic rules. It pays to be aware of those rules from the beginning. Every scientific journal publishes a set of instructions to authors that apply to all of the articles it publishes. These instructions are provided by the publisher at least once a year and are usually printed at the back of the issue and appear on the journal's website. These instructions include information about the types of articles the journal publishes, their expected length, organization, and the journal's policies, including its preferred style for abbreviations, tables, figures, and cited references. Research papers can have numerous coauthors as long as the collaborators collectively agree on the order of authorship. Generally speaking, the name of the person who performed the majority of the work is listed first, while those who contributed less are included second, third, etc. The final name on the list of coauthors is often the instructor, advisor, or principal investigator on the project.

Most research papers follow the same basic outline: Title, Abstract, Introduction, Materials and Methods, Results, Discussion, and References. Some writers conceive of the title of their paper first, perhaps to inspire the rest of the manuscript, while others postpone that pleasure until after their writing is complete. All writing is a creative process, so things like titles can be amended at any time. Regardless, the title should be descriptive and succinct, and appropriate for the intended audience.

The Abstract is a summary, usually consisting of a single paragraph of 100–250 words. Most researchers don't try to write the abstract until a first draft of the rest of the manuscript is complete. This is so they know what information to include. The first sentence of the abstract should be a topic sentence, briefly describing the project. The next few sentences provide background information about the project as well as the overall research objective. The next few sentences can be devoted to briefly describing the methodology. The most important results should be summarized next with a brief conclusion. The final sentence of the abstract can summarize the project as a whole, perhaps hinting at implications or future plans.

The Introduction should contain information that is pertinent to all the different aspects of the project while staying close to the topic of the paper. In the case of this research project, there are four essential topics that should be discussed in the Introduction. First, there should be a description of the gene being isolated (*GAPDH*) and its biological function. Second, there should be information about the plant species whose DNA is being studied. Details might include the natural habitat of the plant, its geographical origin, unusual characteristics, and its usefulness to people. Third, there should be a short narrative about the techniques used to isolate the *GAPDH* gene. Since these techniques will be discussed in detail in the Materials and Methods section, they need be only summarized in the Introduction. Just let the reader know what basic strategy was used to isolate this *GAPDH* sequence. Fourth, finish the Introduction with a sentence or two stating the specific objectives for the study.

The next section of the research paper should be Materials and Methods. In this section, there should be a description of how the genomic DNA was extracted from the plant, how it was purified, amplified, ligated into the vector, and used to transform bacterial cells. The specific vector and its source should also be mentioned. There should also be a description of how the transformed cells were screened and selected, and how plasmid DNA was isolated from specific recombinant colonies and prepared for sequencing. Some things that should be included in this section are: the DNA sequence of the PCR primers used, the PCR conditions, and the procedures for plasmid purification, restriction enzyme digestion, and electrophoresis. If you are able to determine the

concentration of recombinant plasmid DNA in the experiment, that information can be provided here. A brief description of how the resulting DNA sequence was analyzed should also be included.

Be sure to discuss the experimental controls that were used in your experiment. Positive controls are used to verify that the experimental procedure will work in practice when a positive result is expected. This confirms that the reagents and conditions used were the right ones and helps in troubleshooting if a negative result is obtained with experimental samples. Negative controls are used to verify that when a negative result is expected, it in fact is observed, and therefore that positive results in the experiment are not due to an artifact. In PCR, artifacts are common and can arise from contamination of reagents and equipment with positive controls or other samples. More broadly, negative controls are used to ensure that the results are indeed the outcome of the conditions used in an experimental treatment.

The Results section of the research paper is a presentation of the outcome of the experiment with both figures and tables showing data and a thorough description of the results in the text. For instance, the DNA sequence, along with the protein sequence it codes for, can be presented as a figure. Each figure and table should have a short legend, and should be as informative as possible. Other figures can be included that show the results of BLAST searches and the phylogenetic relationship of that plant species with other plant species. The Results section should also include the accession number of the DNA sequence in the GenBank. Most scientific journals require that a researcher submit the DNA sequence to the GenBank or another sequence database before considering publication of a research paper for any previously uncharacterized gene. See Appendix C for information on submitting the sequence to the GenBank.

The writer should also include a title and a legend or caption for each of the tables and figures included in the report. The body text of the results section should help the reader by summarizing the important observations about the data. Set aside speculations or narratives about the implications of the data for the Discussion section of the paper. The Results section should be largely factual descriptions of observations.

The Discussion section of the paper is where the writer can speculate about the implications of the research and how the results coincide with or differ from the research of others. Technical problems that may have occurred during the experiment can be described in this section, as well as recommendations for future research. The original objectives of the research project, as outlined in the Introduction section of the paper, should be thoroughly addressed in the Discussion section. The final few sentences of the discussion should summarize the project as a whole.

At the end of the paper should be a References section, consisting of a complete list of references. List only sources that were actually used and cited in the paper. These may include articles from research journals, the popular press, or book chapters. The standards expected for sources are high, so the sources need to be reputable; the standard is generally a publication that underwent peer review or independent confirmation. Generally, information obtained from websites will not meet acceptable standards. Each of the references listed in the References section needs to be correctly cited, at least once, somewhere in the text. Some references might be cited numerous times in a paper, but usually not more than twice per paragraph. Each journal has a specific way to cite references in the body of the paper. For example, each reference can have a number assigned to it, then with that number used as the citation, or the last name of the authors can be cited (with year) as in this manual. Likewise, there are different ways that sources can be arranged in the References section. References are usually listed either alphabetically or chronologically (in the order that each reference was cited). Refer to the instructions provided by the particular journal of interest to determine which style should be used.

Once the Introduction, Materials and Methods, Results, and Discussion have been written, it is time to go back and write the Abstract for the paper. The Abstract should be a short recapitulation of each of these sections.

Oral Presentations

Another important way of relaying research to others is through oral communications. This might involve speaking to fellow students in a classroom setting, but could also be at a larger venue that others on campus are invited to, or even a formal conference involving researchers from other institutions. Most research talks are organized similarly to written reports, with a Title, Introduction, Materials and Methods, and Results and Discussion. Since audience members cannot go back and review the material (like they can with a written report), it is even more important to be succinct and to arrange the talk in a logical fashion.

There are other important differences between oral and written presentations. It is impractical to distribute copies of the written paper during the presentation, so instead speakers will usually project their presentation on a large screen using transparencies, 35 mm slides, or PowerPoint or similar presentation software. It is important to display the text in as large a font as reasonably possible so the audience can see it. Since the time frame for most oral reports is only 10–15 minutes, there isn't time for the audience to read a lot of text. Instead of showing long paragraphs, speakers will typically encapsulate the salient points in brief sentences or bulleted lists. Be concise and accurate, and give priority to the most important features of the research.

There is not the same distinction between Results and Discussion in an oral presentation as in a written report. When showing a table or graph, it is acceptable to speculate or to discuss implications right there, rather than waiting until later. Because the audience won't be able to return to presented data, it is also important to include a summary in the oral report. It is difficult for the audience to remember salient points, so it is worthwhile repeating them. The summary takes the place of an abstract, which is not normally given during a research talk. Only a minimal number of references should be provided in the talk, and might be integrated into the body of the talk, rather than listed at the end.

At the end of the talk, the speaker should ask the audience for questions or comments. This is a good opportunity to get perspectives from other people about the research and presentation, as well as to reiterate important features of the experiment.

Research Posters

Another popular means of research communication is with a scientific poster. This is partially because posters are a sort of hybrid between research papers and oral presentations. They allow the reader to examine the material thoroughly, as with a written paper, but since the researcher is also expected to stand by the poster for a certain period of time, it allows the reader to ask questions, as with an oral presentation. This provides the opportunity for both parties to have a more informal and in-depth discussion about the research than the short question and answer periods that follow oral presentations. For this reason, it is not uncommon to have more researchers presenting posters at state, regional, national, and international research conferences than are giving formal talks.

The outline of the research poster is identical to the written paper: Title, Abstract, Introduction, Materials and Methods, Results, Discussion, and References. There will not be room on the poster to include all of the narrative that is included in a paper, so only the more important points should be provided. The use of bulleted lists is encouraged.

Research posters can vary from 3' x 4' to 4' x 8' in size and can be displayed in several different ways depending on the situation. Individual pages can be printed on 8.5" x 11" sheets of paper or cardstock and mounted on large, flat sheets of white foamboard or trifold poster boards. These boards can usually be purchased at local art shops, teacher-supply stores, or campus bookstores. The sheets can be mounted with thumbtacks or attached with glue. An alternative,

and more attractive, substitute is to print the posters in their entirety on a single sheet of paper. Many commercial and campus printing shops can do this, although it is relatively expensive. If this approach is taken, the author would compose the poster ahead of time using software like Microsoft PowerPoint or Adobe Illustrator. In this scenario, instead of making the presentation a series of consecutive slides, all the material is placed on a single slide that is formatted to have very large dimensions. All of the research material is then copied and pasted into that single page and arranged in an effective manner. If this option is chosen, it is worth checking with printing shops on their software formatting requirements and necessary lead time.

The text of the poster needs to be legible from at least 6 feet away, so the font needs to be quite large and easy to read. Serif font styles like Bookman, Times New Roman, Garamond, Palatino, Helvetica, and Geneva are generally more legible from a distance. Generally, the title of the poster should have a font size of 80 point or more, headings should 36 point, and text font should be 18–24 point.

Organize the different sections of the poster so the reader's eyes will run up and down the poster in columns (how we read a newspaper) rather than left to right (how we read books). This is to allow readers to spend more time standing in one place while perusing the poster.

It is a good idea to draw a rough draft of the poster on a sheet of paper in order to envisage how it will look when complete. The general rule of thumb is to maximize the use of graphics and minimize the use of text. Perhaps the procedures described in the Materials and Methods could be replaced by a flow chart or a few tables. The Results can be dominated by tables and figures accompanied by well-written legends.

Feel free to include photos and other attractive visuals (like campus logos) in the poster, but try to remain faithful to a single stylistic theme involving just a few different colors. Be sure to leave plenty of blank space in the poster so the reader doesn't feel overwhelmed. Always check for spelling and grammatical errors, and make sure that the writing style is consistent from beginning to end. It is a good idea to show your poster to a fellow student, colleague, or instructor before printing it, as feedback from a third party can be quite helpful.

Resources for Research Articles

A valuable source of free, full-length research articles about a specific gene or a specific plant is PubMed Central. Just like the GenBank, this literature repository is maintained by the National Center for Biotechnology Information (NCBI) of the U.S. National Institutes of Health (NIH) and has an agreement with more than 330 life science journals to make available their research articles free of charge. All of these journals are peer-reviewed, and therefore considered to be of very high quality. To search for articles in PubMed Central, go to pubmedcentral.nih.gov and type in a search term(s). When conducting a search on PubMed Central, it is better to use the full name of the gene (glyceraldehyde 3-phosphate dehydrogenase) rather than the acronym (GAPDH), because the articles will focus more on that gene or enzyme. Searches for GAPDH will list articles that are primarily concerned with other, broader research topics, and that only mention the *GAPDH* gene or GAPDH enzyme. This is because researchers studying tissue-specific genes will often use *GAPDH* as an experimental control, since it is an essential gene expressed in all cells and organisms. When searching for literature about a specific plant, conduct two different searches. First, do a search using the common name of the plant (for example, corn), if available. This search will collect a broader range of research papers published about this plant, and will at least provide the scientific name of the plant. Second, do a search using the scientific binomial name of the plant (in this case, *Zea mays*). This will likely result in a different collection of papers, although there may be some overlap.

Two other valuable search engines exist for finding research articles: PubMed and Google Scholar. Although maintained by the same NIH agency, PubMed is a repository for more than 33,400 peer-reviewed scientific journals published throughout the world. To search for articles in PubMed, go to ncbi.nlm.nih.gov/sites/entrez?db=PubMed. To search for articles in Google Scholar, go to scholar.google.com. The primary disadvantage of PubMed and Google Scholar is that they do not provide the full-length articles that result from searches. PubMed will usually include the abstract, as well as the journal name, date of publication, and page numbers. Occasionally there will be a link to the article of interest. Google Scholar is less formal and has many different forms of articles. If the article looks interesting and is not available, it might be possible to find a free download from the journal at a local college or university library. Most campuses allow catalog searching on the Internet to save unnecessary trips to the library. Once it is confirmed that the library carries that journal, the writer can visit the library and photocopy the article. Libraries have agreements to exchange literature with each other so the writer might also be able to arrange for an interlibrary loan. Most colleges and universities offer interlibrary services free of charge to students and faculty at that school. Interlibrary loan forms are often available on the campus library website. Most articles that are retrieved from another library are transmitted electronically, so can be received within days. A much more expensive alternative to an interlibrary loan is to purchase the article directly from the journal. This can cost \$10–30, but the article is typically sent immediately via email. Another approach would be to contact the designated authors directly for a copy of the article. This last approach would be the best prospect, although this can be difficult because email addresses are not published by PubMed, and the authors might not even be available to answer your request. If this approach is taken, it is important to be polite and specific about which article is being requested. Many authors publish on the same topic numerous times every year, so you will need to specify exactly which article is being requested.

APPENDIX E: CUSTOM BLAST SEARCH SERVICES SET UP BY CUT AND PASTE

As mentioned in the Instructor's Advanced Preparation section, BLAST search services will need to be set up within Geneious Prime in order for students to carry out their bioinformatics steps. Because this setup requires internet for downloading the installation files, this step can be time-consuming if there are a lot of student computers, especially if your internet connection is slow.

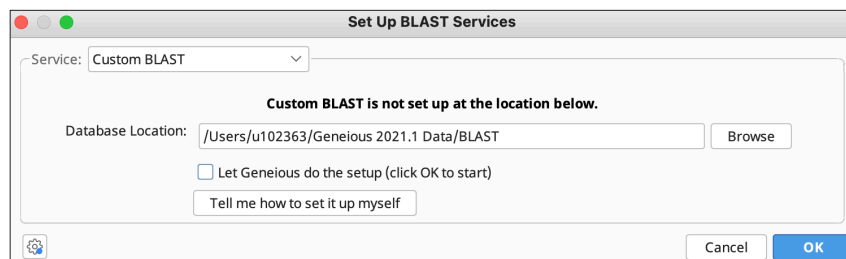
If you already have one computer that has undergone successful setup, you can copy the needed installation files onto a memory stick from this computer and transfer to all the other computers without requiring the internet. This method may be speedier if your internet connection is slow. If you do not have a computer with custom BLAST search service set up, you can still download the installation files from the internet and transfer the files with your memory stick.

If the custom BLAST search service has already been set up:

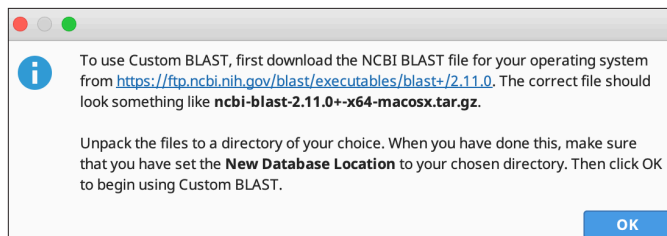
- 1) Navigate to the Geneious Prime Data folder on your computer that stores the installation files:
 - In your hard drive (C:), go to the Users folder, then your username folder (the example used here is dgong), then the Geneious Prime Data folder. The filepath should look something like this:
C:\Users\dgong\Geneious 2021.1 Data\
- 3) The files for the Custom BLAST search service reside in the BLAST folder. Copy the contents of the BLAST folder onto a memory stick.
- 4) On the other student computers, navigate to the Geneious Prime Data folders. Replace their existing BLAST folders (or if they don't have one, this will be their BLAST folder) with the one you have on your memory stick.

If you DO NOT have a computer that already has the custom BLAST search service properly set up:

- 1) In the Geneious Prime program, in the main toolbar, go to **Tools**, then **Add/Remove Databases**, then **Set Up BLAST Services**.
 - a) Select **Custom BLAST for Service**. A default database location will appear in the text field, which is where your original installation files went, and should be something like this: C:\Users\dgong\Geneious 2021 Data\BLAST
 - b) Now, you want to create a different folder where the correct files should live. Append this folder name with something to distinguish it from the original. For example, add a 2 to the end: C:\Users\dgong\Geneious 2021 Data\BLAST2
 - c) Uncheck the box that says **Let Geneious Prime do the setup** and click **Tell me how to do it myself**:



- d) You'll then see a window with brief instructions. The ftp link is the weblink you will click in order to get the appropriate files from NCBI. In this window, take note of the filename listed in bold. This will be the file you will look for in the list of downloads:

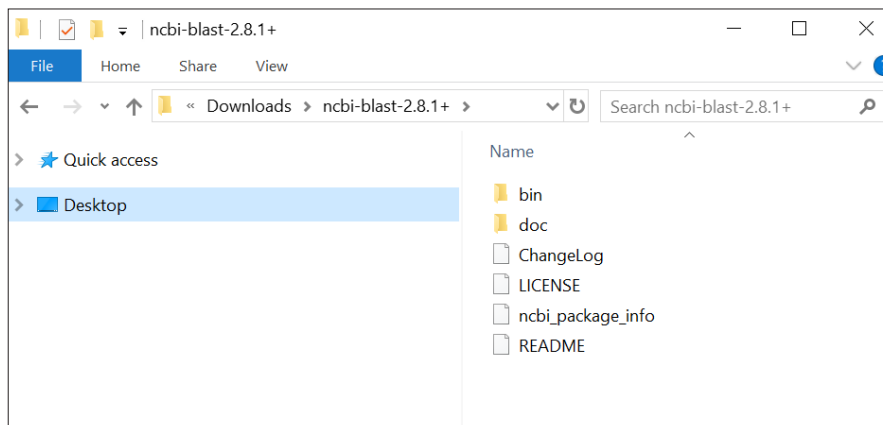


When you click the blue link (or copy and paste into your web browser), it should take you to a page with even more links for downloads. Find the file that was listed in bold from the previous window.

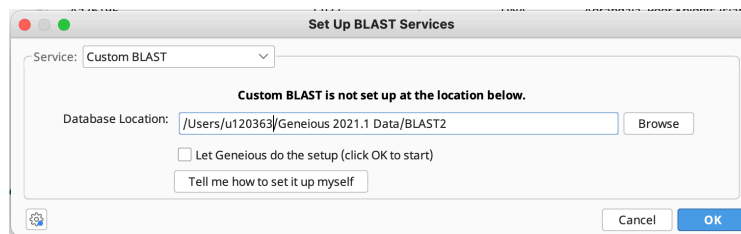
2)

Name	Last modified	Size
Parent Directory		-
ChangeLog	2020-11-03 09:49	85
ncbi-blast-2.11.0+-1.src.rpm	2020-11-03 09:49	51M
ncbi-blast-2.11.0+-1.src.rpm.md5	2020-11-03 09:49	63
ncbi-blast-2.11.0+-1.x86_64.rpm	2020-11-03 09:49	180M
ncbi-blast-2.11.0+-1.x86_64.rpm.md5	2020-11-03 09:49	66
ncbi-blast-2.11.0+-src.tar.gz	2020-11-03 09:49	56M
ncbi-blast-2.11.0+-src.tar.gz.md5	2020-11-03 09:49	64
ncbi-blast-2.11.0+-src.zip	2020-11-03 09:49	60M
ncbi-blast-2.11.0+-src.zip.md5	2020-11-03 09:49	61
ncbi-blast-2.11.0+-win64.exe	2020-11-03 09:49	89M
ncbi-blast-2.11.0+-win64.exe.md5	2020-11-03 09:49	63
ncbi-blast-2.11.0+-x64-linux.tar.gz	2020-11-03 09:49	229M
ncbi-blast-2.11.0+-x64-linux.tar.gz.md5	2020-11-03 09:49	70
ncbi-blast-2.11.0+-x64-macosx.tar.gz	2020-11-03 09:49	139M
ncbi-blast-2.11.0+-x64-macosx.tar.gz.md5	2020-11-03 09:49	71
ncbi-blast-2.11.0+-x64-win64.tar.gz	2020-11-03 09:49	89M
ncbi-blast-2.11.0+-x64-win64.tar.gz.md5	2020-11-03 09:49	70
ncbi-blast-2.11.0+.dmg	2020-11-03 09:49	142M
ncbi-blast-2.11.0+.dmg.md5	2020-11-03 09:49	57

- a) When you click this link, the zipped files will be downloaded to your computer (look in whichever default folder you use for internet downloads). The folder should be called something like "ncbi-blast-2.11.0"...
- 3) Once you unzip these files, your folder contents should look something like this:






- 4) Now, navigate back to your Geneious Prime 2021 Data folder on your computer:
C:\Users\dgong\Geneious 2021 Data\
 - a) Drag the unzipped ncbi-blast-2.11.0 ... folder into this location on your computer, and rename **BLAST2**.
- 5) Go back to the window in Geneious Prime, and click **OK** to exit out of all windows. If you go back to Tools, then Add/Remove Databases, then **Set Up Blast Services**, the same window will appear and tell you that Custom BLAST is set up at the new location (BLAST2):



APPENDIX F: TIPS AND TRICKS TO NAVIGATE THE GENEIOUS PRIME INTERFACE

The Geneious Prime program allows you to perform your bioinformatics workflow using a simple desktop interface. Chapter 9 of this Cloning and Sequencing Explorer Series Instruction Manual includes a tour of the Geneious Prime platform to help familiarize you with the various features of the program that are essential for analyzing the sequences you cloned from your plant samples. Here are some extra features that may be helpful in getting the most out of your experience with Geneious Prime.


Split View

If you want to view your document in more than one way at the same time, you can use the Split View function  to view two different tabs for the same document at the same time. When the view is split, the regions of sequence and the annotations selected are synchronized between the viewers. You can access Split View by clicking **View** on the Geneious Prime toolbar, then selecting **Split Viewer Left/Right**, which lays out the views side by side (see figure below), or **Split Viewer Top/Bottom**, which lays out the views with one on top of the other. You can also use the Split View icon  in the upper right corner of the document view panel. To close Split Views, click on the X  for one of the windows.



Split Viewer Left/Right. In this Geneious Prime window, the view of the document has been split so different elements can be viewed side by side.

Expanded View

If you need more viewing space for your document, the document view panel can be expanded to fill up the main Geneious Prime window. Expanded View hides the sources panel on the right and the document table above. You can access Expanded View by clicking **View** on the Geneious Prime toolbar, then selecting **Expand Document View**. You can also use the Expanded View icon  in the upper right of the document viewer. Clicking this icon again will return the layout to its original state.



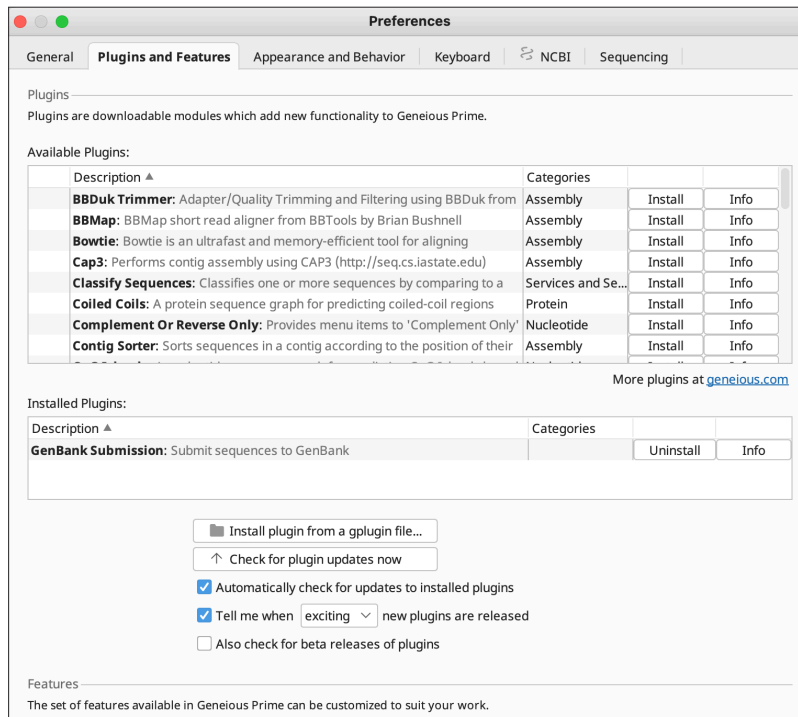
Expanded View. In Expanded View, the sources panel on the right and the document table above can be hidden to maximize the viewing space for your document. Expanded View can be toggled using the icon on the upper right of the document view window.

New Window

Another way to increase viewing space for your data is to open a new window. This allows you to have several documents open at once. You can open a new window by clicking **View** on the Geneious Prime toolbar, then selecting **Open Document in New Window**. You can also open a new window by double clicking on your document in the document table. Note that there will not be a menu bar at the top of the new window.

Plugins

Geneious Prime contains many plugins, written by the company and the Geneious Prime user community to extend its functionality. You can navigate to the Geneious Prime plugins interface by selecting **Tool** from the Geneious Prime toolbar, and then choosing **Plugins**. This will take you the Plugins and Features tab, which lists the plugins currently available for download from the Geneious Prime website (that are not already installed). Each plugin is listed with a status, which can be a star (for exciting plugins), New, or Beta. Click the Info button to read more about the plugin, or click **Install** to download and install it. Once you've installed a plugin it will be listed within the Installed plugin window. Click the uninstall button next to a plugin to remove it.



Plugins available for download from Geneious Prime. There are many available plugins for download that add extra functions for Geneious Prime. You can access the list of available plugins by clicking **Tools** in the Geneious Prime toolbar, then clicking **Plugins**.

APPENDIX G: BLAST SEARCHING ON THE NCBI WEBSITE

In the bioinformatics workflow in Chapter 9, the trimmed sequences from the cloned plant samples are compared to other sequences in the GenBank database using a program called BLAST. Geneious Prime allows you to run BLAST searches directly from its interface to help streamline your bioinformatics workflow.

In general, the amount of time it takes to retrieve BLAST results will vary depending on how many searches NCBI BLAST is asked to run at any particular moment from researchers around the world. In some cases, searches performed through Geneious Prime are not as fast as performing the BLAST searches directly through NCBI. If you have short class periods (50 minutes or less) and need the BLAST results as soon as possible to enable data analysis, you have the option of performing the BLAST searches directly from NCBI's website.

Formatting Your Sequences for BLAST Searching

To run your BLAST searches, GenBank requires that sequences be either pasted or uploaded as a file in FASTA format. FASTA is a text-based format for representing sequences such that nucleotides (or amino acids) are denoted using single-letter codes. Since your instructor has enabled the Fasta View plugin in your Geneious Prime software (step 6 from the Instructor's Advance Preparation section), you can simply copy and paste from this view into the search box on the BLAST webpage. If your version of Geneious Prime does not have Fasta Nucleotide View enabled, you can also manually export your sequence files for upload onto the BLAST webpage.

Copy and paste your sequence from Fasta Nucleotide View

In the document table, click to select your file. In the document viewer, look for the tab named Fasta Nucleotide View.

[illegible]

In Fasta Nucleotide View, your sequence is displayed in FASTA format. The first line is a single line description of your sequence that begins with a greater-than (>) symbol. The text following the symbol is called the identifier of the sequence, and the rest of the text is the description. There should be no spaces between the > symbol and the first letter of the identifier. The lines beneath the single-line description is your sequence.

You can also view multiple sequences in Fasta Nucleotide View to enable a batch blastn search on the NCBI website. Click to select multiple sequences from the document table. Using the Fasta Nucleotide View tab in document view, you will see all the FASTA-formatted sequences on the page. Each > symbol marks the start of a new sequence, so GenBank will be able to separate your sequences for you when you copy them all at once for the BLAST search.

4 of 4 selected

<input checked="" type="checkbox"/>	Name ▼	Sequence Length	% LQ	% HQ	Post-Trim Le...	% GC
<input checked="" type="checkbox"/>	Cbroccoli_1_pJETSEQ_Reverse.ab1	1,149	3.5%	79.6%	952	44.0%
<input checked="" type="checkbox"/>	Cbroccoli_1_pJETSEQ_Forward.ab1	1,086	6.4%	66.4%	844	43.8%
<input checked="" type="checkbox"/>	Cbroccoli_1_GAPSEQ_Reverse.ab1	1,136	0.7%	87.5%	562	38.9%
<input checked="" type="checkbox"/>	Cbroccoli_1_GAPSEQ_Forward.ab1	1,135	0.6%	92.1%	481	41.1%

Sequence View

Annotations

Virtual Gel

Text View

Fasta Nucleotide View

Info

>Cbroccoli_1_pJETSEQ_Reverse.ab1

NNNNNNNNNNNNNNNNNNANCTTCTANNAAGNNCTACTGGTGTCTTCACTGACAAGGACAAAGCTGCTGCTCACTTGAA
GGTTTGTGTTTTATATTGTTAGTCTTGGATGATTCTTCAGTAATTAGTAGACTTGTGATGCTTCAACTGATTGATTGGTG
AATTTGTTTGTTCAGGGTGGTGCTAAGAAGGTTGTCTCTGCTCCGAGCAAGGACGCTCCCATGTTGTTGTTGGTGT
TAACGAGCACGAGTACAAGTCCGACCTTGACATTGTCTCCAACGCTAGTTGCACCACTAACTGCCTTGCTCCCTTGCCA
AGGTAAGACCTCATGCGCTGTTTATAATTGAATTTGATAGTGTATGCTGAACCTCTTCCAATTTGGGTTGCTTCT
TCCTTTAGGTATCAACGACAGGTTGGAATTTGTTGAGGGTCTCATGACTACCGTCCACTCTATCACTGGTAAATTTCTC
AGTCTTCTAAAAATGTTAAACCGAACTGTTGCTATATTAGTTAACTCTGATCTAATGGGTTCTGCTTTTATGGTATCA
GCTACTCAGAGAGACTGTTGATGGACCATCAATGAAGGACTGGAGAGGTGGGAGAGCTGCTTCATTCAACATCATTCCAG
CAGCACTGGAGCTGCAAGGCTGTGCGAAAGGTGCTCCACAGCTCAATGGAAGTTGACAGGAATGTCCTTCCGTGTTCC
CCACGTTGATGTTCTCAGTTGTTGACCTCACGGTTAGACTCGAGAAAGCTGCTACCTACGACCATCAAGAAGGCTATC
AAGTAAGCTTTCGGTTCAGTTAACTAGTTTGATCAAACTTTTGAAGATTAACTAACTGATTGGATTGTTACACAGG
GAGGAATCTGAGGGCAAGCTAAAGGGAATCCTTGGTTACACCGAGGATGATGTTGTCTCGACTGACTTCGTTGGTGACAA
CAGGTCAAGCATCTTTGATGCCAAGGCNNNTCTTGCTGNAAAACTCGAGCCNNNCCNNNNNNNGGNNGCNNNTCTNCCT
NNAGNNNNNNCTATTACCCGNNNNGNATTGGGNTTNNNGCANNNNNGTTAANGNNNTTGNTTTNNNTCCNNNNNNNNNA
AAAAANNNNNAAGGNAACNNNNNNNCAGTT

>Cbroccoli_1_pJETSEQ_Forward.ab1

NNNNNNNNNNNNNNNNNNNGGNTCNGTTNNNNNGAAGATCAGCCTNTGGCATCAAAGATGCTTGACCTGTTGTCACCAAC
GAAGTCAGTCGAGACAACATCATCTCGGTGTAACCAAGGATTCCTTTAGCTTGCCCTCAGATTCCTCCCTGTGTAACA
ATCCAATCAGTTAGTTAAATCTTACAAAGATTTGATCAAACTAGTTAACTGGAACCGAAAGCTTACTTGATAGCCTTCT
TGATCTGGTGGTAGGTAGCAGCTTTCTCGAGTCTAACCGTGAGGTCAACACTGAGACATCAACGGTGGGAACACGGAAG
GACATTCCTGTCAACTTTCCATTGAGCTGTGGAAGCACCTTCCGACAGCCTTGGCAGCTCCAGTGCTGCTGGGAATGAT
GTTGAATGAAGCAGCTCTCCACCTCTCCAGTCCTTCATTGATGGTCCATCAACAGTCTTCTGAGTAGCTGATACCATAA
AAGCAGAACCCATTAGATCAGAGTTAACTAATATAGGCAAGCACTTCCGTTTAACTTTTGAAGACTGAGAAATTTAG

Sequence from first document

Sequence from the second document...

When you copy your sequence to paste into the search box on the BLAST website, highlight your sequence from Fasta Nucleotide View. Click Control-C to copy, or go to the Edit tab in the Geneious Prime toolbar and select **Copy**. Then, navigate to the nucleotide blast page (see below) to paste your sequences.

Manually export sequences files in FASTA format

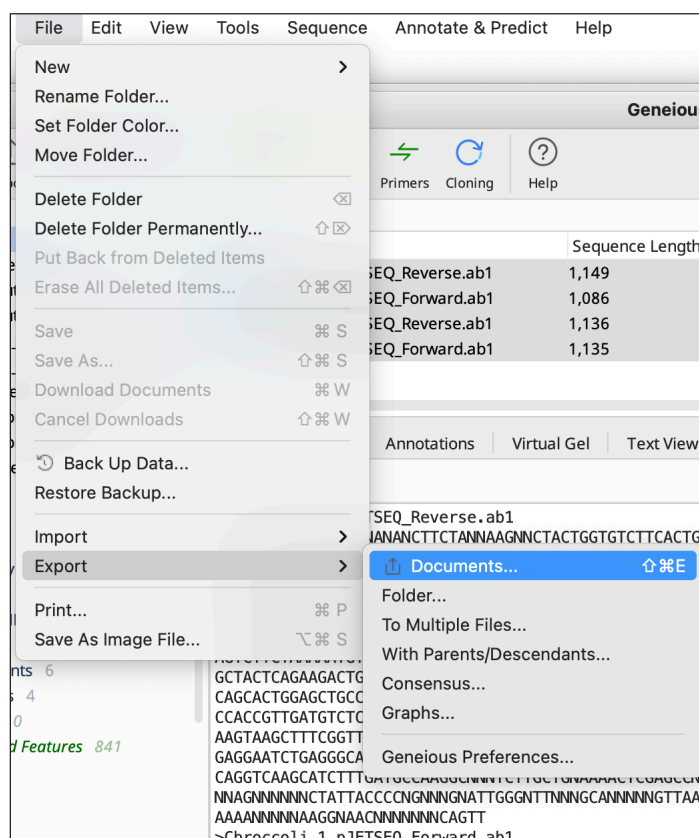
If you prefer to generate sequence files in FASTA format for BLAST searching or other purposes, Geneious Prime provides an easy way to do this using the Export function.

There are several ways you can export your data.

- Export one file at a time
- Batch export—export multiple individual files at the same time
- Export multiple files into a single document

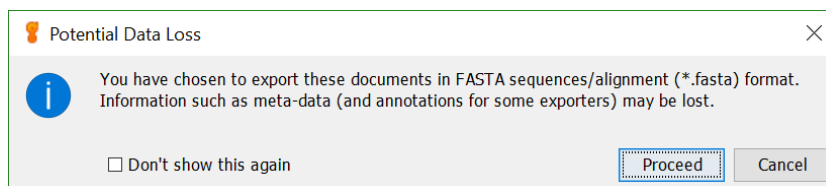
1. Export one file at a time (or export all four files together as one document):

Click to select your file(s) in the document table. In the Geneious Prime toolbar, click **File**, then **Export**, then **Selected Documents**.

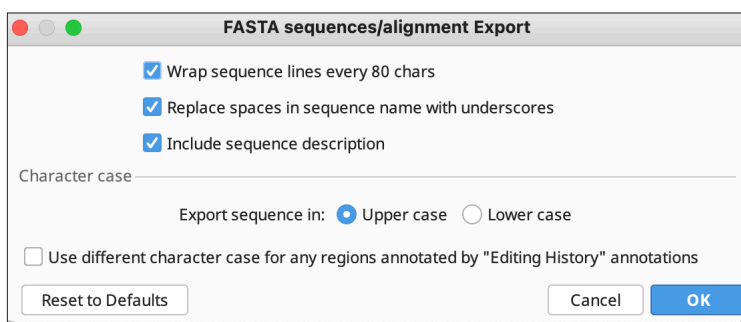


In the new window, select the location where you would like to store your file. In the example below, the file will be stored in a new folder named “cbroccoli export” located on the computer desktop. For Files of Type, select FASTA sequences/alignment (*.fasta) from the dropdown menu. If you are exporting a single file, the file name will be the name of the sequence itself. If you are exporting all four sequences into one document, the file will be automatically have the prefix “4 documents from” followed by the folder name that contains your documents.

If you are exporting four sequences into one file, you may see another window appear notifying you that there may be potential data loss. This is because FASTA format saves only the sequence itself rather than any metadata that Geneious Prime adds onto your data, such as annotations. Click **Proceed**.

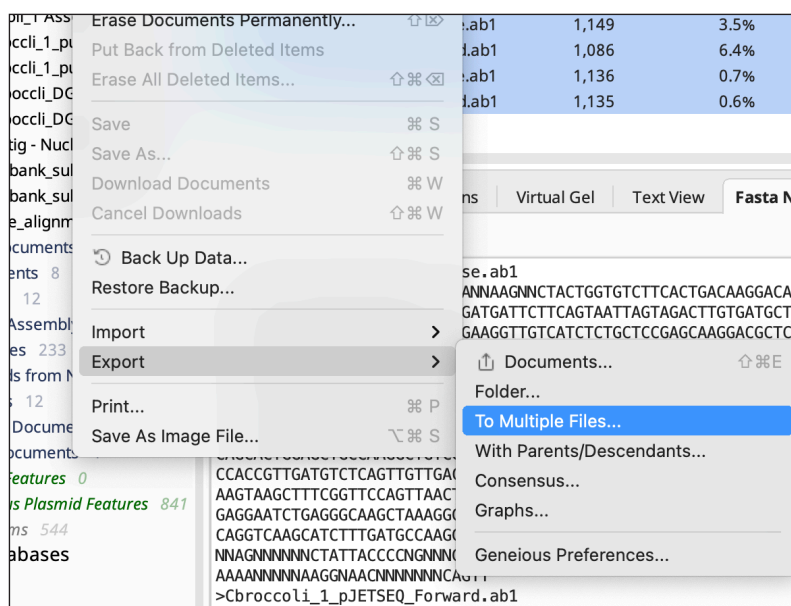


Another window will appear, this time offering you options on how the FASTA file will be organized. Keep the first three boxes checked and click **OK**.

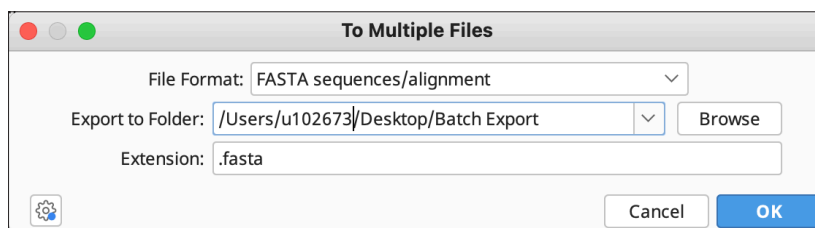


2. Export multiple individual files at the same time.

Click to select your files in the document table. In the Geneious Prime toolbar, click **File**, then **Export**, then **To Multiple Files**.

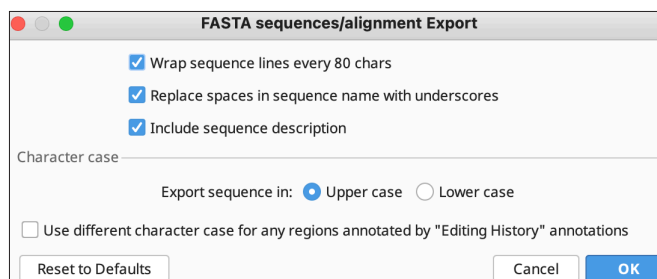


A new window will appear to set your save options. Select **FASTA** from the File Format dropdown menu. Extension should automatically populate with **.fasta**. Select where you'd like to export your file. Click **OK**.



For the FASTA export window, check the first three boxes and click **OK**.

Your file will now be saved in the location you chose.

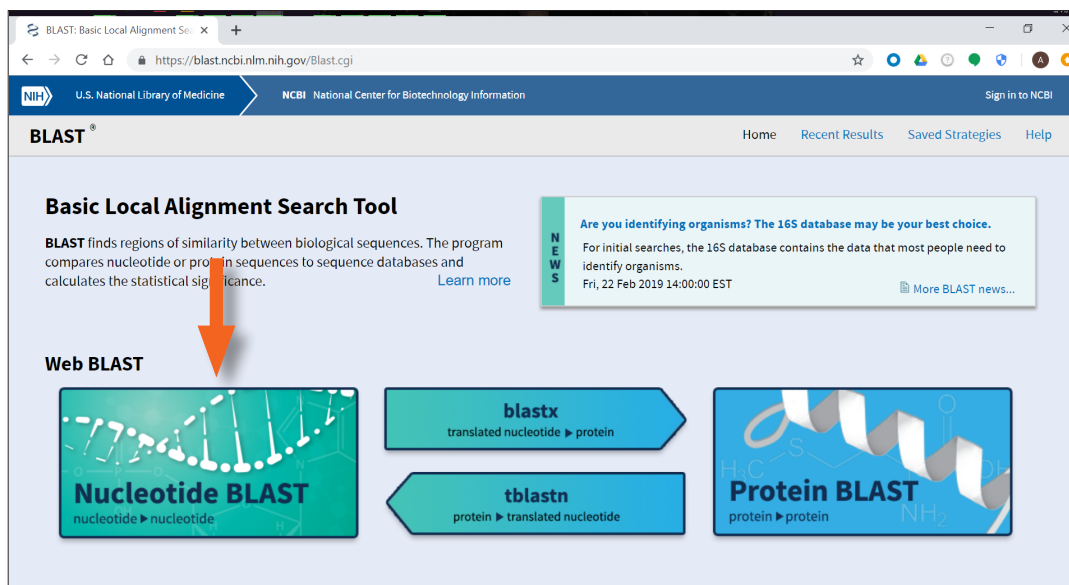


Your FASTA formatted file is now saved in the location you specified.

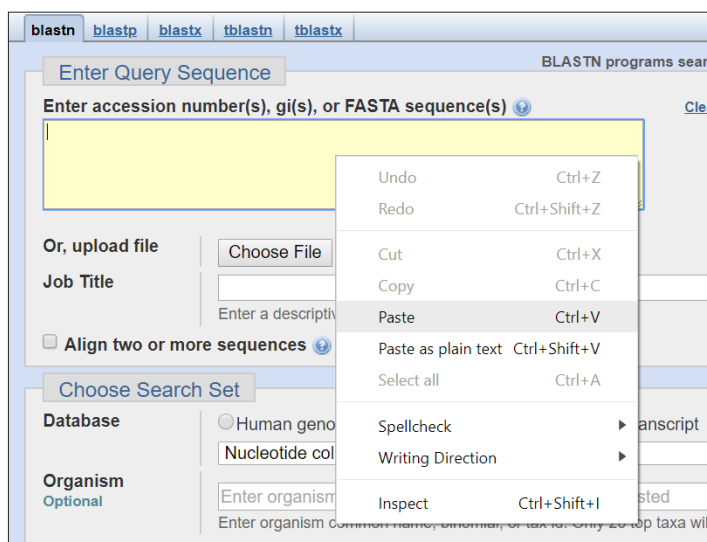
BLAST Searching on the NCBI BLAST Webpage

Now that you have copied your sequences or exported your sequence files in FASTA format, you are ready to run your BLAST search on NCBI's BLAST webpage.

Open the NCBI BLAST homepage blast.ncbi.nlm.nih.gov/Blast.cgi. Click on the large icon for Nucleotide BLAST.



On the next page, enter your query sequence(s) in the top section titled Enter Query Sequence. If you copied the sequence from Fasta Nucleotide View from Geneious Prime, click on the search box at the upper right. The window will turn yellow when selected. Right click on the box and select **Paste**.



If you exported your sequences as FASTA files, click the **Choose File** button and select your exported file for upload and click **OK**. The name of your file will appear beside the Choose File icon.

If you pasted in your sequence, click the expand icon in the bottom right corner of the search box to make sure your entire sequence is pasted in. Under **Choose Search Set**,

- Since you have plant sequences, select the radio button for **Others** (nr etc.) as your database
- In the dropdown menu, select the specific database listed in the directions for the section of Chapter 9 in your bioinformatics workflow. This could be either **Reference Genomic Sequences (refseq_genomic)** or **Nucleotide Collection (nr/nt)**
- For Organism, enter **Plant** and a dropdown list will appear. You can select the category that best fits your sample.

Nucleotide BLAST: Search nucleotides

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s)

>cbroccoli_DG_Consensus
AGCCTTGGCATCAAAGATGCTTGACCTGTGTGACCAACGAAGTCAGTCGAGACAACATCATCTCGGTGTAA
CCAAAGGA
TTCCCTTTAGCTTGCCCTCAGATTCTCCCTGTGTAACAATCCAATCAGTTAGTTAAATCTTACAAAAGATTT
GATCAAA

Clear

Query subrange

From

To

Or, upload file

Choose File 4 documents...1_DG.fasta

Job Title

cbroccoli_DG_Consensus

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

Reference genomic sequences (refseq_genomic)

Organism

Optional

plant

green plants (taxid:33090)

Plant rhabdovirus subgroup A (taxid:11305)

higher plants (taxid:3193)

land plants (taxid:3193)

plants (taxid:3193)

vascular plants (taxid:58023)

Plantactinospora (taxid:673534)

Plantactinospora Qin et al. 2009 (taxid:673534)

Plantibacter (taxid:190323)

Exclude

Optional

Limit to

Optional

Entrez Query

Optional

Program Selection

Optimize for

exclude

shown

Create custom database

Cloning and Sequencing Explorer Series

- Under Program Selection, select **Somewhat similar sequences (blastn)**. Beneath the blue BLAST icon, click **Algorithm parameters** to open more parameter options.
 - Select **50** for Max target sequences
 - Select **15** for Word size
 - Keep the remaining selections with their default values
 - Click the blue BLAST icon at the bottom of the page to begin your blastn search

Nucleotide BLAST: Search nucleotide sequences

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

BLAST Search database Reference genomic sequences (refseq_genomic) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences ♦ 50 Select the maximum number of aligned sequences to display

Short queries ☒ Automatically adjust parameters for short input sequences

Expect threshold 10

Word size ♦ 15

Max matches in a query range 0

Scoring Parameters

Match/Mismatch Scores 2,-3

Gap Costs Existence: 5 Extension: 2

Filters and Masking

Filter ☒ Low complexity regions ☐ Species-specific repeats for: Homo sapiens (Human)

Mask ☒ Mask for lookup table only ☐ Mask lower case letters

BLAST Search database Reference genomic sequences (refseq_genomic) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

A new page will open to display the progress of your blastn search.

BLAST » blastn suite » RID-8DKH4AB4014 [Home](#) [Rec](#)

Format Request Status

[Formatting options]

Job Title: cbroccoli_DG_Consensus

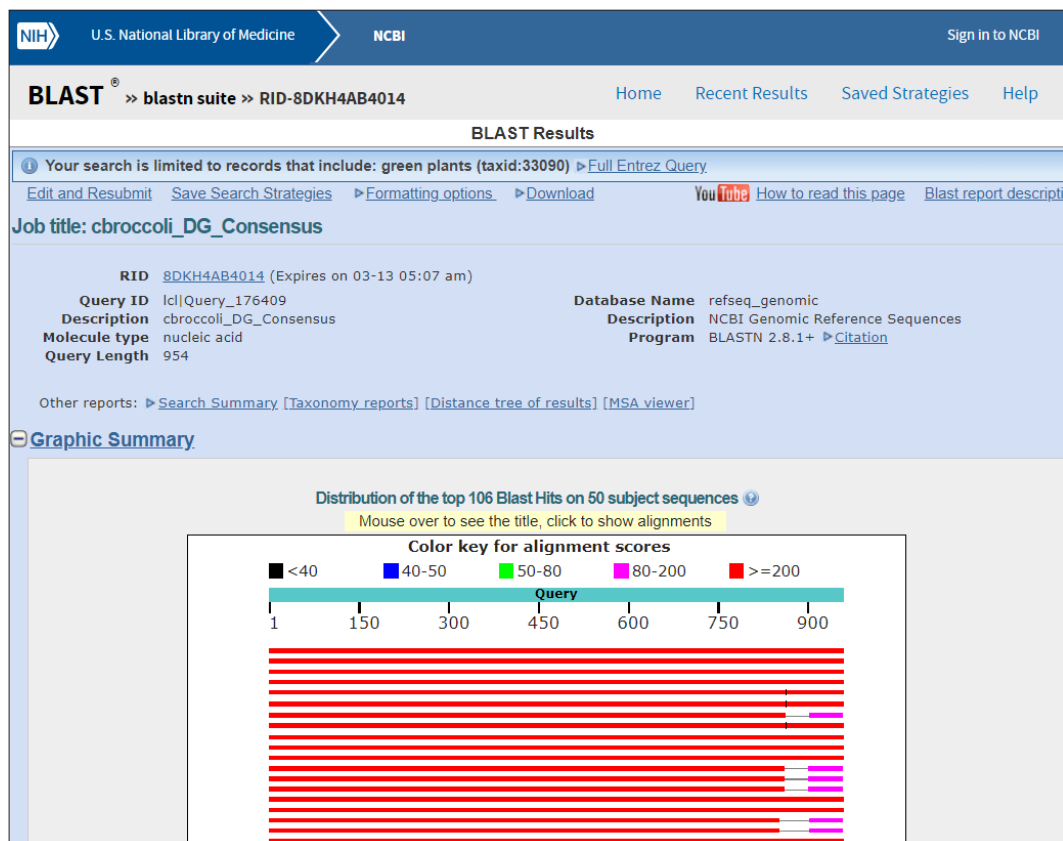
Request ID	8DKH4AB4014
Status	Searching
Submitted at	Mon Mar 11 17:07:48 2019
Current time	Mon Mar 11 17:07:56 2019
Time since submission	00:00:07

This page will be automatically updated in 2 seconds

BLAST is a registered trademark of the National Library of Medicine

NCBI
National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

When your results appear, the page will be divided into a few main sections. At the top, the Graphic Summary shows different regions and degrees of similarity between your query sequence and the sequences found in the genomic database. This is similar to the Geneious Prime Query Centric View. If the resulting sequences do not continuously match your query sequence, there will be a gray line that connects those regions.



Next on the blastn results page is a section titled Descriptions containing a list that summarizes the statistics. Each row contains a matching sequence with the best-matching sequence at the top of the table. This is similar to the Geneious Prime Hit Table. From this table you can get a feeling for whether your cloned gene was in the ballpark of a *GAPDH* gene in plant organisms.

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) [Selected 0](#)

Alignments	Description	Max score	Total score	Query cover	E value	Ident	Accession
Download	GenBank	Graphics	Distance tree of results				
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C3 genomic scaffold_BOL_C3 whole genome shotgun sequence	1532	1532	100%	0.0	95.14%	NW_013617408.1
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C3_BOL whole genome shotgun sequence	1532	1532	100%	0.0	95.14%	NC_027750.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 chromosome C5 genomic scaffold_Bra_napus_v2.0 C05 whole genome shotgun sequence	1482	2423	100%	0.0	93.92%	NW_019168549.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 chromosome A3 genomic scaffold_Bra_napus_v2.0 A03 whole genome shotgun sequence	1174	1174	100%	0.0	86.23%	NW_019168537.1
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C5 genomic scaffold_BOL_C5 whole genome shotgun sequence	1041	2116	100%	0.0	87.13%	NW_013617410.1
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C5_BOL whole genome shotgun sequence	1041	2116	100%	0.0	87.13%	NC_027752.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 unplaced genomic scaffold_Bra_napus_v2.0 scaffold2977 whole genome shotgun sequence	1028	1112	95%	0.0	87.00%	NW_019169570.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 chromosome A6 genomic scaffold_Bra_napus_v2.0 A06 whole genome shotgun sequence	1023	2021	100%	0.0	86.89%	NW_019168540.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 unplaced genomic scaffold_Bra_napus_v2.0 scaffold42243 whole genome shotgun sequence	1012	1012	100%	0.0	82.16%	NW_019169015.1
<input type="checkbox"/>	Brassica rapa cultivar Chifu-401-42 unplaced genomic scaffold_Brapa_1.0 Scaffold000191 whole genome shotgun sequence	1003	1003	100%	0.0	81.96%	NW_008732848.1
<input type="checkbox"/>	Raphanus sativus cultivar WK10039 unplaced genomic scaffold_Rs1.0 whole genome shotgun sequence	959	959	100%	0.0	80.52%	NW_017353143.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 unplaced genomic scaffold_Bra_napus_v2.0 scaffold1056 whole genome shotgun sequence	953	1048	95%	0.0	84.29%	NW_019168631.1
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C1 genomic scaffold_BOL_C1 whole genome shotgun sequence	949	1044	95%	0.0	85.18%	NW_013617406.1
<input type="checkbox"/>	Brassica oleracea var. oleracea cultivar TO1000 chromosome C1_BOL whole genome shotgun sequence	949	1044	95%	0.0	85.18%	NC_027748.1
<input type="checkbox"/>	Brassica napus cultivar ZS11 unplaced genomic scaffold_Bra_napus_v2.0 scaffold1670 whole genome shotgun sequence	944	944	100%	0.0	81.04%	NW_019168768.1
<input type="checkbox"/>	Raphanus sativus cultivar WK10039 unplaced genomic scaffold_Rs1.0 whole genome shotgun sequence	942	1777	100%	0.0	80.78%	NW_017353144.1
<input type="checkbox"/>	Brassica rapa cultivar Chifu-401-42 chromosome A1 genomic scaffold_Brapa_1.0 whole genome shotgun sequence	938	1031	94%	0.0	84.67%	NW_008711244.1

Additionally, in the Alignments section beneath the Description section, you will see how each query result aligns to your subject sequence on a nucleotide level. This is similar to the Geneious Prime Text View when a single blast result is compared with your query sequence.

If you submitted multiple sequences for the blastn search, you will be able to see the results for each sequence by selecting from the “Results for” dropdown menu at the top of the page.

Appendices

APPENDIX H: DETERMINE SEQUENCE IDENTITIES OF INDIVIDUAL SEQUENCES USING BLAST

After you complete Section 1 (View Sequence Traces and Review the Quality of the Sequencing Data), this is an optional step to compare all your trimmed sequences to the sequences in a selected database. This is called a basic local alignment search tool (BLAST) search. BLAST will compare each sequence to the selected database. The objective in this step is to become familiar with BLAST and how to assess alignments, and to make a preliminary determination of which plant *GAPDH* genes most closely resemble the gene that you cloned.

Biological sequences have evolved over time from common ancestors. Comparing a sequence with other known sequences, using an inexact alignment method to find potential relatives, will help you identify the function of an unknown or new sequence. BLAST programs are designed to find short (local) regions where pairs of sequences match. A *blastn* search compares a query sequence in turn to each sequence in a nucleotide sequence database. The result of a *blastn* search will be a set of matching and potentially related sequences ranked according to similarity.

Here, *blastn* will be used to compare your .ab1 chromatogram sequences to the GenBank database of all genomic nucleotide sequences. Once the search is complete, *blastn* counts all the nucleotides in the matching regions and awards two points for every pair of bases that match. If one sequence has an insertion, a deletion, or a gap (more than one base missing) and the other does not, BLAST deducts points from this score. The net result is that a *blastn* score is more or less twice the length of the matching region, depending on how many points were deducted.

The completed search will return a *blastn* score and an E-value for each match of your query sequence to a sequence in the GenBank database. The results also include an alignment of your sequence to each match in the database so that you can compare them. The meaning of the *blastn* scoring will be explained in more detail in step 2 below (Understanding the *blastn* results).

You may easily run single or multiple BLAST searches (that is, a batch search) using NCBI's BLAST within the Geneious Prime program. Geneious Prime supports searching for RNA, DNA, and protein sequence.

IMPORTANT NOTE: Regarding BLAST searches using Geneious Prime: In general, the amount of time it takes to retrieve BLAST results will vary depending on how many searches NCBI BLAST is asked to run at any particular moment from researchers around the world. In some cases, searches performed through Geneious Prime are not as fast as performing the BLAST searches directly through NCBI.

If you have short class periods (50 min or less) and need the BLAST results as soon as possible to enable data analysis, you have the option of performing the BLAST searches directly from NCBI's website for this section.

Please consult your instructor as to whether you will be performing the BLAST search using Geneious Prime or using NCBI's BLAST website. Use Appendix G for protocol steps on how to export sequences as FASTA files for BLAST searching directly on the NCBI website.

1. Using BLAST to perform a sequence search at NCBI using Geneious Prime.

The most efficient way to confirm your results is to do a BLAST search with all your sequences from one miniprep clone in one batch, but BLAST searches could also be performed on individual sequences. Be sure that your computer is connected to the Internet to perform BLAST searches.

Note: Instructions on the batch (step 1.1) or the single (step 1.2) BLAST steps are listed below. You will need to perform only one or the other, you choose!

1.1 Perform a BLAST search on multiple sequences against the NCBI sequence database.

1.1.1 Select more than one sequence in the document table. These will be your vector- and quality-trimmed sequences, which are also referred to as your queries.

1.1.2 Select the BLAST icon in the menu bar. A new dialog box will open:

BLAST

Query: ☒ Batch search of 4 nucleotide sequences
☐ Selected region
☐ Enter unformatted or FASTA sequence

Database: Nucleotide collection (nr/nt) (AA or nt)

Program: blastn - similar matches (DNA query, DNA)

Results: Hit table ?

Retrieve: Matching region

Maximum Hits: 20

☒ Low Complexity Filter Max E-value: 0.05
☒ Mask for lookup table Word Size: 15
☐ Human Repeats Filter Gap cost (Open Extend): 5 2

Scoring (Match Mismatch): 2 -3 Max Target Seqs: 100

Entrez Query:

Other Arguments:

BLAST search dialog box. A new window will open after clicking the BLAST icon in the menu bar. Fill in the dialog box for your BLAST search with the same information you see in this image.

The Query will be listed as “Batch search of 4 nucleotide sequences” (or however many sequences you selected).

- Select NCBI Reference Genomic Sequences from the Database dropdown menu
- Select **blastn** for Program
- Select **Hit table** for Results
- Select **Matching region with annotations (slow)** for Retrieve
- Enter **20** under **Maximum Hits**
- Click **More Options** at the bottom left of the window
- Select **15** (or the largest number available) for Word Size
- Keep default values for other selections
- Click **Search**

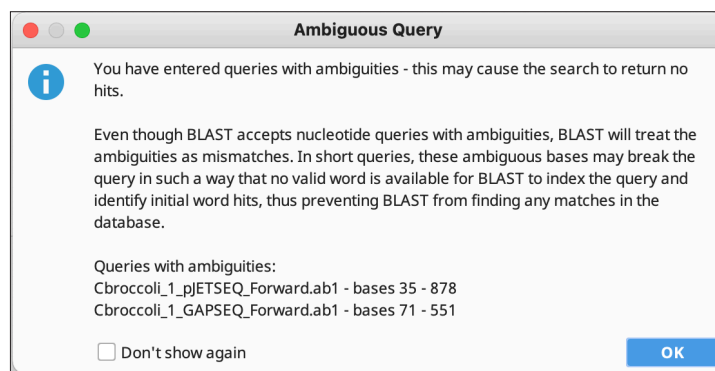
1.2 Perform a BLAST search on a single sequence against the NCBI sequence database using Geneious Prime.

1.2.1 Select one sequence in the document table.

1.2.2 Select the BLAST icon in the menu bar. A new dialog box will open.

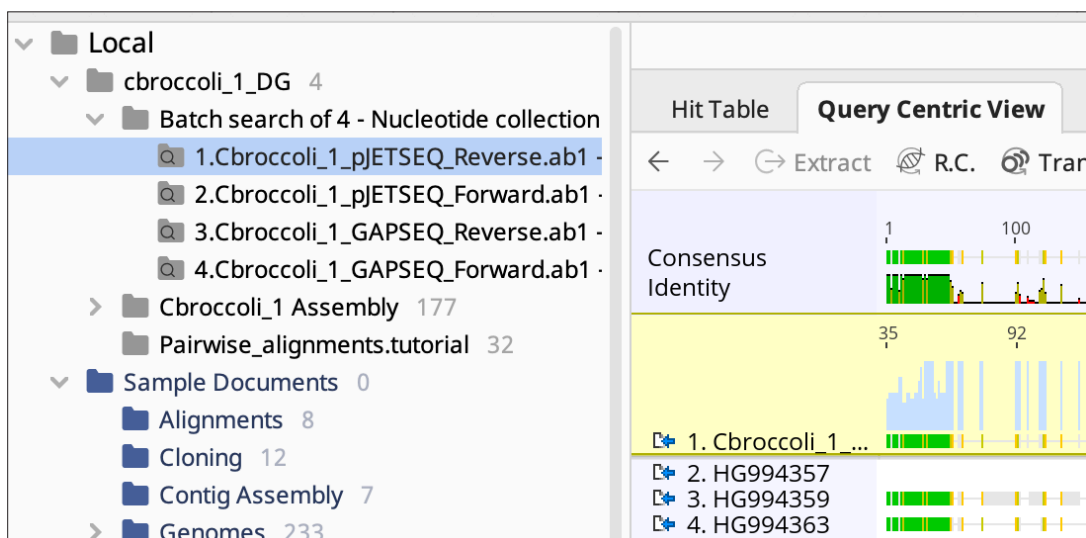
1.2.2.1 Follow step 1.1.2 to set up your BLAST search.

Note: If the Ambiguous Query dialog box appears, it means there are still some ambiguous bases (Ns) in your single sequences, which may interfere with the BLAST search. You may click **OK** to go on. Geneious Prime will send your query to the NCBI and create a New Search folder. The time it takes for the search to be completed will depend on a number of variables: Internet speed, number selected for maximum hits, and which Results parameter was selected.



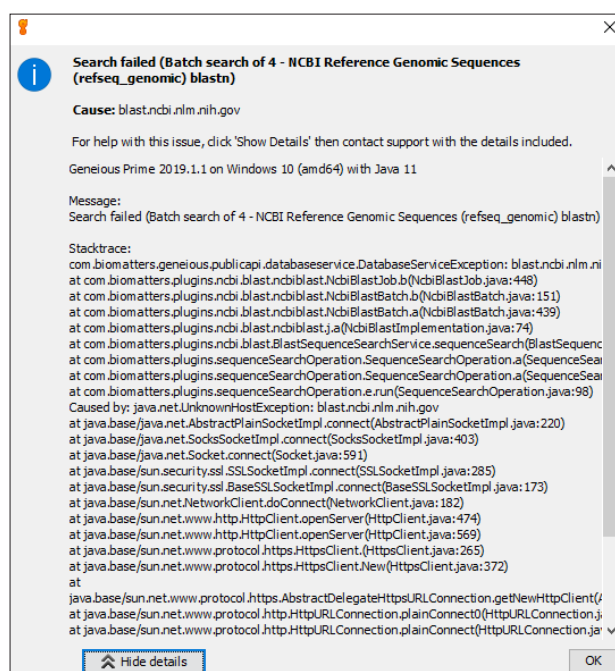
Tip: When results start downloading, a magnifying glass icon will appear on the folder. You can begin your analysis as soon as results start showing up in your folder.

1.2.3 A new folder will be created, called “Batch search of 4 -,” which contains one folder for each single sequence (if you searched on a single sequence, the Batch folder will not exist). When one of these folders is selected, you will see that the document table has a few new tabs, including one labeled Hit Table.



Each sequence in the batch search has its own folder containing the results from the BLAST search.

- 1.2.4 If a new window opens with the title “Search failed,” read the description to learn the cause. In this case, not all the results were recovered from NCBI:



- 1.2.4.1 Click on each result folder. If you don't see a tab in the document table for Query Centric View, the BLAST search will need to be rerun. Repeat section 1.2 to rerun the BLAST, but this time set Maximum Hits to 10. You will need at least ten results to get a feel for what gene most closely resembles your clone. Think of BLAST as doing an experiment. To get relevant results, you need to optimize the experimental conditions settings.

2. Understanding the blastn results using the Hit Table.

The results from a blastn search include many different kinds of information and statistics. These bits of information include the size of the database, length of each query sequence, statistics that describe the number and percent of matching bases, a BLAST score, and the E value.

The sequences in the example shown below come from a *GAPDH* cloning experiment with a plant from the genus *Brassica* (cabbage). Your results may differ from those shown in this manual.

On the Hit Table tab of the document table, you will find summary statistics for the search results. Each row contains a matching sequence with the best-matching sequence at the top of the table. The column labeled **E value** indicates the expected frequency of an alignment's occurrence by chance. The thing to remember is that the smaller the number, the better the match.

For example, the top hit 1.54e-168, which is equal to 1.54×10^{-168} , this is a very small number and indicates that it is highly unlikely that this alignment would ever occur by chance. You may even have examples where the E value reads 0.0000. This is telling you that statistically there is no likelihood that this alignment has happened by chance. Use these statistics as a guide; alignments that have larger E values, and thus may appear far less significant, can still be interesting. The exact values will change as the database size increases.

In addition to the E value, there is also a column labeled **% Pairwise Identity**. (You may need to scroll sideways to find this column.) Drag this column over next to the E value. It is also useful as it will indicate how similar the sequence found in the database is to the one you used as a query. Note that sorting by E value or % Pairwise Identity can produce a different ordering because statistical significance is related to alignment length as well as identity, but identity relates only to the aligned region. For example, consider the alignment in the figure below. In this alignment, the % Pairwise Identity is 96% with *Oryza sativa* (rice). However, when you examine the matching regions in more detail, you find that the region where 96% of the bases match is only 28 nucleotides long. This is a good example of how short sequences can give a good match that is not meaningful.

```

Features in this part of subject sequence:
  hypothetical protein

Score = 46.4 bits (50), Expect = 0.007
Identities = 27/28 (96%), Gaps = 0/28 (0%)
Strand=Plus/Minus

Query  6          AGCCTTGGCATCAAAGATGCTCGACCTG  33
      |||||
Sbjct 23549447    AGCCTTGGCATCAAAGATGCTGGACCTG  23549420

```

Alignment of rice sequence with query sequence. This is a sequence that has a high % Pairwise Identity score (96%). However, it is a very short sequence, so the match is not meaningful.

Tip: It is useful to look at alignments in isolation, but looking at alignments together reveals more information. The Geneious Prime screen called **Query Centric View** provides a multiple alignment-style visualization of the BLAST hits mapped against the original query sequence. This isn't a true multiple sequence alignment, but instead a mapping of the individual BLAST hits against the query sequence. It is a useful way to see where the conserved regions in the BLAST search are lining up against the query. Keep in mind that BLAST alignments are local alignments.

Nucleotide BLAST: Search nucleotide sequences

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

BLAST

Search database Reference genomic sequences (refseq_genomic) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: ♦ 50 (Select the maximum number of aligned sequences to display)

Short queries: ☒ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: ♦ 15

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: ☒ Low complexity regions ☐ Species-specific repeats for: Homo sapiens (Human)

Mask: ☒ Mask for lookup table only ☐ Mask lower case letters

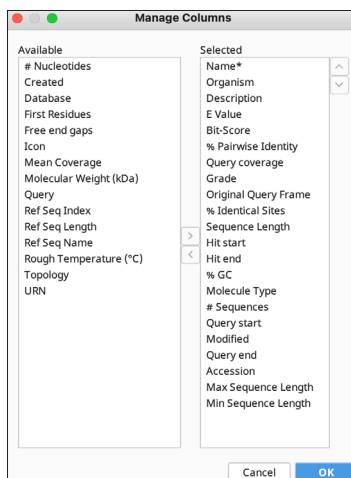
BLAST

Search database Reference genomic sequences (refseq_genomic) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Example of BLAST results in a Hit Table.

2.1 Summary of the major categories on the Hit Table. There are many columns that can be displayed or hidden. You might have to scroll to the right to find some of these columns, or they may not be automatically displayed at all. To display/hide additional column headers, choose the icon that looks like a small data table just above the scroll bar on the right (the popup bubble will tell you this icon is called "Change the visible columns"). This will reveal all the column options that are available.



Manage Columns lets you choose to view the data that will be most helpful to you. Click the small data table icon, then select Manage Columns from the list. A dialog box will open in which you can select your options.

Name: A sequence's name is its accession number, which is the unique identifier of your sequence within a database. The main public database that is used for storing and distributing sequence data is NCBI's GenBank. Accession numbers are also used to report sequences in scientific papers and journals.

Description: A brief description of hit matches, including the scientific name of the organism and the chromosomal location if known.

Query coverage: The length of your sequence covered by the one found with the BLAST search.

E Value: The expect (E) value is the number of hits one can expect to see by chance when searching a database. The size of the database being searched will affect E value calculations. For example, an E value of 1 means that in a database of the current size one might expect to see one match with a similar score purely by chance. The E value describes the random background noise in a match, so it decreases exponentially as the score (S) (see Bit-Score), the assessment of an alignment's overall quality, of the match increases.

Generally, the closer the E value is to zero, the more significant the match. An exception is that virtually identical short alignments have relatively high E values because shorter sequences have a higher probability of occurring in a database purely by chance.

Tip: The E value can be a convenient way to create a significance threshold for reporting results. You can change the E value threshold within Geneious Prime easily. Raising the E value threshold will produce a longer list, but more of the hits will have low scores.

Bit-Score: The score (S) describes the overall quality of an alignment; higher values correspond to greater similarity. The bit-score takes the statistical properties of the scoring system into account to normalize an alignment's score (S). Every search is unique, but a bit-score allows alignment scores (S) from different searches, which may have been conducted with different algorithms using different values, to be compared.

Grade: A percentage calculated by Geneious Prime by weighting the query coverage, E value, and identity value (0.5, 0.25, and 0.25 respectively) for each hit. This allows you to sort hits so that the longest, highest identity hits are at the top.


% Pairwise Identity: This is the value, expressed as a percentage, of how similar two sequences (nucleotide or amino acid) are.

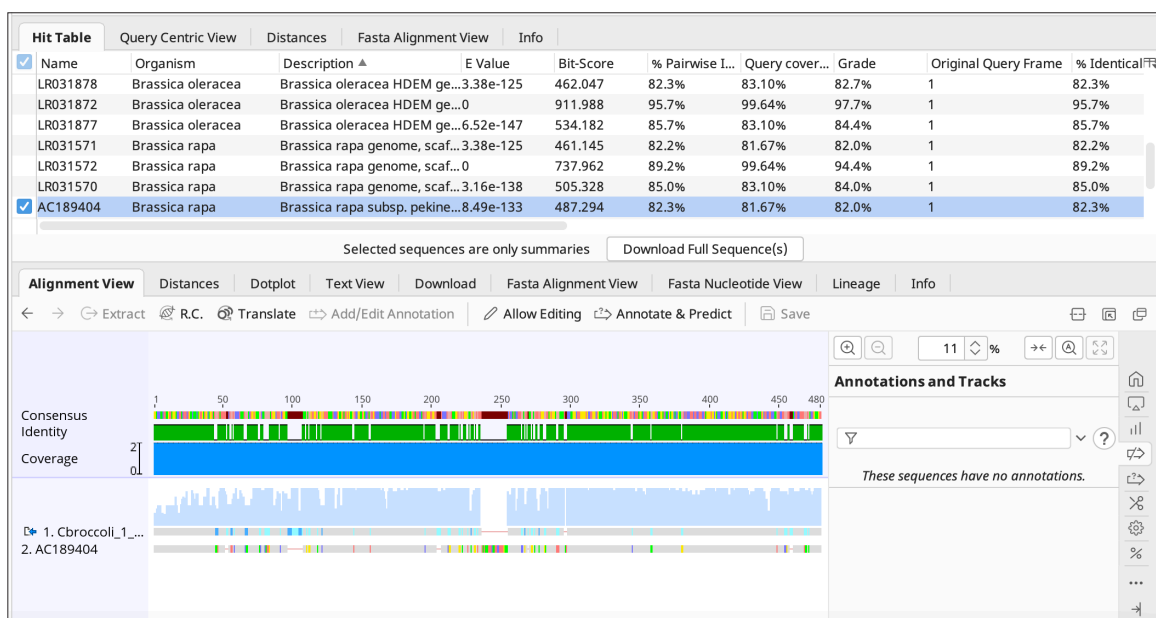
% Identical Sites (PID): The percentage identity for two sequences can be variable and depends on many factors. The alignment method and parameters used to compare the sequences will affect the sequence alignment. PID is strongly length-dependent, which means that the shorter a pair of sequences is, the higher the PID you might expect by chance.

3. Looking at the alignments in the document viewer panel.

Now that you have a set of search results, you will need to look at some alignments. You can click any sequence in the Hit Table list and Geneious Prime will display the pairwise alignment for that hit in the document viewer window.

- 3.1** Click on one BLAST result from the Hit Table (let's say one with a description of *Brassica rapa*) and look in the Alignment View tab in document viewer. You will see a zoomed out view of the query aligned to the BLAST hit. From top to bottom in the document viewer you will see:

- The consensus sequence, displayed on top
- The depth of coverage chart, displayed in blue. If the depth of coverage chart is not visible, go to the Graphs tab  in the options panel and check the box labeled Coverage
- A graphical representation of % pairwise identity, displayed in green
- Your query sequence, then the sequence for the hit



Alignment view for one BLAST hit. A *Brassica rapa* query result is chosen for comparison with the GAP SEQ F read from cbroccoli. In Alignment View, you can see information such as the consensus sequence, depth of coverage, identity chart, and known annotations from *Brassica rapa*.

- 3.2** In the Annotations and Tracks tab ⇨ in the options panel, make sure that the box for Show Annotations is checked. This way, you will see whether there are any *GAPC* genes within this *Brassica* chromosome.

Tip: Mousing over the annotations will display a pop up with more information, including the gene name and gene product, if known. You can select text from within this pop up and copy the text into another file or an electronic lab notebook.

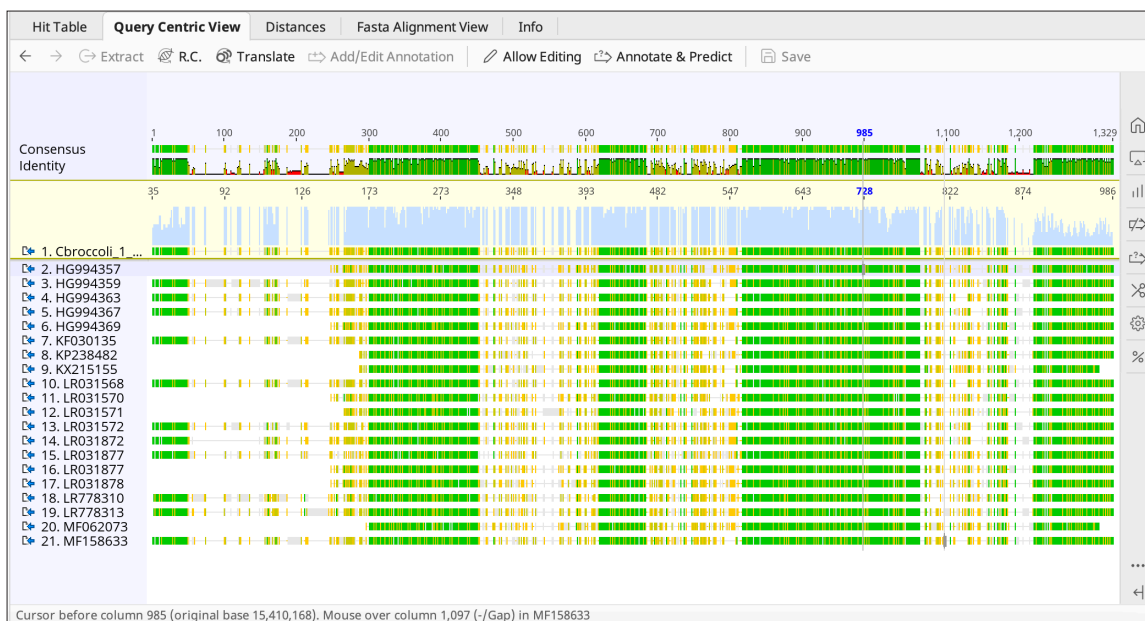
- 3.3** A consensus sequence is an alignment that occurs, with minor variations, across many genetic locations or organisms. It is constructed from the order of the nucleotides appearing most frequently at each position of a sequence alignment. The consensus is the same length as the contig (which includes only untrimmed bases), and shows which nucleotides are conserved and which are variable. For a nucleotide to be selected for the consensus, it must reach a minimum threshold of occurrence in that position in a variety of sequences. The consensus sequence is available when viewing alignments or contig documents, and is displayed when the box for Consensus is checked in the General tab ⓘ in the options panel.

Tip: Ambiguity codes, such as an R designation for a nucleotide that could be either an A or a G, are counted as fractional support for each nucleotide in the ambiguity set (A and G, in this case). Thus, two rows with Rs are counted the same as one row with an A and one row with a G.

Tip: When “Ignore gaps” is checked (in the Display tab ⓘ of the options panel), the consensus is calculated as if each alignment column consisted only of the non-gap characters. Otherwise, the gap character is treated like a normal nucleotide. However, mixing a gap with any other nucleotide in the consensus always produces the total ambiguity symbol (N for nucleotides and X for amino acids).

- 3.4** Depth of coverage represents the number (often an average) of nucleotides contributing to a portion of assembled sequences. On a whole-genome basis it means that each base has been sequenced, on average, a particular number of times (for example, 10x, 20x, etc.). For a specific nucleotide, it represents the number of reads that contributed information about that nucleotide. The depth can vary depending on the genomic region being sequenced. In the figure above (the single-hit alignment to *Brassica*), the depth of coverage across the contig is 2.

- 3.5** Click the Query Centric View tab in the document table. You will see all the hits with annotations aligned to your query sequence displayed. This gives you a quick survey of how many hits have annotations for the *GAPC* family of genes:



Viewing the BLAST query results in Query Centric View.

4. Using Text View in the document viewer (when Hit Table is selected).

You can use the Text View tab to take a quick look at the information from the BLAST hit.

- At the top of the page is the accession number and the description of the BLAST hit
- Beneath that is a summary of the statistics for the hit, such as length, bit-score, E value, etc.
- Next is the alignment in text format. Your query is located at the top, followed by the consensus sequence in the middle and the BLAST hit on the bottom
 - The numbers at the ends of the sequence refer to the gene's location on the chromosome
 - No letter means there is no match between the query and the BLAST hit
 - A dash means a gap

Selected sequences are

Alignment View Annotations Distances Dotplot **Text View** Download Fasta Alignment View Fasta Nucleotide View Lineage Info

Sequence format: Don't reformat

```
>NC_003074 Arabidopsis thaliana chromosome 3 sequence
Length = 23459830

E-value = 0, Score = 911, Bitscore = 822.721, Identities = 799/1003 (79%),
Positives = 799/1003 (79%), Gaps = 60/1003 (5%)
Frame = +1

Contig          1 AGCCTTGGCATCAAAGATGCTTGACCTGTTGTCCACCAACGAAGTCAGTCGAGACAACATC      60
                AGCCTTGGC TCAAA ATGCT GACCTGTTGTC CCAACGAAGTCAGT GAGACAACATC
NC_003074      1083023 AGCCTTGGCGTCAAAAATGCTCGACCTGTTGTGCCAACGAAGTCAGTTGAGACAACATC 1082964


Contig          61 ATCCTCGGTGTAACCAAGGATTCCCTTTAGCTTGCCCTCAGATTCTCCCTGTGTAAACA      120
                ATCCTCGGTGTA CCAAGGATTCCCTT AG TTGCC TC GATTCTCCCT TGTA
NC_003074      1082963 ATCCTCGGTGTATCCAAGGATTCCCTTGAGTTGCGCTTCGGATTCTCCCTATGTATTGG 1082904

Contig          121 TCC---AATCAGTTAGTTA--AATCTTA---CAAAG--ATTGATCAAACTAGTTAACT      170
                T   AA CA TT GT A AAT T A   C A G AT T A AAACT T A CT
NC_003074      1082903 TGGTAAACCACTTGGTGAGTAATTTGATCACTACTGGAATATAAGTAACT--TAATCT 1082846

Contig          171 GGAACCG----AAAGCTTACTTGATAGCCTTCTTGATCTGGTCGTAGGTAGCAGCTTTCT      226
                G A G   AAAGCTTACTTGATAGCCTT TTGAT T TCGTAGGTAGCAGCTTTCT
NC_003074      1082845 GTCATTGCTCAAAAGCTTACTTGATAGCCTTTTGGATTTTCATCGTAGGTAGCAGCTTTCT 1082786

Contig          227 CGAGTCTAACCGTGAGGTCAACAACTGAGACATCAACGGTGGGAACACGGAAGGACATTC      286
                CGAGTCT AC GT AGGTCAACAACTGAGACATCAACGGT GG ACACGGAA GACATTC
NC_003074      1082785 CGAGTCTGACAGTAAGGTCAACAACTGAGACATCAACGGTTGGGACCGGAAGACATTC 1082726

Contig          287 CTGTCAACTTTCCATTGAGCTGTGGAAGCACCTTTCCGACAGCCTTGGCAGCTCCAGTGC      346
                C GTCAACTTTCC TT AG   TGGAAGCACCTTTCCGACAGCCTTGGCAGCTCCAGTGC
NC_003074      1082725 CAGTCAACTTTCCGTTAAGAGCTGGAAGCACCTTTCCGACAGCCTTGGCAGCTCCAGTGC 1082666
```

Text view of the Cbroccoli_1_GAPSEQF sequence compared with the *Arabidopsis thaliana* BLAST query result. The alignment between the cbroccoli_1_GAPSEQF and the *Arabidopsis* sequence (NC_003074) is displayed in Text View. The Text View was expanded by clicking the Expand Viewer button  in the options panel, which hides the Sources panel on the left and the document table at the top of the Geneious Prime window.

5. BLAST searching the contig and recording matches.

For your contig, record the top three matches and their statistics in the charts below.

pJET SEQ F Sequence

Description	E value	Bit-Score	% Pairwise Identity	Query Coverage	Grade

pJET SEQ R Sequence

Description	E value	Bit-Score	% Pairwise Identity	Query Coverage	Grade

GAP SEQ F Sequence

Description	E value	Bit-Score	% Pairwise Identity	Query Coverage	Grade

GAP SEQ R Sequence

Description	E value	Bit-Score	% Pairwise Identity	Query Coverage	Grade

6. Verification of sequence.

It is possible that PCR products have been generated as a result of contamination by the control PCR reactions. To verify that your sequence is not an *Arabidopsis* gene, you must compare it against the BLAST results.

Note: If your goal was to clone an *Arabidopsis thaliana* gene, then this is not necessary).

6.1 In the Hit Table, sort the results by clicking on the column headers for Bit-Score, E value, and Grade.

6.2 View your BLAST results in the sorted Hit Table. Using the Accession and Description columns, look to see whether any of your top hits come from *Arabidopsis thaliana*. In some cases, the results can be ambiguous and you may have to wait to verify which gene has been cloned until a contiguous sequence (contig) of all sequences from one clone have been assembled and there are more data to work with (which will be done in Section 2).

A number of scenarios can occur, depending on how your cloning reaction took place:

- If the top match to your sequence IS NOT an *Arabidopsis* sequence, it is unlikely that you have cloned an *Arabidopsis* gene
- If your top match IS an *Arabidopsis* sequence, then look at the sequence alignment of your novel sequence with the *Arabidopsis* sequence
- If the aligned sequence is broken up into two or more sections, this suggests there is a region that does not match the subject sequence, indicating your sequence IS NOT from an *Arabidopsis* gene
- If the entire query sequence aligns in a single block with *Arabidopsis thaliana*, then look at the green Identity graph at the top of the sequence alignment, just beneath the blue depth of coverage chart. This graph indicates the homology of the query sequence with the subject sequence. Mousing over the chart will display the value as a percentage; clicking a base will display the value as a fraction in the bottom left corner of the window
- If the Identity value is below 90%, then it is unlikely to be an *Arabidopsis GAPC* gene
- If your sequence has low homology or has gaps that have no homology, it is unlikely to be an *Arabidopsis* gene
- If your sequence has high homology with the same *Arabidopsis* gene, it is likely that the gene is from *Arabidopsis* and may have been cloned accidentally. However some plant species close to *Arabidopsis* may have close homology. You will need to determine how closely your plant is related to *Arabidopsis* and whether to continue analysis with this plasmid

APPENDIX I: GLOSSARY

Agar – A jelly-like substance obtained from seaweed. It is made of linked sugars (a polysaccharide) and is used in making medium for growing bacteria.

Ampicillin – A penicillin-like bactericidal antibiotic that inhibits the synthesis of the peptidoglycan component of bacterial cell walls, especially in gram-positive bacteria but also in some gram-negative bacteria such as *E. coli*.

Annealing – Binding of single-stranded DNA to complementary DNA sequences. Oligonucleotide primers bind to single-stranded (denatured) template DNA.

Annotating – The process of identifying the protein coding sequences and other biological features within genomic DNA sequences and adding such information to the sequence.

Antibiotic – A chemical that prevents or reduces the growth of bacteria or other microbes.

Antiparallel strands – DNA strands oriented in opposite directions, such that the 5'-phosphate end of one strand is aligned with the 3'-hydroxyl of the other strand.

Assembly – Aligning and merging shorter sequences of a much longer DNA sequence in order to reconstruct the original sequence. Assembly is used to generate a significant portion of a genomic DNA sequence because current technology only allows for sequencing of 600–800 base pair (bp) fragments of DNA with high fidelity.

Bacteria – Single-cell microorganisms with no nucleus.

Bactericidal – A term used to describe an antibiotic or other agent that kills bacteria.

Bacteriostatic – A term used to describe an antibiotic or other agent that prevents the growth of bacteria.

Base call – Assigning a base to a peak when reading a DNA sequencing chromatograph.

Base pairs – Complementary nucleotides held together by hydrogen bonds. In DNA, adenine is bonded by 2 hydrogen bonds with thymine (A–T) and guanine with cytosine by 3 hydrogen bonds (G–C). Because of the 3 H-bonds between G and C (compared to the 2 between A and T), the GC bonding is stronger than the AT bonding.

Binary fission – A process by which most bacteria reproduce asexually by duplicating their DNA and dividing into two equal halves.

Bioinformatics – Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

BLAST – Abbreviation for Basic Local Alignment Search Tool, a suite of computer programs used to compare DNA and protein sequences to those in libraries of databases to search for similarities (www.ncbi.nlm.nih.gov/blast/Blast.cgi).

Chloroplast – An organelle found in plant cells, responsible for photosynthesis. A type of plastid.

Chromatogram – A visual representation of the signal peaks detected by a sequencing instrument. The chromatogram contains information on the signal intensity as well as the peak separation time.

Chromatography – A technique for separating molecules based on their physical characteristics, such as size (size exclusion chromatography), charge (ion exchange chromatography), or hydrophobicity (hydrophobic interaction chromatography). Affinity chromatography is another type of chromatography that separates molecules based on their specific activity, cofactor, or chemistry (such as sugars on a glycoprotein).

Clone – An organism or cell, or group of organisms or cells, produced asexually from one parent cell or organism, to which they are genetically identical.

Codon bias – Although there are multiple codons for many amino acids, specific organisms tend to favor particular codons instead of using all indiscriminately.

Codon – A set of 3 nucleotides (in DNA or mRNA) that code for a single amino acid. Many amino acids have multiple codons.

Colony – A bacterial colony is a group of bacteria on growth media that usually have grown from a single bacterium. A colony is thus a clone of identical organisms.

Complementarity – Binding of two strands of nucleotides derived from the base pairing of the nucleotides (AT and GC). The two DNA strands of the double helix have complementarity, as do DNA or RNA primers bound to template DNA.

Consensus sequence – A sequence derived from the alignment of similar DNA, RNA, or protein sequences. Normally, each position in the consensus sequence is determined by the base or amino acid that predominates in that position in the majority of the aligned sequences. Consensus sequences can be used to design primers for PCR.

Constitutive gene expression – Refers to genes that are transcribed continuously, such as housekeeping genes.

Contig – A sequence that has been constructed by comparing and merging the information from sets of overlapping DNA segments.

Cytosol – The internal fluid of the cell where a portion of cell metabolism occurs. Proteins within the cytosol play an important role in signal transduction pathways and glycolysis.

Degenerate primers – A mixture of PCR primers that are similar, but not identical. They may be designed based on a consensus sequence derived from similar organisms, with substitutions of different bases at one or more locations in the primers.

Denaturation – With respect to DNA, separation of complementary strands of DNA into single stranded DNA. Denaturation of DNA is also sometimes referred to as “melting”. In vivo, DNA is denatured by enzymes, but in PCR, DNA is denatured by heat.

Depth of coverage – Multiple reads of the same sequence. Two methods for obtaining multiple reads are: 1) using different primers to sequence the same clone of a gene, or 2) sequencing unique clones of the same gene.

Disagreement – Differences in base calls between two or more different sequences of the same clone or between different clones of the same gene. Synonym for “discrepancy”.

DNA ligase – An enzyme that repairs single-strand breaks in double-stranded DNA.

DNA polymerase – An enzyme involved in DNA replication. DNA polymerase links a new nucleotide to a growing strand of DNA. The new nucleotide is connected via a phosphodiester bond to the 3'-hydroxyl group of the most recently incorporated nucleotide on the new strand.

DNA replication – The process of copying a DNA molecule. A double-stranded DNA molecule is replicated to form two identical double-stranded DNA molecules.

E. coli – *Escherichia coli* is a gram-negative facultative anaerobic bacillus bacterium. It inhabits the intestines of animals and humans and may benefit them by producing vitamin K and preventing the spread of harmful bacteria. Harmless genetically weakened forms of *E. coli* such as the HB101 strain used in this kit are used in many scientific applications. Normally *E. coli* is harmless but a few strains, such as O157:H7, can cause disease.

Electrophoresis – A technique for separating molecules based on their relative migrations in an electric field. DNA and RNA are usually separated using agarose gel electrophoresis, and proteins are separated using a polyacrylamide matrix (PAGE or SDS-PAGE).

Endosymbiotic theory – The theory that plastids and mitochondria exist as subcellular organelles in modern eukaryotic cells as a result of an ancient symbiotic event between eukaryotes and prokaryotes. For example, plastids are believed to have been derived from photosynthetic cyanobacteria that were engulfed by eukaryotic cells more than a billion years ago.

Exon – A segment of a eukaryotic gene that is transcribed to RNA and retained after RNA processing. An exon becomes part of the mRNA that gets translated to protein. Exon can refer to either the DNA sequence or the RNA transcript. Exons are separated in DNA and in the primary RNA transcript by introns (see definition for Intron). Exons are also known as the protein coding sequences of genes, while introns are known as noncoding regions.

Exonuclease – An enzyme that removes nucleotides from the ends of DNA strands. Exonuclease I, used in this experiment, removes nucleotides in a stepwise manner from the 3'-hydroxyl end of single-stranded DNA.

Extension – The step in PCR in which the new strand is extended (or elongated) from the primer by addition of dNTPs to the 3'-end of the growing strand.

FASTA format – A format used for submitting sequence data (DNA bases or amino acids) to alignment programs. The first line is a description of the data, beginning with the greater-than (>) symbol and ending with a paragraph break, with no spaces within the line. FASTA format uses single letter codes for the sequence, with no spaces or paragraph breaks within the sequence.

Finishing – A process in which researchers examine the contigs to look for misassemblies or regions that require additional coverage.

GAPDH – (glyceraldehyde-3-phosphate dehydrogenase) – The enzyme that catalyzes the sixth reaction of glycolysis, oxidizing glyceraldehyde-3-phosphate into 1,3-bisphosphoglycerate.

GC clamp – One to three G or C bases at the 3'-end of a primer. If there is a GC clamp at the 3'-end of the primer, the primer will form a more stable complex with the template DNA.

GenBank – The sequence database maintained by the NIH. As of February 2008, the GenBank contained 85,759,586,764 bases in 82,853,685 sequence records. To access GenBank, go to <http://www.ncbi.nlm.nih.gov/Genbank/>

Gene duplication – The duplication of a region of DNA containing a gene. Gene duplications have been important during evolution, as once the genes are duplicated, the gene copy can mutate to create a different gene. Groups of genes that have resulted from gene duplication events, such as the genes that code for GAPDH, are called gene families.

Genome – The total genetic material of an organism.

Genomic DNA (gDNA) – All of the chromosomal DNA found in a cell or organism.

Glycolysis – The pathway by which glucose is converted into pyruvate in a series of ten enzymatic reactions, producing energy for the cell as well as precursors for many biological molecules.

Homologous – Genes that are similar because they share a common ancestor.

Homology (of DNA or proteins) – Regions of protein or DNA that have a high level of sequence similarity due to shared ancestry. However, sequence similarity does not necessarily indicate homology, especially if the similar sequences are short.

Housekeeping gene – A gene that is expressed constitutively in cells. A housekeeping gene is necessary for the cell to survive.

Indel – A sequence discrepancy due to either an inserted or a deleted base.

Intron – A eukaryotic gene segment that is transcribed to RNA and spliced out from the primary transcript during RNA processing. Introns are interspersed between exons (see definition for Exon) and are also known as noncoding regions of DNA.

Isozyme – Enzymes that catalyze the same reaction, but differ in amino acid sequence or another physical characteristic. Isozymes are coded at different loci in the organism's genome and result from gene duplication events during evolution.

LB – Lysogeny broth (sometimes called Luria Bertani broth) is composed of yeast extract, tryptone, and sodium chloride and commonly used to culture bacteria.

Lysis – The process of rupturing a cell to release its contents. Once lysed, the mixture of the cell and lysis solution is called a lysate.

Master mix – A premixed reagent solution for chemical or biological reactions. A PCR master mix contains all components needed for PCR (dNTPs, primers, buffer, salts, DNA polymerase, and Mg^{2+}) except for the template DNA.

Melting temperature (T_m) – The temperature at which the two strands of a DNA molecule dissociate.

Nested PCR – A variation of PCR in which two sequential rounds of PCR are performed, each with a different set of primers. The first set of primers binds outside of the region of interest, and the second set binds within the region amplified by the first set of primers.

Nucleotide – A fundamental unit of DNA and RNA. Molecules composed of a sugar, a phosphate group, and one of four bases: adenine, guanine, cytosine, and thymine (DNA) or uracil (RNA).

Okazaki fragment – Short pieces (100–2,000 bases) of DNA synthesized on the lagging strand during DNA synthesis. The fragments are later linked by DNA ligase.

Oligonucleotide (oligo) – A short segment (often 10–30 base pairs) of DNA or RNA that is usually synthesized synthetically. Frequently used as primers for PCR or sequencing.

Organic extraction – A technique used to separate molecules based on their differential solubility in solutions that do not mix, such as an aqueous solution and an organic solution. For extraction of DNA using this method, DNA will remain in the aqueous solution, and contaminants such as proteins will be in the organic phase.

Origin of replication – A particular sequence on a molecule of DNA where DNA replication is started. Circular plasmid DNA normally has a single origin of replication (ORI), but eukaryotic genomic DNA has many origins on each molecule.

Orthologous – Genes that share a high level of homology but are from different species.

Paralogous – Genes that share a high level of homology and are from the same genome.

PCR – Polymerase chain reaction. A technique for rapidly creating multiple copies of a segment of DNA utilizing repeated cycles of DNA synthesis.

Penicillin – A bactericidal antibiotic that inhibits the synthesis of the peptidoglycan component of bacterial cell walls, especially in gram-positive bacteria. Penicillin was discovered by Alexander Fleming in 1928, and was the first antibiotic to be used medically.

Petri dish – Small, round, flat containers made of glass or plastic. It is commonly used to hold media used to culture microbes. Petri dishes were invented by microbiologist Julius Petri, an assistant to Robert Koch.

Plastid – A double membrane-bound organelle found in the cytoplasm of plants. Different plastids perform different roles in plants, including photosynthesis and storage of metabolites and pigments. Plastids include chloroplasts, leucoplasts, and chromoplasts.

Primer specificity – The degree to which the primer sequence complements the template sequence. The more specific the primer is for the template, the higher the temperature at which the primer and template will anneal to each other. Conversely, if there are mismatches between primer and template, the annealing temperature will be lower; in fact, the primer may dissociate from the template prior to amplification.

Primer – Short, single-stranded oligonucleotide (usually 18–24 bases in length) designed to bind to DNA template strands at the end of the sequence of interest and serve as the starting point for DNA synthesis. Primers can be either single-stranded DNA or RNA.

Quality score (or value) – A numerical value used in DNA sequencing data indicating the confidence level for base calls. A higher quality value means higher confidence that the base call is correct. A lower quality value indicates the base call is less reliable.

Query sequence – The input sequence (or other type of search term) with which all of the entries in a database are to be compared (ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html).

Query – This is a program written in SQL (see definition for SQL) that is used to extract information from the database.

Read – Sequences of bases that contain information about the parent chromatogram. As long as the base sequence is linked to the chromatogram, it can be considered a read.

Relational database – A database consisting of multiple tables of information, based on a model of the data and the relationships between different types of data (for example, a DNA sequence that is related to and linked to a chromatogram and also to information about the sample).

Replication fidelity – The number of mistakes made by DNA polymerase as it adds the bases during DNA replication. High-fidelity DNA polymerase makes few errors. For example, eukaryotic DNA polymerase, which makes one mistake every 10,000–100,000 base pairs replicated is considered to be high-fidelity. *Taq* DNA polymerase has low replication fidelity.

Sequence – The ordered list of bases that make up a DNA strand. When linked with a chromatogram, this would be considered a read.

SQL – Structured query language, a programming language used to extract relationships between different data sets in a relational database.

Subject sequence – A sequence found by BLAST to have similarity to a sequence entered by the user (the query sequence).

Taq DNA polymerase – A DNA polymerase that is stable at high temperatures. *Taq* DNA polymerase is commonly used in PCR. The enzyme was originally isolated from the thermophilic bacterium *Thermus aquaticus*, which can tolerate high temperatures.

Template DNA – The “target DNA” that contains the sequence to be amplified by PCR.

Thermal cycler – An instrument used in PCR that automates the repeated cycles of heating and cooling.

Transcription – Synthesis of mRNA from a DNA template.

Translation – The synthesis of amino acids from mRNA, producing proteins.

APPENDIX J: REFERENCES

- Allison LA (2007). *Fundamental Molecular Biology*. (Malden, MA: Blackwell Publishing).
- Altenberg B and Greulich KO (2004). Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes. *Genomics* 84, 1014–1020.
- Baxeavanis AD (2006). The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics* 1: Chapter 1 Unit 1.1.
- Baxeavanis AD and Ouellette BF, ed. (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. (Hoboken: John Wiley & Sons).
- Benson DA et al. (2007). GenBank. *Nucleic Acids Res* 35, D21–D25.
- Biesecker G et al. (1977). Sequence and structure of d-glyceraldehyde 3-phosphate from *Bacillus stearothermophilus*. *Nature* 266, 328–333.
- Boyle JA (2004). Bioinformatics in undergraduate education. *Biochem Mol Bio Educ* 32, 236–238.
- Bray EA et al. (2000). Responses to Abiotic Stresses (Chapter 22). In: *Biochemistry and Molecular Biology of Plants*. Buchanan BB, Gruissem W, Jones RL, ed. (Rockville, MD: American Society of Plant Physiologists).
- Brinkmann H et al. (1989). Cloning and sequence analysis of cDNAs encoding the cytosolic precursors of subunits GAPA and GAPB of chloroplast glyceraldehyde-3-phosphate dehydrogenase from pea and spinach. *Plant Mol Biol* 13, 81–94.
- Campbell AM (2003). Public access for teaching genomics, proteomics, and bioinformatics. *Cell Biol Educ* 2, 98–111.
- Caspers ML (2003). An undergraduate biochemistry laboratory course with an emphasis on a research experience. *Biochem and Mol Biol Educ* 31, 303–307.
- Clermont S et al. (1993). Determinants of coenzyme specificity in glyceraldehyde-3-phosphate dehydrogenase: Role of acidic residue in the fingerprint region of the nucleotide binding fold. *Biochemistry* 32, 10178–10184.
- Dennis DT and Blakely SD (2000). Carbohydrate Metabolism (Chapter 13). In: *Biochemistry and Molecular Biology of Plants*. Buchanan BB, Gruissem W, and Jones RL, ed. (Rockville, MD: American Society of Plant Physiologists).
- ExPASy (Expert Protein Analysis System) Proteomics Server (2007). Glyceraldehyde 3-phosphate dehydrogenase active site. www.expasy.org/cgi-bin/nicedoc.pl?PS00071. Accessed October, 2015.
- Figge RM et al. (1999). Glyceraldehyde-3-phosphate dehydrogenase gene diversity in eubacteria and eukaryotes: Evidence for intra- and inter-kingdom gene transfer. *Mol Biol Evol* 16, 429–440.
- Galewsky S (2000). Sequencing cDNAs: An introduction to DNA sequence analysis in the undergraduate molecular genetics course. *Bioscene* 26, 23–25.
- Gammie AE and Erdeniz N (2004). Characterization of pathogenic human MSH2 missense mutations using yeast as a model system: A laboratory course in molecular biology. *Cell Biol Educ* 3, 31–48.
- Handelsman J et al. (2004). Scientific Teaching. *Science* 304, 521–522.
- Honts JE (2003). Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biol Educ* 2, 233–247.
- Huang X and Madan A (1999). CAP3: a DNA sequence assembly program. *Genome Res* 9, 868–877.

- Kasimova MR et al. (2006). The free NADH concentration is kept constant in plant mitochondria under different metabolic conditions. *Plant Cell* 18, 688–698.
- Kim JW and Dang CV (2005). Multifaceted role of glycolytic enzymes. *Trends in Biochem Sci* 30, 142–150.
- Kima PE and Rashe ME (2004). Sex determination using PCR. *Biochem Mol Biol Educ* 32, 115–119.
- Kresge N et al. (2005). Otto Fritz Meyerhoff and the elucidation of the glycolytic pathway. *J Biol Chem* 280, e3.
- Liaud MF (1990). Differential intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. *Proc Natl Acad Sci USA* 87, 8918–8922.
- Lissemore JL et al. (2005). Isolation of *Caenorhabditis elegans* genomic DNA and detection of deletions in the *unc-93* gene using PCR. *Biochem Mol Biol Educ* 33, 219–226.
- Lodge J et al. (2007). *Gene Cloning, Principles and Applications*. (New York: Taylor & Francis).
- López-Juez E (2007). Plastid biogenesis, between light and shadows. *J Exper Bot* 58, 11–26.
- Lyons E and Freeling M (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* 53, 661–673.
- Martin W et al. (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastic phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99, 12246–12251.
- Maxam AM and Gilbert W (1977). A new method for sequencing DNA. *Proc Natl Acad Sci* 74, 560–564.
- Metzenberg S (2007). *Working with DNA*. (New York: Taylor & Francis Group).
- Meyer-Gauen G et al. (1994). Molecular characterization of a novel, nuclear-encoded, NAD⁺-dependent glyceraldehyde-3-phosphate dehydrogenase in plastids of the gymnosperm *Pinus sylvestris* L. *Plant Mol Biol* 26, 1155–1166.
- Mychaleckyj JC (2007). Genome mapping statistics and bioinformatics. *Methods Mol Biol* 404, 461–488.
- National Research Council, Committee on Undergraduate Education to Prepare Research Scientists for the 21st Century (2003). *Bio2010: Transforming undergraduate education for future research biologist*. (Washington, DC: NRC).
- National Science Foundation (1996). *Shaping the future: New expectations for undergraduate education in science, mathematics, engineering, and technology*. (Arlington, VA: National Science Foundation).
- NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) (2006). *Enzyme Nomenclature*. www.chem.qmul.ac.uk/iubmb/enzyme/. Accessed October, 2015.
- Nichols AJ et al. (2003). Incorporating bioinformatics into the biology classroom through DNA sequence analysis. *Bioscene: Journal of College Biology Teaching* 29, 9–15.
- Olsen KW et al. (1975). Sequence variability and structure of d-glyceraldehyde-3-phosphate dehydrogenase. *J Biol Chem* 250, 9313–9321.

- PDB, Protein Data Bank. 2006. www.rcsb.org/pdb/home/home.do. Accessed October, 2015.
- Pérusse JR and Schoen DJ (2004). Molecular evolution of the GapC gene family in *Amsinckia spectabilis* populations that differ in outcrossing rate. *J Mol Evol* 59, 427–436.
- Petersen J and Cerff R (2003). Origin, evolution, and metabolic role of a novel glycolytic GAPDH enzyme recruited by land plant plastids. *J Mol Evol* 57, 16–26.
- Plaxton WC (1996). The organization and regulation of plant glycolysis. *Annu Rev Plant Physiol Plant Mol Biol* 47, 185–214.
- Primrose SB and Twyman RM (2006). *Principles of Gene Manipulation and Genomics*. (Malden, MA: Blackwell Publishing).
- Russell DA and Sachs MM (1991). The maize cytosolic glyceraldehyde-3-phosphate dehydrogenase gene family: organ-specific expression and genetic analysis. *Mol Gen Genet* 229, 219–228.
- Sambrook J and Russell DW (2001). *Molecular Cloning, A Laboratory Manual*, Volume 1. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).
- Sanger F et al. (1977). DNA Sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74, 5463–5467.
- Sansom C (2000). Database searching with DNA and protein sequences: an introduction. *Brief Bioinform* 1, 22–32.
- Schendel MS (1999). An integrated biology-chemistry freshman laboratory project in biotechnology. *Bioscene* 25, 13–19.
- Selzer PM et al. (2004). *Applied Bioinformatics: An Introduction*. (New York: Springer).
- Sirover MA (1999). New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochim Biophys Acta* 1432, 159–184.
- Skarzynksy T (1987). Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Å Resolution. *J Mol Biol* 193, 171–187.
- Song SY (1999). Structure of active site carboxymethylated d-glyceraldehyde-3-phosphate dehydrogenase from *Palinurus versicolor*. *J Mol Biol* 287, 719–725.
- Streicher H and Brodte A (2002). Introducing students to DNA. *Biochem Mol Biol Educ* 30, 104–105.
- Tatton WG (2000). Glyceraldehyde-3-phosphate dehydrogenase in neurodegeneration and apoptosis signaling. *J Neural Transm, Suppl* 60, 77–100.
- Teich R and Petersen J (2007). Intron distribution in *Plantae*: 500 million years of stasis during land plant evolution. *Gene*, 394, 96–104.
- Walsh C (1979). Enzymatic Oxidations and Reductions via Apparent Hydride Transfers: Nicotinamide Coenzymes (Chapter 10). In: *Enzymatic Reaction Mechanisms*. (San Francisco: W.H. Freeman and Company).
- Xia X (2007). *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*. (New York: Springer).

Bioinformatics Internet Resources

National Center for Biotechnology Information (NCBI) homepage: ncbi.nlm.nih.gov/

NCBI's BLAST homepage for homology searching: ncbi.nlm.nih.gov/blast/Blast.cgi

Program for sequence assembly and generation of contigs (CAP3): doua.prabi.fr/software/cap3

Alternative CAP3 Program for sequence assembly and generation of contigs:
mobyle.pasteur.fr/cgi-bin/portal.py#forms::cap3

Six-frame translation program from MolBio: molbiol.ru/eng/scripts/01_13.html

Six-frame translation program from the EMBL-EBI: ebi.ac.uk/Tools/st/

Protein databank website to view and download 3-D protein structures: rcsb.org

Legal Notices

Copyright © 2021 Bio-Rad Laboratories, Inc.

SYBR is a trademark of Thermo Fisher Scientific Inc.

BIO-RAD and MINI-SUB are trademarks of Bio-Rad Laboratories, Inc. in certain jurisdictions. All trademarks used herein are the property of their respective owner.



**Bio-Rad
Laboratories, Inc.**

Life Science
Group

Website bio-rad.com **USA** 1 800 424 6723 **Australia** 61 2 9914 2800 **Austria** 00 800 00 24 67 23 **Belgium** 00 800 00 24 67 23 **Brazil** 4003 0399
Canada 1 905 364 3435 **China** 86 21 6169 8500 **Czech Republic** 00 800 00 24 67 23 **Denmark** 00 800 00 24 67 23 **Finland** 00 800 00 24 67 23
France 00 800 00 24 67 23 **Germany** 00 800 00 24 67 23 **Hong Kong** 852 2789 3300 **Hungary** 00 800 00 24 67 23 **India** 91 124 4029300
Israel 0 3 9636050 **Italy** 00 800 00 24 67 23 **Japan** 81 3 6361 7000 **Korea** 82 080 007 7373 **Luxembourg** 00 800 00 24 67 23
Mexico 52 555 488 7670 **The Netherlands** 00 800 00 24 67 23 **New Zealand** 64 9 415 2280 **Norway** 00 800 00 24 67 23 **Poland** 00 800 00 24 67 23
Portugal 00 800 00 24 67 23 **Russian Federation** 00 800 00 24 67 23 **Singapore** 65 6415 3188 **South Africa** 00 800 00 24 67 23
Spain 00 800 00 24 67 23 **Sweden** 00 800 00 24 67 23 **Switzerland** 00 800 00 24 67 23 **Taiwan** 886 2 2578 7189 **Thailand** 66 2 651 8311
United Arab Emirates 36 1 459 6150 **United Kingdom** 00 800 00 24 67 23

