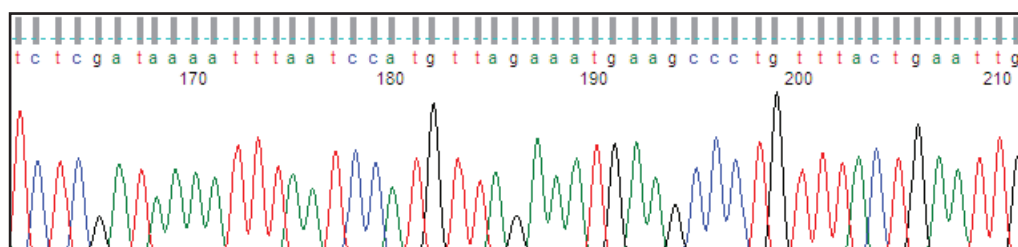

Sequencing and Bioinformatics Module

Instruction Manual

Catalog #166-5025EDU

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 Biotechnology Explorer kit curricula.

This kit is shipped on blue ice and contains temperature sensitive reagents. Open immediately upon arrival and store reagent bag at -20°C .



BIO-RAD

Dear Educator,

The Genomics Revolution. With the advent of automated sequencing technologies, DNA sequencing has revolutionized biological research. Now we can contemplate elucidating the root genetic causes of diseases, specific mutations that permit viruses to escape therapies, and even genes responsible for traits of dog breeds — such as the genes that give the Rhodesian ridgebacks their ridges! However, this kind of research generates enormous data sets. For example, the dog genome is 2.5 billion base pairs long. Manually, with pen and paper, it's extremely difficult to organize or analyze that amount of data.

The science of bioinformatics developed in response to the needs of scientists, as a way first to store and organize data and then to analyze it. After all, base calls of genomes aren't of much use if you can't get any information from them. Some of the many questions that can be answered through bioinformatics include comparing gene sequences between different species to determine evolutionary relationships, identifying genes by comparing them to genes in other organisms, predicting protein structure, and finding mutations in genes that might contribute to disease.

This kit guides students through DNA sequencing and subsequent data analysis by a local sequencing facility. The class' DNA can be sequenced by a local sequencing facility or by Eurofins MWG/Operon which offers reduced pricing for educators. Thanks to a collaboration with Biomatters, Inc., students have access to Geneious, a bioinformatics software platform through which they can store, organize, and assess their data. The investigation is open-ended and the depth of analysis is left to the instructor's discretion.

The Sequencing and Bioinformatics module is part of the Bio-Rad Cloning and Sequencing Explorer Series (catalog #166-5000EDU) and can also be used as a stand-alone kit to sequence DNA and explore bioinformatics. The module includes DNA that can be used either to obtain sequences for analysis or as a control for sequencing along with your own DNA samples.

This curriculum was developed in collaboration with Dr Sandra Porter from Digital Biology World and Dr Kristi DeCourcy of the Fralin Biotechnology Center at Virginia Tech. We thank Drs Porter and Decourcy for their invaluable guidance and contributions to this curriculum.

The Biotechnology Explorer Team
Life Science Group
Bio-Rad Laboratories
6000 James Watson Dr.
Hercules, CA 94547

Create context. Reinforce learning. Stay current.

New scientific discoveries and technologies create more content for you to teach, but not more time. Biotechnology Explorer kits help you teach more effectively by integrating multiple core content subjects into a single lab. Connect concepts with techniques and put them into context with real-world scenarios.

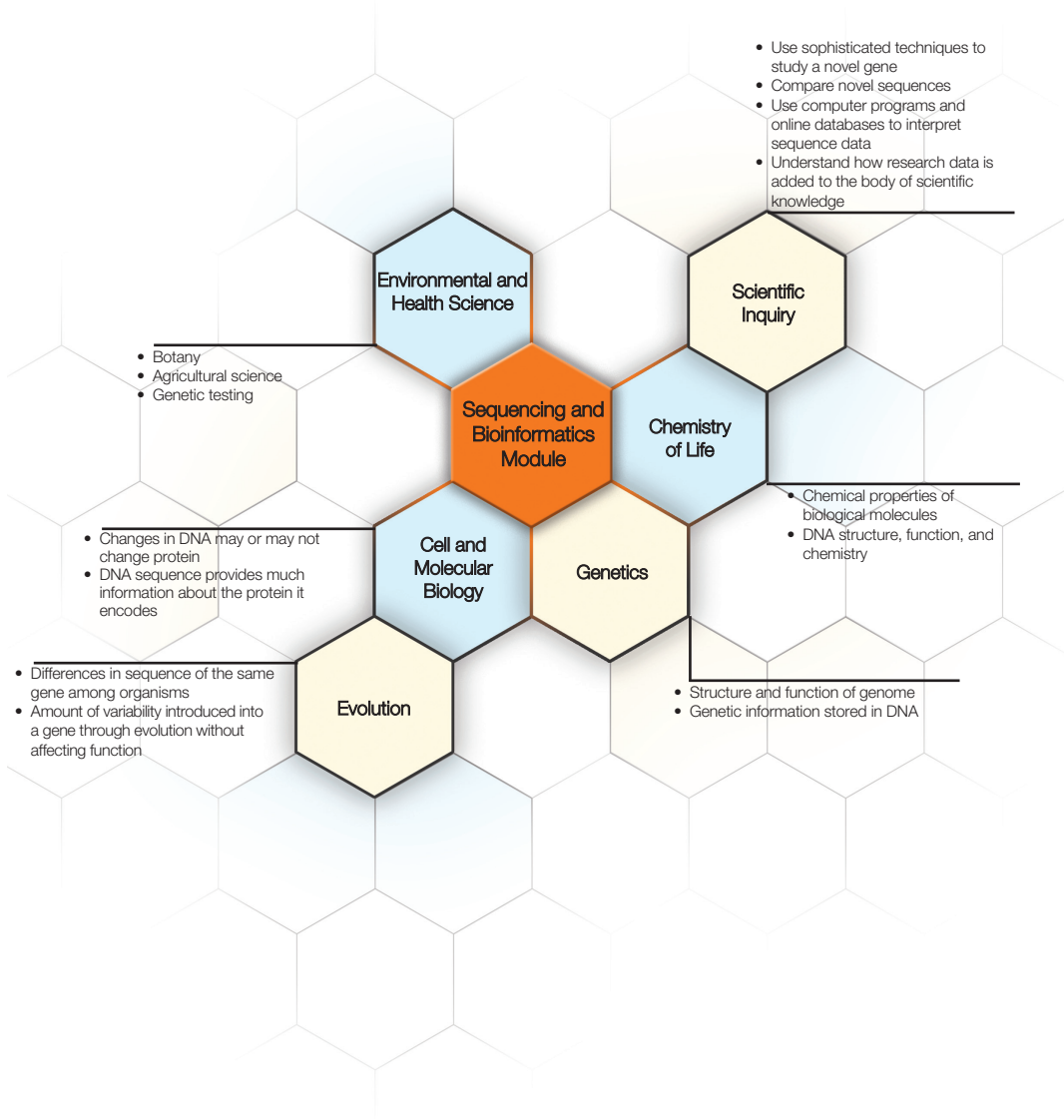


Table of Contents

	Page
Introduction	1
Lesson Time Line	3
Curriculum Fit	3
Kit Inventory Checklist	4
DNA Sequencing	5
Background for Instructor.....	5
Instructor's Advance Preparation	9
DNA Sequencing – Quick Guide.....	12
Student Protocol	13
Bioinformatics	18
Background for Instructor.....	18
Instructor's Advance Preparation	23
Protocol.....	33
A Tour of the Geneious Platform.....	34
1. View Sequence Traces and Review the Quality of the Sequencing Data	36
2. Determine Sequence Identity Using BLAST	60
3. Assemble the Sequences and Correct Mistakes in the Base Calls	75
4. Conduct a BLAST Search on the Contig Sequence to Verify Identity of the Cloned Gene.	89
5. Determine Gene Structure (Intron/Exon Boundaries) Using BLAST — Build a Gene Model.....	97
6. Predict an Amino Acid Sequence from the Cloned Gene (blastx).....	120
Appendix A: GenBank — Searching and Submitting Sequences	129
Appendix B: Instructor Answer Guide	132
Appendix C: Glossary	134
Appendix D: References	136

Introduction

DNA sequencing has become an integral part of the biological sciences. Many organisms have had their entire genomes sequenced, generating enormous data sets. For example, the human genome has approximately 3 billion bases and 20,000–25,000 genes. As our abilities to generate large datasets of biological information have expanded, tools to store, organize, and analyze data have become increasingly important, leading to the development of the field of bioinformatics.

In this module, students will sequence DNA and then use computational programs to store and evaluate the quality of the data, look for similar DNA sequences in DNA databases, construct a contiguous sequence from the class' sequences, and more.

The Sequencing and Bioinformatics module is part of Bio-Rad's Cloning and Sequencing Explorer Series. The Cloning and Sequencing Explorer Series is a sequence of individual modules that have been designed to work in concert to give students the real world experience of a molecular biology research workflow. Below is a typical workflow for cloning and sequencing a gene. The Sequencing and Bioinformatics module can sequence plasmids generated using the Ligation and Transformation module, or the entire Cloning and Sequencing Explorer Series, as well as independently generated plasmids.

The specific modules of the Cloning and Sequencing Explorer Series that are available to perform the additional steps are indicated below and can be purchased separately. Further information on the separate modules is available in the Biotechnology Explorer™ catalog or from explorer.bio-rad.com.

- Isolate genomic DNA from plant of choice¹
- Amplify gene of interest using PCR²
- Purify PCR product³
- Ligation of PCR product into pJet1.2 plasmid⁴
- Transform ligated plasmid into bacteria⁴
- Culture bacteria and grow minipreps⁵
- Purify plasmid from minipreps⁶
- Analyze plasmid by restriction digestion⁴
- Electrophorese restriction digest reaction⁷
- **Sequence plasmid and analyze sequence data**

¹ Nucleic Acid Extraction module (catalog #166-5005EDU) extracts genomic DNA from plant tissues.

² *GAPDH* PCR module (catalog #166-5010EDU) amplifies a fragment of the *GAPDH* gene from a preparation of plant genomic DNA.

³ PCR Kleen™ Spin module (catalog #732-6300EDU) purifies 25 PCR products.

⁴ Ligation and Transformation module (catalog #166-5015EDU) contains reagents required to ligate PCR products into pJet1.2 plasmid vector, transform ligations, and analyze resultant plasmids by restriction digest analysis.

⁵ Microbial Culturing module (catalog #166-5020EDU) contains all required reagents for culturing bacteria for ligating and transforming bacteria using the Ligation and Transformation module.

⁶ Aurum Plasmid Mini Purification module (catalog #732-6400EDU) contains reagents to purify plasmid DNA from 100 minipreps.

⁷ Electrophoresis modules contain reagents to analyze plasmid restriction digests.

This module is designed for 96 sequencing reactions. This module also includes thorough bioinformatics exercises, including a three month Geneious account in which the sequencing data can be stored, accessed, and analyzed by your students.

The module includes a pGAP control plasmid, which is the pJet1.2 vector containing a partial *Arabidopsis GAPC* genomic sequence. The primers provided in the kit are a forward and a reverse primer—pJET SEQ F and pJET SEQ R respectively—that anneal to either side of the vector and permit sequencing of the control plasmid or any PCR product cloned into the pJet1.2 cloning site using the Ligation and Transformation module. In addition, two primers that are specific for *GAPC* sequences in the control plasmid are included, allowing four separate sequences from the control plasmid to be generated and assembled together.

The Sequencing and Bioinformatics module can be adapted for the following multiple learning objectives:

- Introduce students to sequencing and bioinformatics by generating sequences from the combination of the control plasmid with the sequencing primers. Upon receipt of the data from a sequencing facility, students can analyze the sequences using the Geneious account to learn about BLAST, generate contigs, determine gene structure, and predict mRNA sequences. Please note: the Cloning and Sequencing Explorer Series instruction manual contains additional background information on *GAPC* and additional bioinformatics activities that can enrich your lessons
- Sequence PCR products by ligating them into the pJet1.2 plasmid vector using the Ligation and Transformation module and purifying plasmid from resultant miniprep clones using the Aurum Plasmid Mini Purification and Microbial Culturing modules. Then perform bioinformatics research on your novel data using Geneious and associated tools and submit novel sequences to GenBank
- Sequence PCR products generated when using the entire Cloning and Sequencing Explorer Series. When using the entire series please be sure to use the Cloning and Sequencing Explorer Series instruction manual
- Sequence independently generated plasmids using your own sequencing primers, perform bioinformatics research using Geneious and associated tools, and submit novel sequences to GenBank

Lesson Time Line

Pre-lab activity	Setting the stage: DNA structure, DNA sequencing, and genome sequencing	1–2 days, lecture and homework
Lesson 1	Set up sequencing reactions and transfer to 96-well plate	45 minutes in lab
	Sequencing results available	Up to 2–3 weeks depending on sequencing facility
Lesson 2	Bioinformatics	Basic analyses take from 5–8 hours*

* The time needed to complete the bioinformatics module will vary, depending on the extent of analyses performed and student experience.

Storage Instructions

Open the kit as soon as it arrives and remove the bag of perishable components. Store these components in the freezer at -20°C . Save the shipping container and blue ice pack for shipping to the sequencing facility.

Intended Audience

This kit is appropriate for students with some experience in molecular biology. It would fit in a molecular biology or biotechnology curriculum at the high school level or at the two- or four-year college level. The background of the students doing the experiment can be used to determine the appropriate depth of the bioinformatics analysis.

Curriculum Fit

- Students use advanced technology to solve a novel problem
- Students develop an understanding of the current techniques in biological computing
- Students engage in a scientific enterprise where they contribute to the scientific body of knowledge
- Students develop an understanding of biological evolution as they see how genes can change over time
- Students develop abilities to conduct inquiry-based experiments

Kit Inventory Checklist

This section lists the equipment and reagents needed for setting up sequencing reactions of genes in the pJet1.2 plasmid in your classroom or teaching laboratory. The Sequencing and Bioinformatics module (catalog #166-5025EDU) supports 12 student workstations, with 2–4 students per station. Open the kit as soon as it arrives and place the bag of perishable components in the freezer (–20°C). Save the shipping containers and blue ice pack for shipping to the sequencing facility.

Kit Components	Number/Kit	(✓)
pJET SEQ F, forward sequencing primer, 50 µl	1	<input type="checkbox"/>
pJET SEQ R, reverse sequencing primer, 50 µl	1	<input type="checkbox"/>
GAP SEQ F, GAPC forward sequencing primer, 50 µl	1	<input type="checkbox"/>
GAP SEQ R, GAPC reverse sequencing primer, 50 µl	1	<input type="checkbox"/>
pGAP control plasmid, 20 ng/µl, 100 µl	1	<input type="checkbox"/>
Bar-coded 96-well plate	1	<input type="checkbox"/>
Sealing film	1	<input type="checkbox"/>
Colored micro test tubes, 2.0 ml	120	<input type="checkbox"/>
Foam box	1	<input type="checkbox"/>
Ice pack	1	<input type="checkbox"/>
Instruction Manual	1	<input type="checkbox"/>

Required Accessories (Not Provided in Kit)	Number/Kit	(✓)
Adjustable volume micropipets, 0.5–10 µl (catalog #166-0550EDU or 166-0505EDU)	1–12	<input type="checkbox"/>
Pipet tips, 0.5–10 µl (catalog #223-9354EDU)	1–12	<input type="checkbox"/>
Computer and Internet access	1–12	<input type="checkbox"/>

Optional Accessories	Number/Kit	(✓)
Additional sequencing primers	1–2	<input type="checkbox"/>
Ligation and Transformation module (catalog #166-5015EDU)	1	<input type="checkbox"/>
Aurum™ Plasmid Mini Purification module (catalog #732-6400EDU)	1	<input type="checkbox"/>
Microbial Culturing module (catalog #166-5020EDU)	1	<input type="checkbox"/>
EZ Load™ Precision Molecular Mass Ruler (catalog #170-8356EDU)	1	<input type="checkbox"/>

Refills Available Separately

Sequencing Kit Refill Pack (catalog #166-5026EDU), includes pJET SEQ F primer, pJET SEQ R primer, GAP SEQ F primer, GAP SEQ R primer, and pGAP control plasmid.

DNA Sequencing

Background

The Development of DNA Sequencing

Sequencing DNA means to determine the exact order of nucleotides (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 the same time. Although we take DNA sequencing for granted now, when researchers started sequencing DNA in the 1970s 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 lab facilities where DNA is sequenced for them using an automated sequencer. The researchers can 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 and other fungi, plants, fruit flies, mosquitoes, zebrafish, and mammals such as human, mouse, rat, opossum, chimpanzee, and dog (a boxer named Tasha). Many more genomes are currently undergoing sequencing.

Maxam-Gilbert Sequencing Method

Maxam and Gilbert, working in the United States, developed a chemical degradation method for DNA sequencing. The steps in sequencing by chemical degradation are:

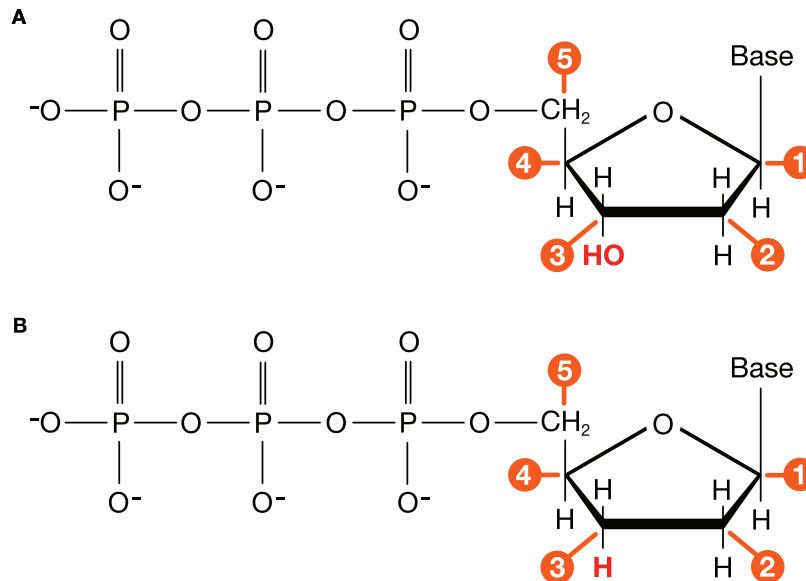
1. Label the 5' end of the DNA to be sequenced with a radioisotope tag. (The labeling can also be performed at the 3' end, but the end selected for labeling must be consistent.)
2. Divide the labeled DNA into four test tubes, each containing different chemicals that cleave the DNA strands after a particular base. The reaction is relatively inefficient and cleaves the DNA at a small percentage of the occurrences of the bases — not at every one. This produces "a nested set of radioactive fragments extending from the labeled end to each of the positions of that base" (Maxam & Gilbert, 1977).
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. 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 on the gel are one base different in size.

The Maxam-Gilbert method could sequence about 100 bases into a DNA fragment, but it used very hazardous chemicals. The method was also 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 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.

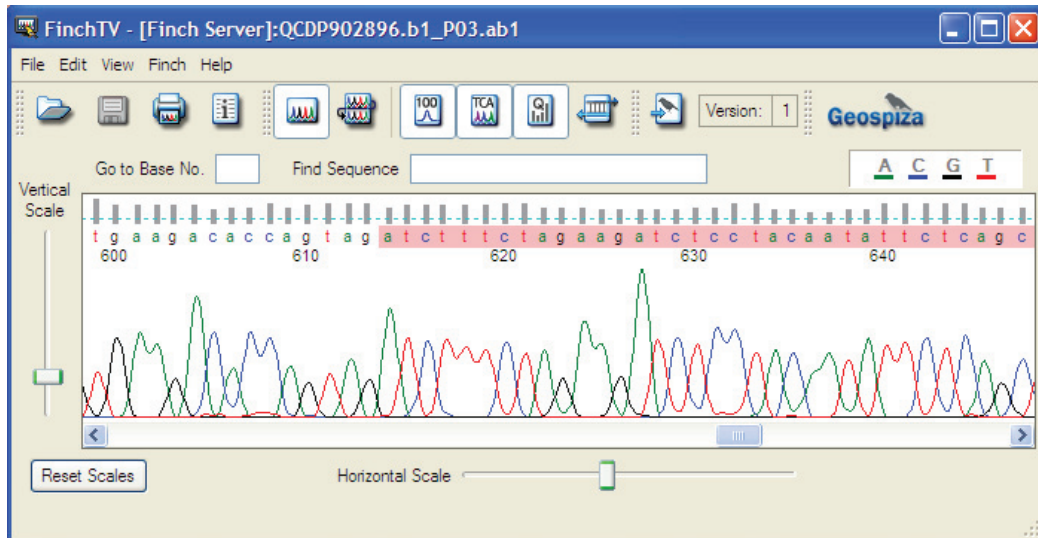
Using the Sanger method with dNTPs labeled with radioisotopes, it is possible to read up to several hundred bases from an autoradiograph, but the process is both time and labor intensive. It was now possible to sequence DNA in the research lab, using either Maxam-Gilbert or Sanger methods, but both methods 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 green 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 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 µl 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.

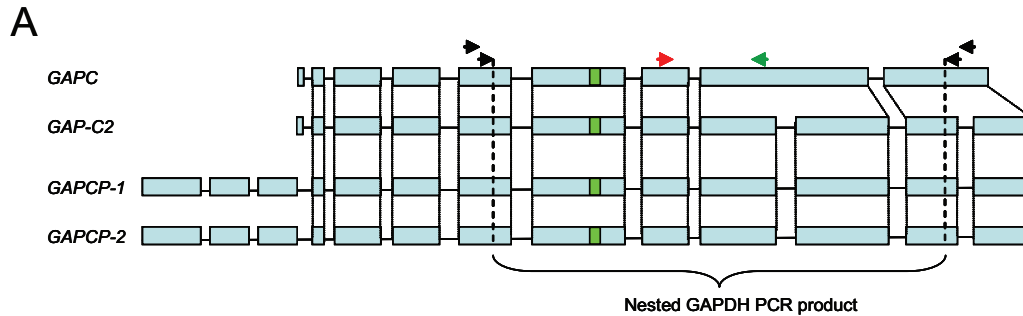
Sequencing Primers

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

- pJET SEQ F, forward sequencing primer, 200 µM (blue)
sequence: CGACTCACTATAGGGAGAGCGGC
- pJET SEQ R, reverse sequencing primer, 200 µM (yellow)
sequence: AAGAACATCGATTTTCCATGGCAG
- GAP SEQ F, *GAPC* forward sequencing primer, 200 µM (red)
sequence: GGHATTGTTGAGGGTCTNATGAC
- GAP SEQ R, *GAPC* reverse sequencing primer, 200 µM (green)
sequence: CCA GTGGTGCTRGGAATGATGTT

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' bioinformatics Geneious 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 is 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 are 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 Biomatters Inc., Geneious account.
4. Retain the foam shipping box for shipping the samples to the sequencing facility.
4. Place the ice pack in the freezer until ready for shipping.
5. Activate your Geneious 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 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

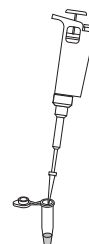
1. Label microcentrifuge tubes with the well numbers your instructor has assigned for your samples.



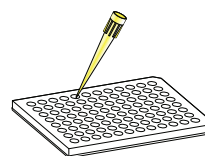
2. Fill in the following table for each sample.

Well Identifier	DNA Sample Name	Sequencing Primer

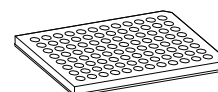
3. Combine 10 μ l of DNA sample with 1 μ l of the appropriate sequencing primer in a microcentrifuge tube. Pipet up and down to mix.



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 web-based bioinformatics tool called Geneious. 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.

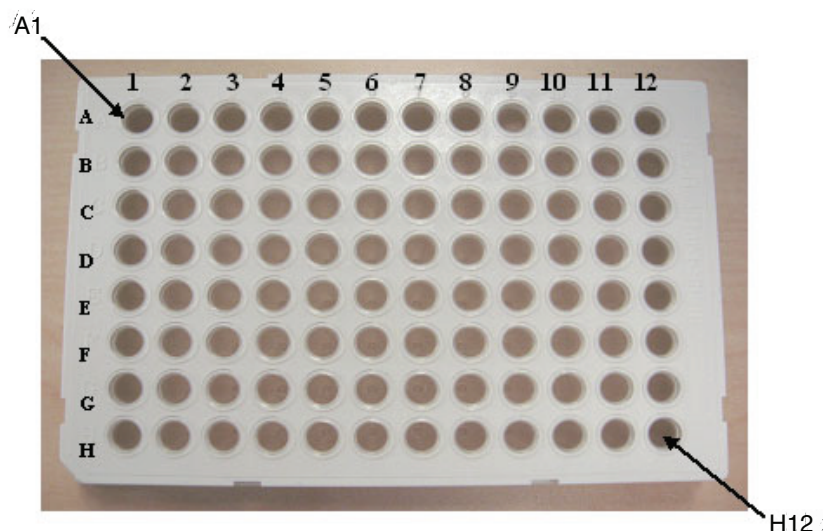
Materials Needed for Each Workstation	Quantity	(✓)
DNA sample cloned into the pJET1.2 plasmid	2	<input type="checkbox"/>
Colored microcentrifuge tubes, 2.0 ml	4	<input type="checkbox"/>
10 µl adjustable-volume micropipet and tips	1	<input type="checkbox"/>

Common Workstation

Material Required	Quantity	(✓)
Forward sequencing primer (pJET SEQ F)	1	<input type="checkbox"/>
Reverse sequencing primer (pJET SEQ R)	1	<input type="checkbox"/>
Bar-coded 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

1. 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 μ l of DNA sample with 1 μ l of sequencing primer. Pipet up and down to mix.
3. Pipet 10 μ l 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 their 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 during transit.

If sequencing with Eurofins MWG/Operon, go to www.operon.com/bio-rad. A pdf with instructions on how to submit samples on their website can be requested from Eurofins MWG/Operon.

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 _____

12								
11								
10								
9								
8								
7								
6								
5								
4								
3								
2								
1								
	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?

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*. Speed and accuracy of gene isolation and sequencing has grown quickly. From 1982, when GenBank was at Release 3, to Release 197 in August of 2013, the number of nucleotide bases in GenBank doubled every 18 months. Release 3 contained only 606 sequences while Release 197 contains more than 167 million 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: a bioinformatics software platform from Biomatters, Inc. that offers data management, 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. 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.

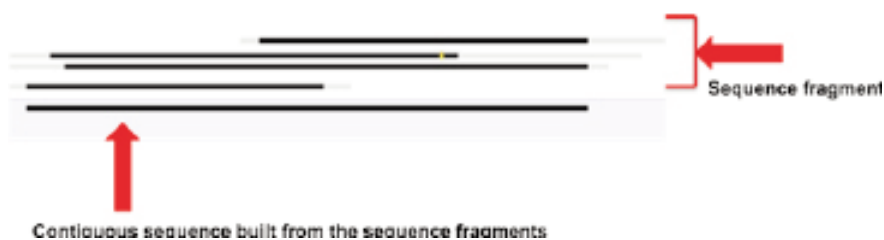


Example of a chromatogram viewed in Geneious.

Geneious Software Tools for Sequence Trimming and Assembly

Geneious 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 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 together 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 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 using Geneious. Individual shorter sequences are compared, aligned, and assembled by programs such as Geneious 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).

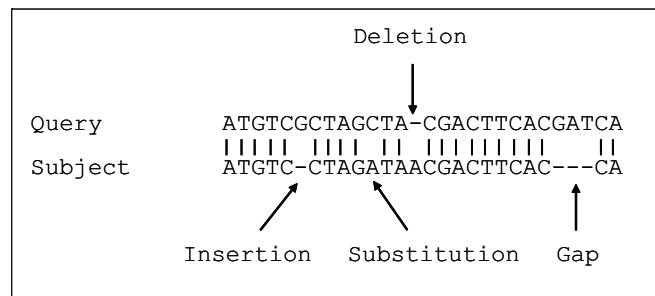
The BLAST search that is used usually depends on the type of sequence data that has been generated. For example, a *blastx* search translates a nucleotide sequence into predicted amino acid sequences and compares these to a database of protein sequences. If a protein sequence is the only information 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 the 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 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 software and the NCBI’s BLAST website directly.

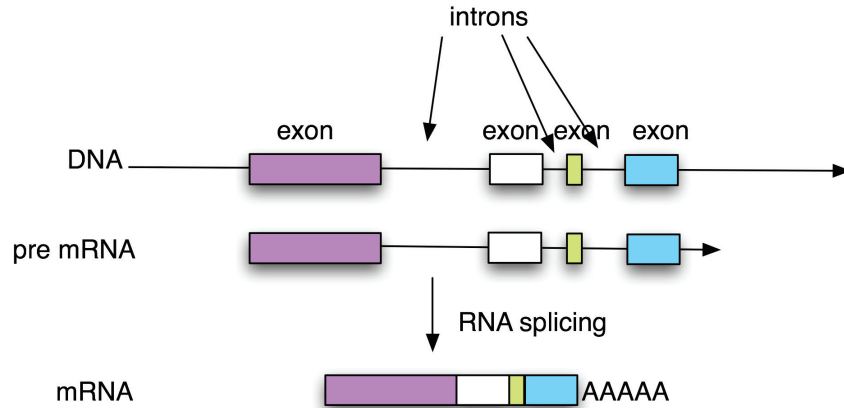


BLAST alignment. BLAST programs compare a user-entered (query) sequence with subject sequences in a database. It scores the match depending on the sequence identity and the number of differences — deletions, insertions, substitutions, and gaps — between the sequences.

Predicting an mRNA Sequence

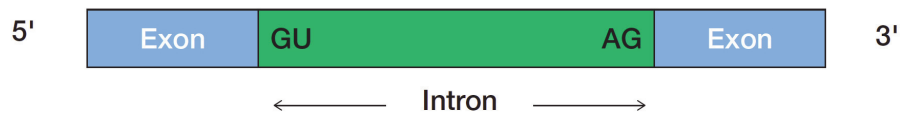
Most eukaryotic genes contain some sequence that does not code for protein. During gene transcription, RNA splicing removes these sequences, known as introns, and fuses (or splices) the sequences that contain coding information (exons) together at what are known as splice sites to form the mRNA sequence. 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 from the reference databases that have been validated by NCBI staff in GenBank.

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, 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 MWG/Operon 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 to look at the quality of individual reads and the class's data as a whole set.
2. Use BLAST (blastn) for a preliminary determination of which GAPDH genes most closely represent the gene that was cloned.
3. Assemble sequences into a contig and correct sequencing errors with Geneious.
4. Verify which GAPDH gene was cloned by conducting a blastn search on the contig sequence against the GenBank genomic sequence database.
5. Annotate a gene by conducting a blastn search against the GenBank mRNA sequence database to predict gene structure (that is, intron/exon boundaries) and mRNA sequence.
6. 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 how to submit the class sequence information to the GenBank database and the GenBank's sequence submission policy.

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 (cbroccoli folder).

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. Web browser software — students should be able to:
 - Enter a URL in the address window to go to a website
 - Use the back button or the history menu to return to a site
 - Open and use multiple windows
 - Refresh a page
 - Copy text from a webpage and paste it into another page
 - Download files from a webpage and locate them afterward
 - Use the Find command to locate a term on a webpage
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.
2. Download Geneious installer and install software on student computers (approx. 1 hr, depending on the number of computers).
3. Activate Geneious software using the license key provided in the kit.
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 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 sample data (6–10 hr).

1. Check for minimum requirements for computers (approx. 30 min).

- 1.1** For Geneious version 8.1.5, check to make sure that computers have one of the following operating system versions before installing Geneious:

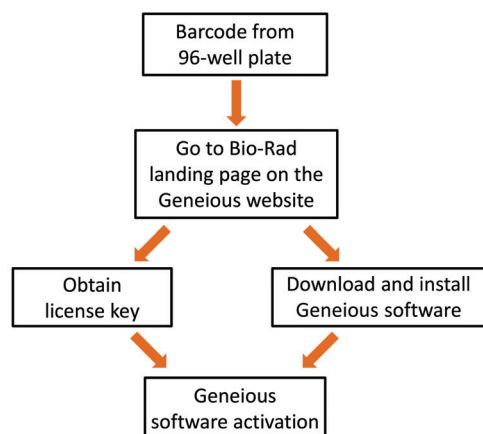
Operating System	System requirements
Windows	XP / Vista / 7 / 8
Mac OS	10.6 / 10.7 / 10.8 / 10.9
Linux	CentOS 6 / RHEL6 / Ubuntu Desktop LTS

- 1.2** It is also recommended that computers have at least the following specifications for running Geneious:

- 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
- **Java 1.6 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**
 - o To check your Java version, click the Apple icon on the upper left in the menu bar > System Preferences. Click the Java icon to open the Java Control Panel. Click **General > About** to find the build version



2. Obtain the Geneious software license key, download the Geneious installer and install software onto student computers (approx. 1 hr).

Download the Geneious software account 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 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 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 the Bio-Rad landing page on the Geneious website. In your web browser, navigate to <http://www.geneious.com/biorad>. Follow the instructions on the webpage to request your license key from Geneious, and download **version 8.1.5** of Geneious.

2.3 **DO NOT** download a newer version of Geneious.

IMPORTANT NOTE: This instruction manual is tailored to Geneious software version 8.1.5. Geneious may prompt you to download a different version. The license key issued to you by Geneious may not work, and the screenshots in this manual may not match versions other than 8.1.5.

3. Activate your Geneious license on student computers.

3.1 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.

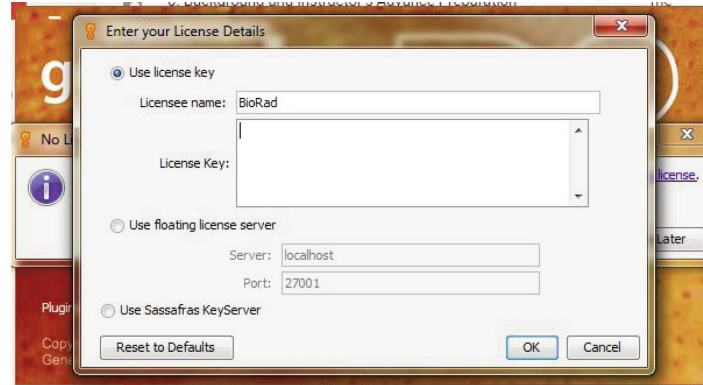
3.1.1 If this is the first time you are installing Geneious onto student computers:

Open the Geneious program. You should again see the dialog box asking about a license. Click Activate a License. A new dialog box will appear.

3.1.1 If this is the **first time** you are installing Geneious onto student computers: Open the Geneious program. You should again see the dialog box asking about a license. Click **Activate a License**. A new dialog box will appear.

3.1.1.1 Select **Use license key**.

3.1.1.2 Enter **BioRad** as the Licensee name.



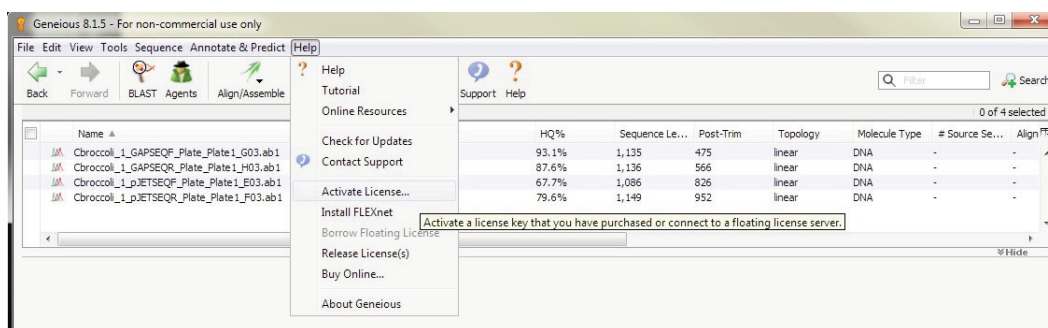
3.1.1.3 Copy the license key from the text file that was emailed to you and paste it into the License Key field. Click **OK**.

A new dialog box will appear to indicate that your registration was successful. You will now have full, unrestricted access to the Geneious software.



3.2 If you have previously installed Geneious onto student computers:

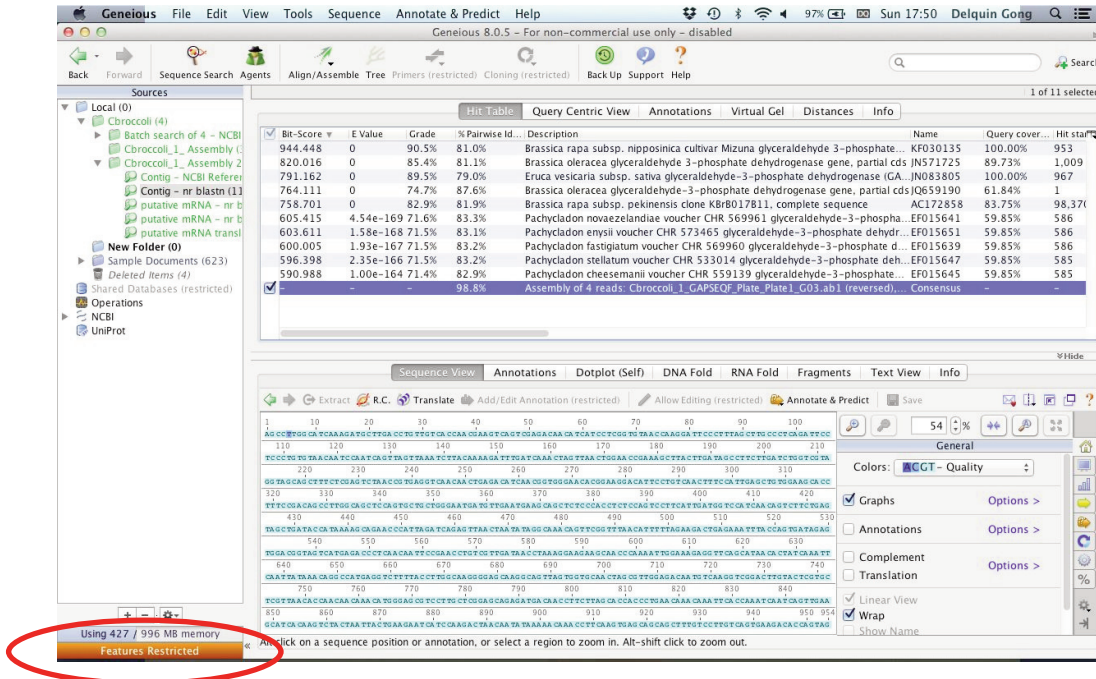
3.2.1 Open the Geneious program. In the program toolbar, click **Help**, then **Activate License**.



3.2.2. A new dialog box will open asking for your license details. Follow the directions from 3.2.1.

3.3 Check to make sure that the licenses are working.

Open the Geneious program and look on the bottom left corner of the Geneious 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 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 for help via phone or email.

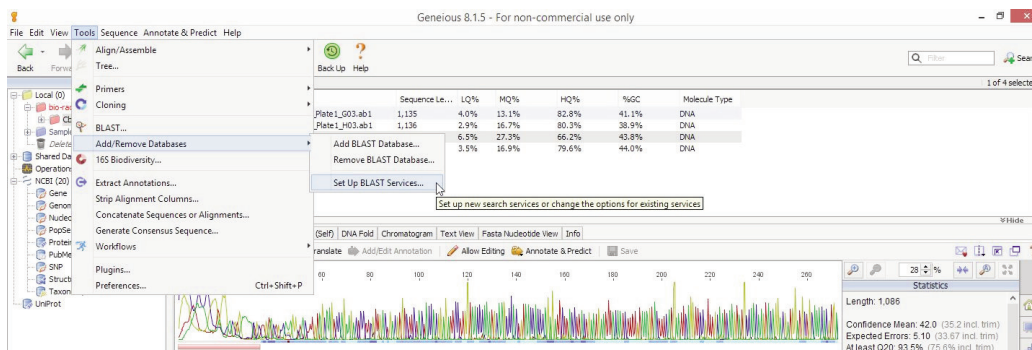
4. Decide how to distribute the class's sequencing data (approx. 30 min).

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 of the individual classroom computers. 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 through Geneious, BLAST services will need to be set up within Geneious. Follow the directions below to set this up for each computer:

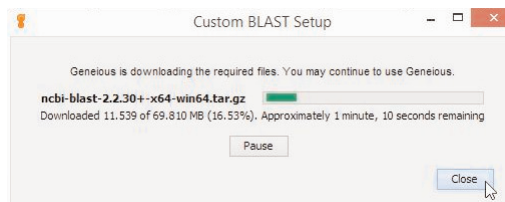
5.1 Open Geneious on your computer. In the main toolbar of the Geneious window, click the Tools tab, select **Add/Remove Databases**, then select **Set Up Search Services**.



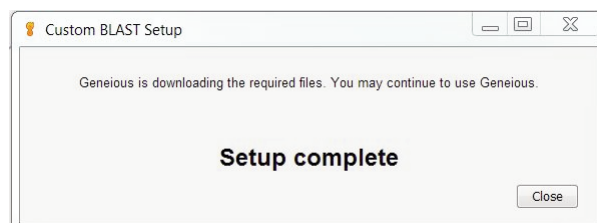
- 5.2 A new dialog box will appear. For Service, choose **Custom BLAST** from the dropdown menu and check the box to let Geneious do the setup for you.



- 5.3 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).



- 5.4 Geneious will let you know when the BLAST search service setup is complete. If the window below does not appear, then the BLAST search service setup was incomplete. Please see Appendix G for troubleshooting help.



- 5.5 Perform steps 4.1–4.4 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:

In general, the amount of time it takes to retrieve BLAST results will vary depending on how many searches NCBI BLAST is being asked to run at that moment from researchers around the world. In some cases, searches performed through Geneious are **not as fast** as performing the BLAST searches directly from NCBI.

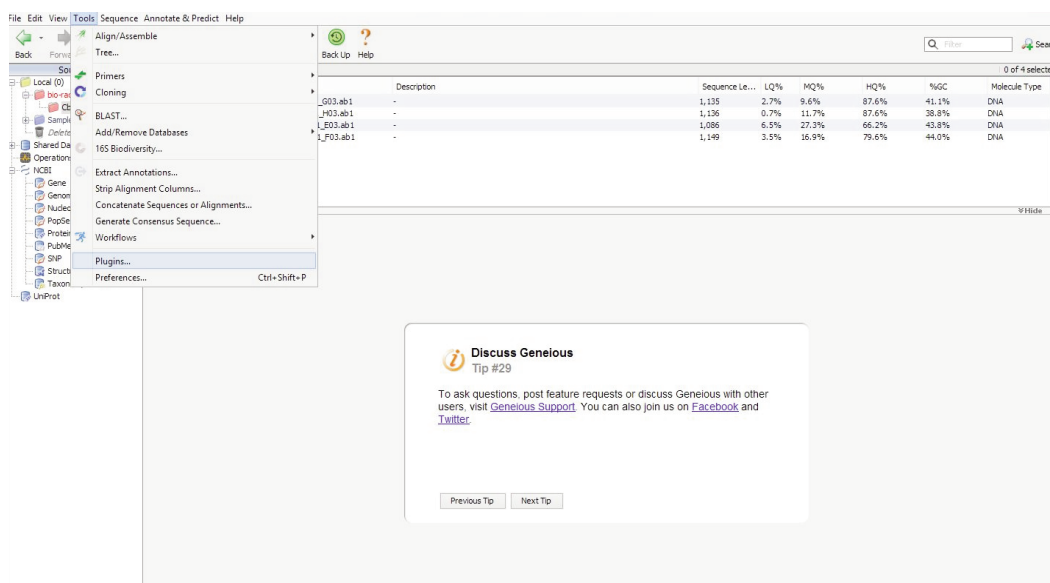
If you have short class periods (50 min or less) and need the BLAST results as soon as possible to enable data analysis, you have the option of having your students **perform the BLAST searches directly from the NCBI website for Sections 2 and 4**. Please refer to Appendix I for protocol steps to export sequences as FASTA files for BLAST searching directly on the NCBI website.

6. Enable the Fasta View custom feature in the Geneious program.

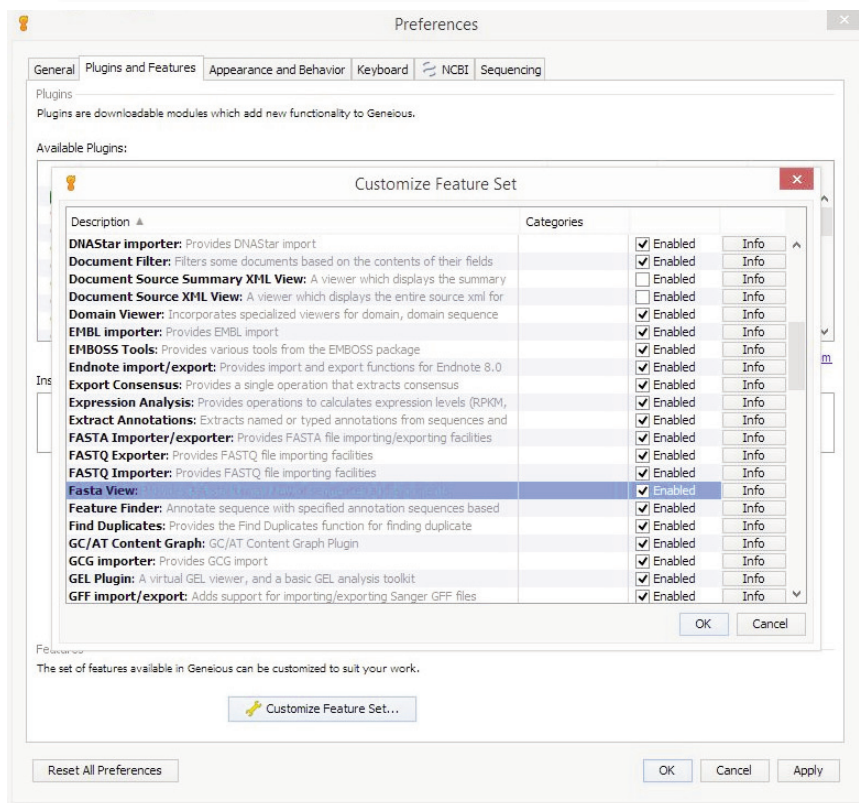
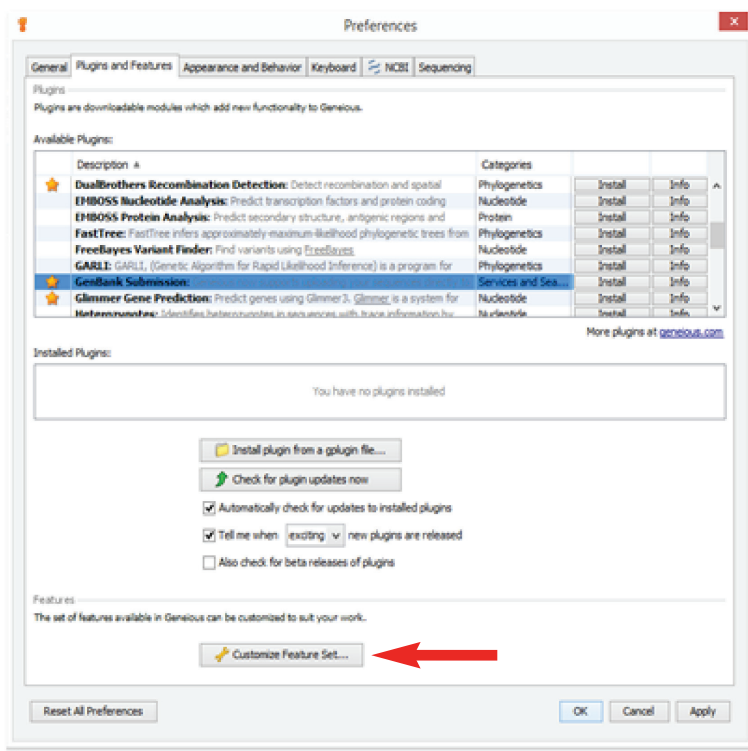
FASTA is a text-based format for representing sequences such that nucleotides (or amino acids) are denoted using single-letter codes. This format also allows for a single line of description for the sequence name and comments about the sequence itself. FASTA is a widely used sequence format in the field 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 will facilitate your GenBank submission process without requiring you to manually export sequence files.

- 6.1 Open the Geneious program. Click the Tools tab in the main Geneious window and select **Plugins**. A new dialog box will appear, named Preferences.



- 6.2 Scroll to the bottom of the window and click **Customize Feature Set**. A Customize Feature Set window will open.



- 6.3** Feature names appear in alphabetical order in the left-hand column. 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.

Please note that 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 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. The first step in analysis is to upload to your computer the raw DNA sequences returned by the sequencing facility so you can import it into the Geneious 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:

- Use Geneious to look at the quality of individual reads
- Use BLAST (blastn) for a preliminary determination of which GAPDH gene was cloned
- Assemble sequences into a contig and finish sequence by correcting sequencing errors with Geneious
- Verify which GAPDH gene was cloned using BLAST (blastn) on the contig sequence against the GenBank 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)

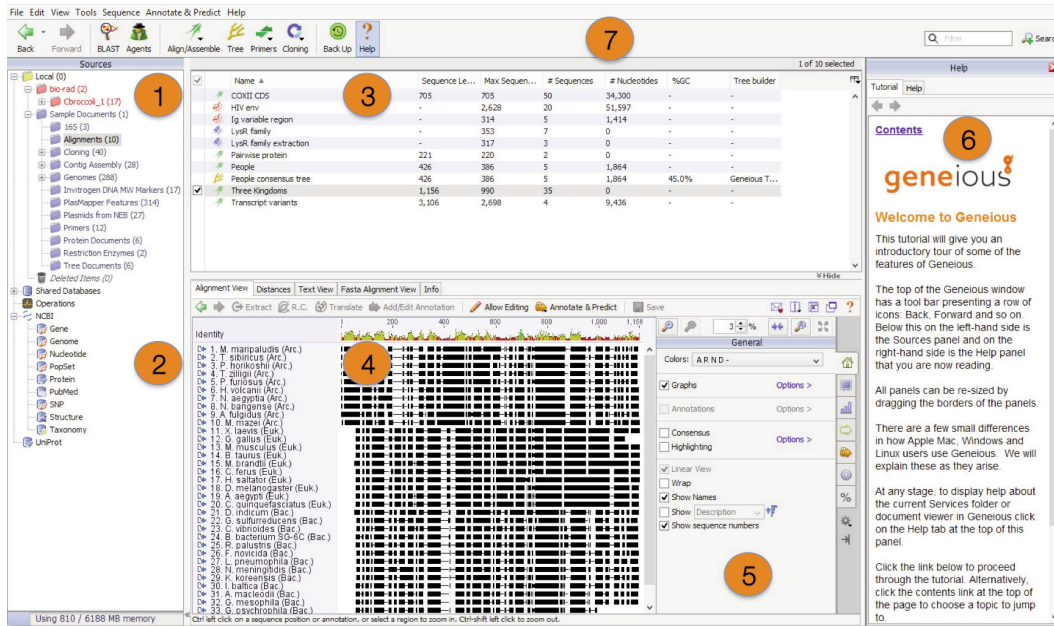
Assembling sequences from the entire class is an optional additional step. After analysis, the gene sequences can be submitted to the GenBank database (see Appendix C).

Materials Required

- Computer with Internet access
- DNA sequence files imported into Geneious software

A Tour of the Geneious Platform

Here is a quick orientation to the Geneious 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.

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 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 help displayed in the help tab changes as you click on different services and choose different viewers. The Tutorial is aimed at first-time users of Geneious and has been included to provide a feel for how Geneious works. It is highly recommended that you work through the tutorial if you haven't used Geneious before. The Help panel can be closed at any time by clicking the red X in its top right corner or by toggling the

Help button  in the menu bar.

7 — Menu Bar

The menu bar contains several icons that provide shortcuts to common functions in Geneious, including BLAST; Agents, which search databases for new content even while you sleep; Align/Assemble; Tree building; Back Up; and Help. You can alter the contents of the menu bar to suit your own needs. The icons can be displayed small or large and with or without their labels. The Help icon  is always available. The Back and Forward options help you move between previous and subsequent views in Geneious and are analogous to the back and forward buttons in a web browser. The Back Up feature  is particularly useful and helps save your data by downloading it to your local hard drive.

For more useful Tips and Tricks on personalizing and navigating the Geneious software, please see Appendix H.

1. View Sequence Traces and Review the Quality of the Sequencing Data

1.1 Using the Geneious software for bioinformatics

In this part of the lab, you will use many of the features of the Geneious software to manually review your sequencing results. When chromatogram files are uploaded to Geneious, several programs act in sequence to analyze and process these data. You will explore some of these data.

1.1.1 Extracting sequences and assessing their quality

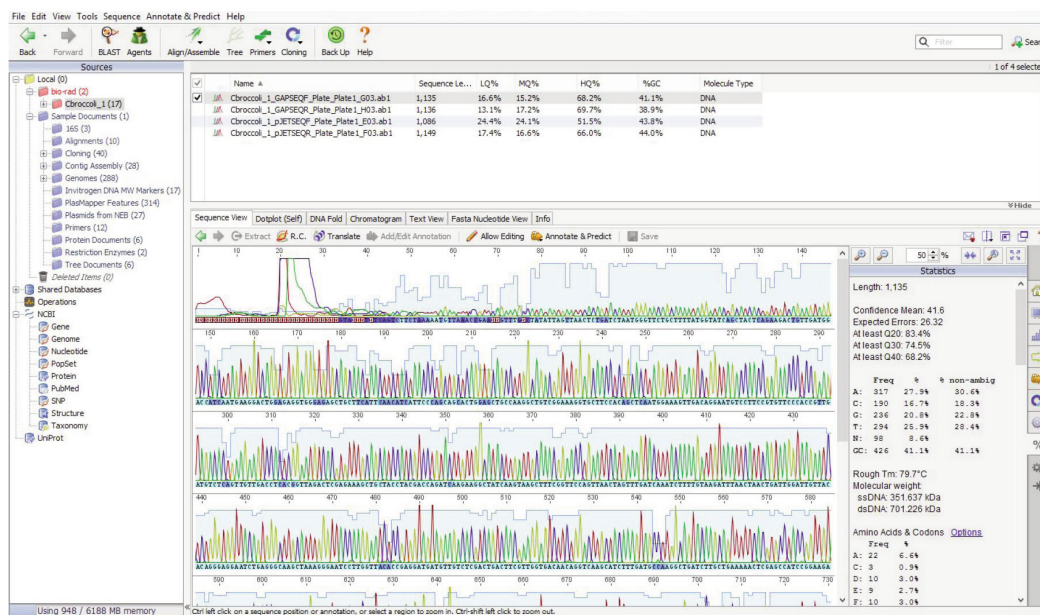
Upon import of .ab1 files, Geneious will read your data files automatically and will display elements like quality scores and DNA sequence information from the file containing the sequencing chromatogram. This chromatogram was generated by the sequencing software after the DNA was sequenced.

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 1/10⁶ chance of a wrong call, 50 a 1/10⁵ chance, 40 a 1/10⁴ 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. 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.

1.1.2 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 the bases at each end of the read have quality scores 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, the trimmed regions are indicated by a light pink annotation.

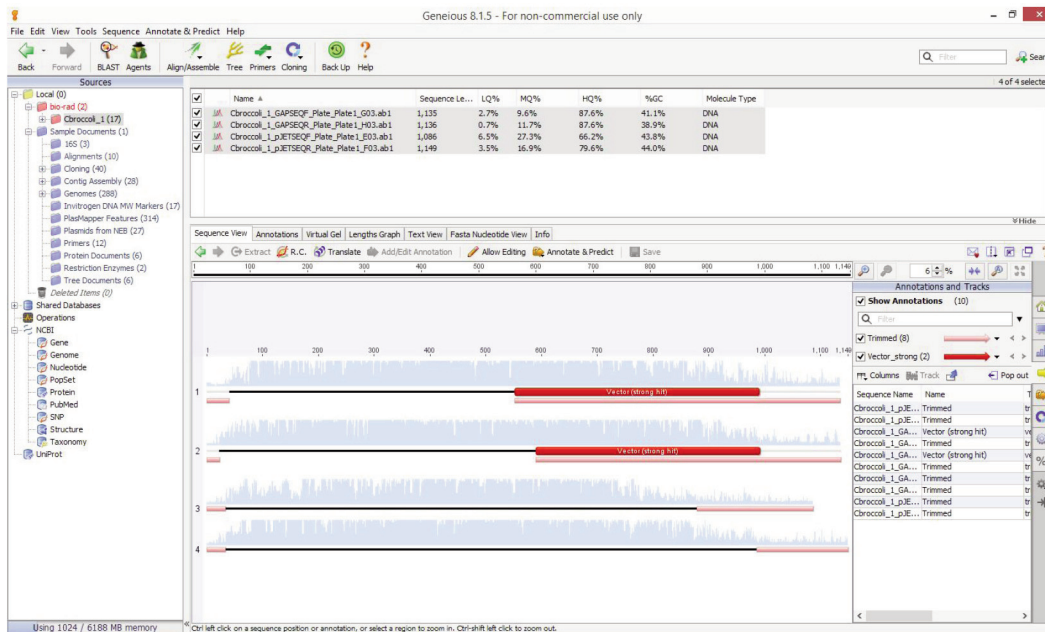


Example of a trimmed region. The light pink lines beneath the sequence indicate low-quality bases at the 5' and 3' ends of a sequence in Geneious.

1.1.3 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 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.

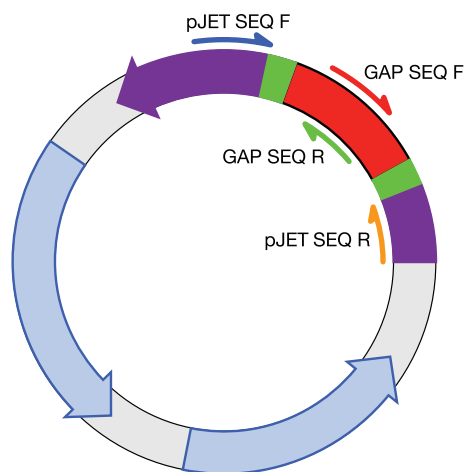
Vector masking is an optional step that can take place when data are imported into Geneious for use in an external program such as BLAST. If this option is selected, the trimmed vector regions in Geneious are also indicated by a red line that will hide it from other DNA analysis programs. This simplifies future analysis of the gene of interest by eliminating interference from the vector sequence.



Chromatograms in Geneious 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.

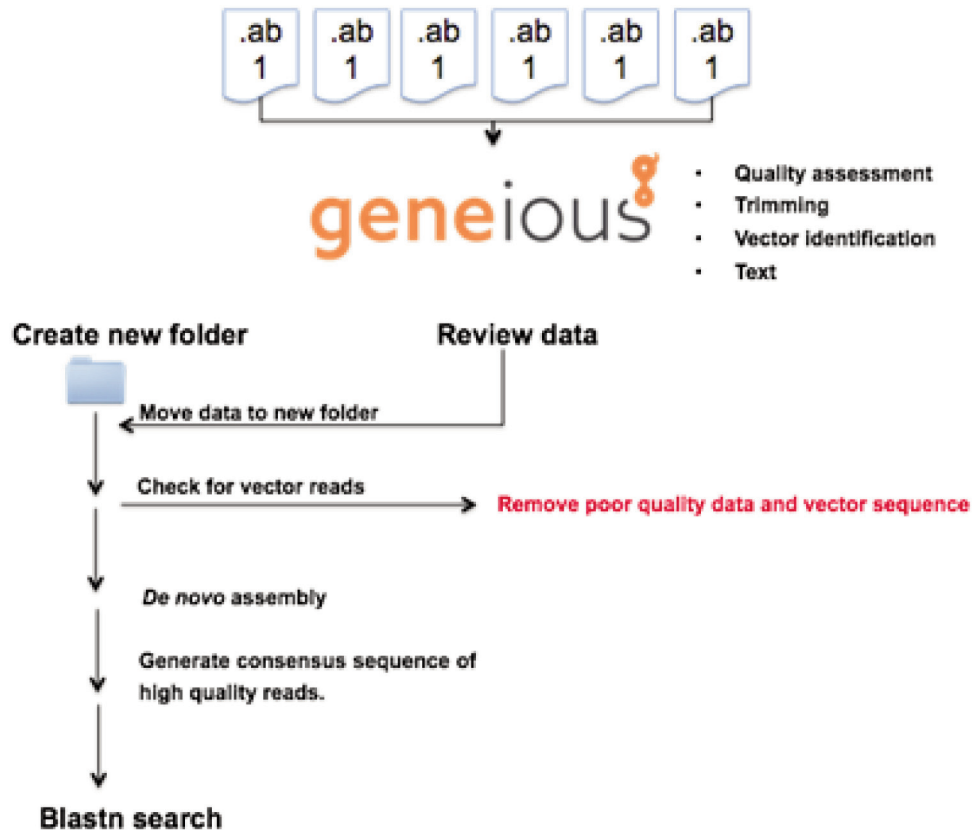
1.2 Review and identify high-quality sequences from your dataset.

In this part of the project, you will 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 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.

1.3 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.

- 1.3.1 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.
- 1.3.2 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.
- 1.3.3 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.

1.3.4 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 MWG/Operon files, zip only the files with the suffix .ab1, as these are the correct format to be used by Geneious. Geneious will automatically unzip your files when you import them.

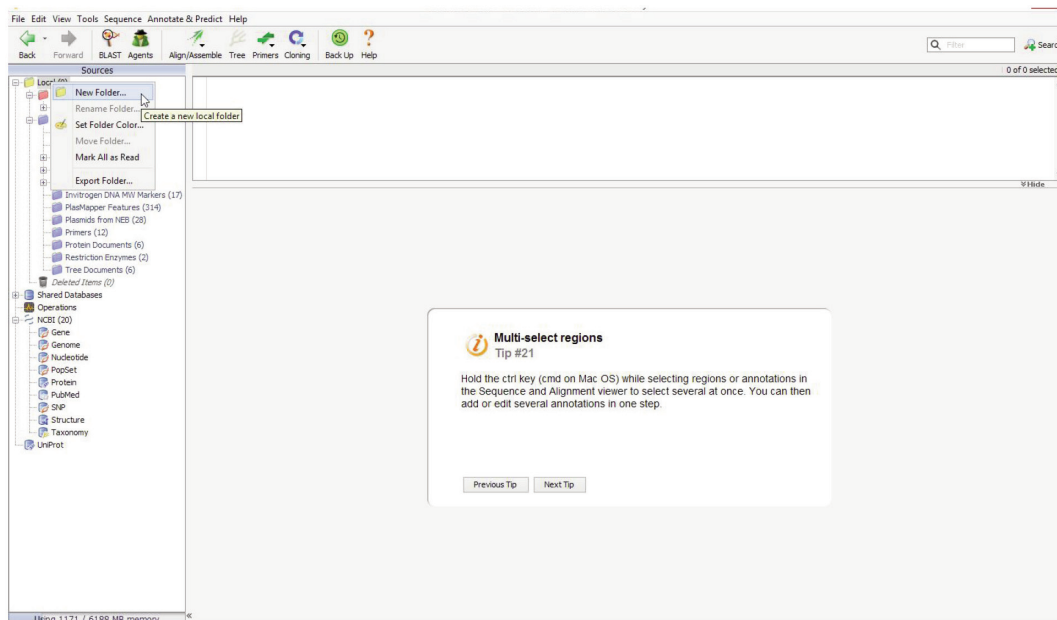
1.4 In Geneious, create a new folder for your data.

1.4.1 In the Sources panel, click to select the Local folder at the top of your directory (see figure).

1.4.2 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.

1.5 Import sequence files and view chromatograms.

There are two ways to import your .ab1 file data into Geneious. Use the File tab at the top of the Geneious program's toolbar or drag and drop your files.

1.5.1 Using the File tab:

1.5.1.1 Click to highlight your folder. The document table shows you the contents of the highlighted folder. Your folder should currently be empty.

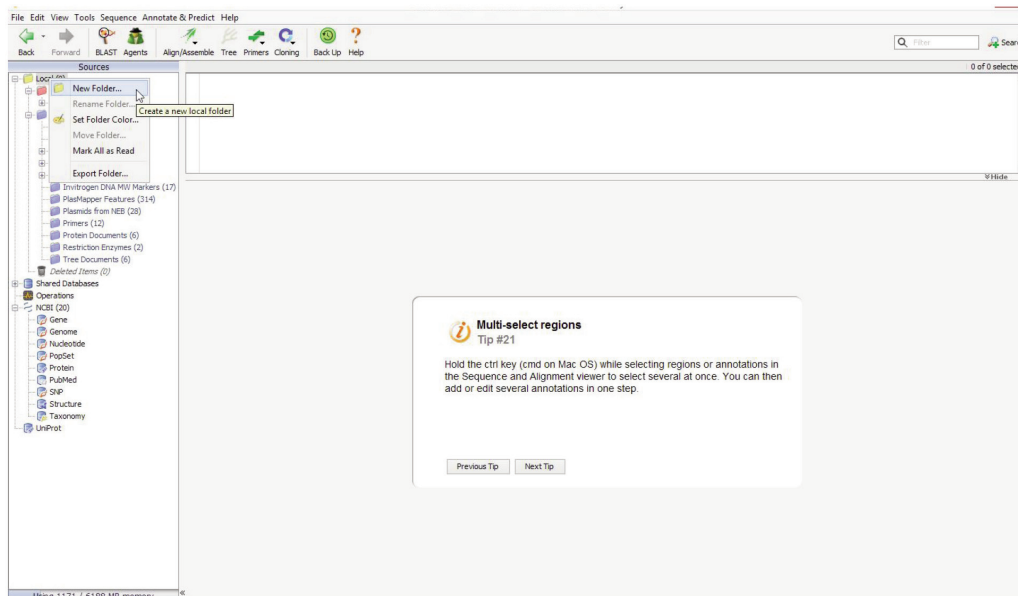
1.5.1.2 On the toolbar for the main Geneious window, click the File tab, and then mouse over Import. A window will appear. Select From File.

1.5.2 To drag and drop your .ab1 files directly into your new folder in Geneious:

1.5.2.1 Click to highlight and select your new folder in the Sources panel.

1.5.2.2 Locate your .ab1 sequencing files on the local hard drive of your computer.

1.5.2.3 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 to drop them into your folder. Your .ab1 files should now be in your folder.






Right click on the Local folder and select New Folder to create a new folder for your data.

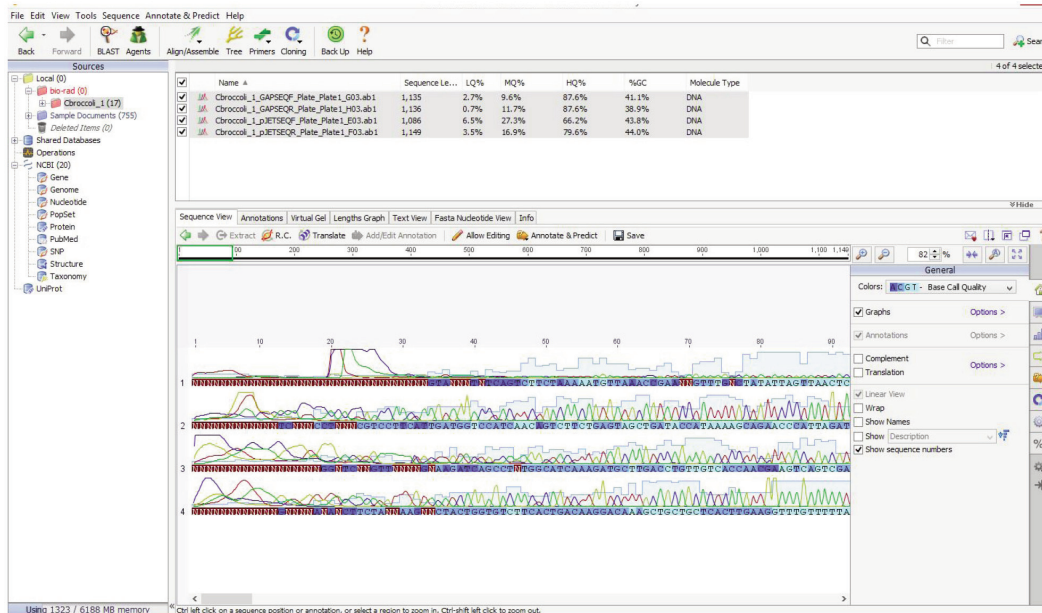
1.6 Examine your chromatogram traces and evaluate data quality.

The document table contains information about your sequencing data in a column format that includes the name of your file, the type of file (a small icon on the left side of the Name; mouse over it and a popup will tell you the kind of file it is), the sequence length, the percent of the sequence of a particular quality (for instance, HQ% = percent high quality, LQ% = percent low quality), etc.


1.6.1 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 will display the chromatogram traces superimposed on the sequence. 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.

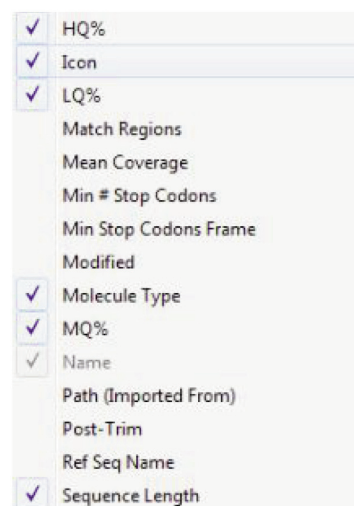


Viewing a chromatogram trace in Sequence View. All four sequences are selected for viewing from base 1, zoomed in at 82%. The files in the document table are all grayed out because they are all selected but the mouse is active in a different panel in Geneious.

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), MQ% (% medium-quality bases), HQ% (% high-quality bases), and Molecule Type.

Do your data include long sequence lengths?

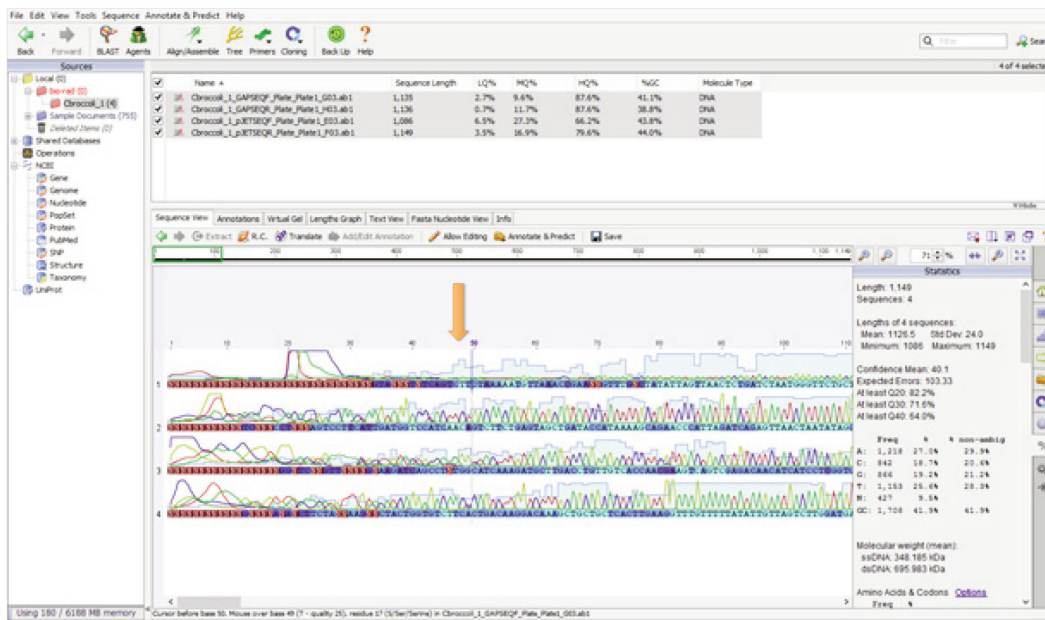
Do all your sequences have a large number of bases that are high quality?



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.

1.6.2 View the quality scores for your sequences. The quality of the base calls are 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.

1.6.2.1 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 (circled in orange) reports the quality score of the base being marked by the line cursor (marked by an orange arrow).

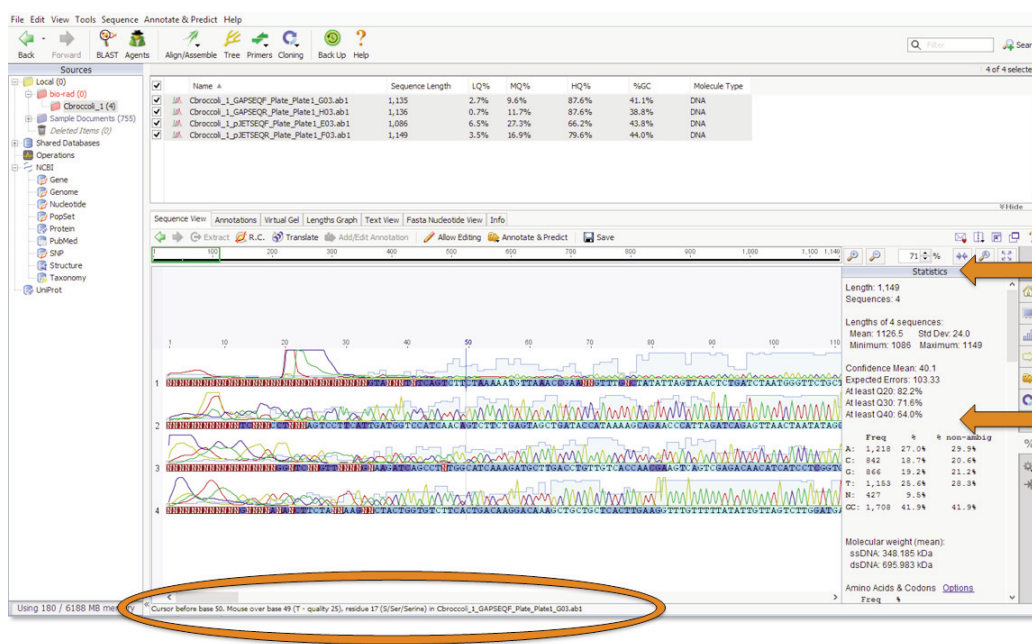
What do multiple peaks at the same location and blunt peaks indicate about a base call's quality score?

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?

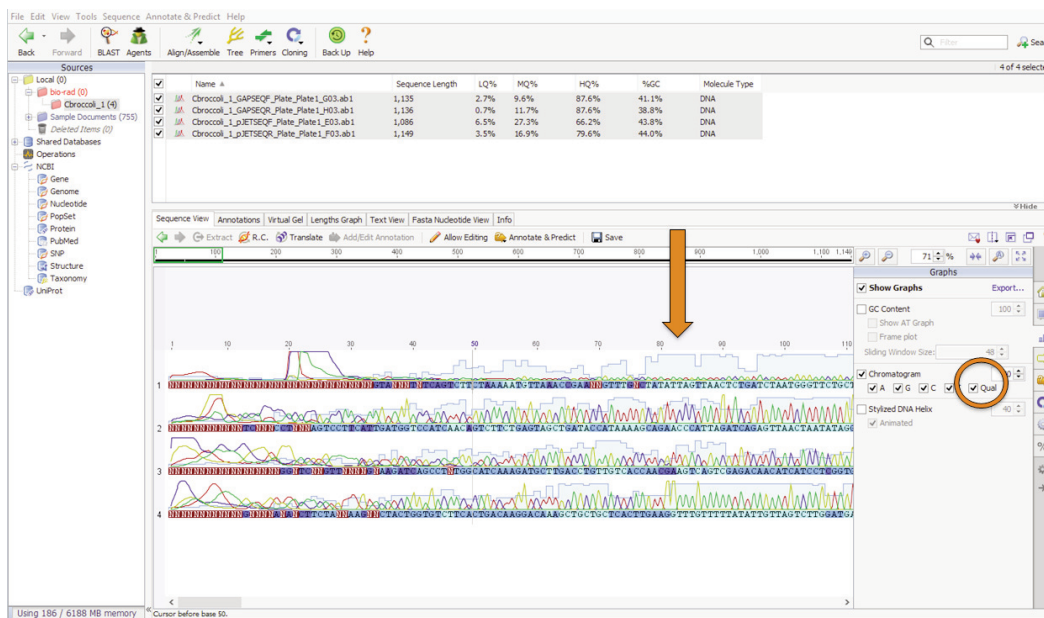
What are the differences, in terms of sharpness of peaks, number of overlapping sequences, and number of individual bases with low quality scores, between sequences with low versus high overall quality scores?

1.6.2.2 The Statistics tab  in the options panel will show the percentage of sequence(s) that fall under certain quality scores:

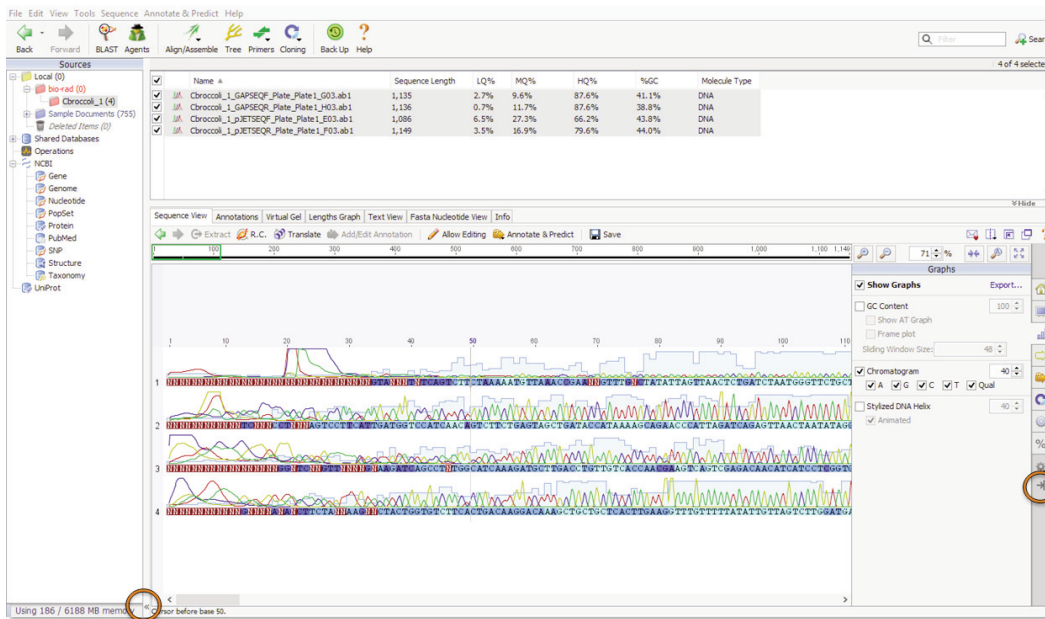


The Statistics tab in the options panel. The Statistics tab displays the percentage of your sequences that have specified quality scores (bottom orange arrow).

1.6.2.3 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) on (orange arrow pointing to light blue histogram) using the “Qual” checkbox (circled in orange).




Zoom in for a closer view of the chromatograms in Sequence View. You can increase or reduce the viewing screen for the chromatograms by toggling the icons to hide or reveal the Sources panel (arrowheads encircled on the lower left), or the Options panel (arrow encircled on the right).

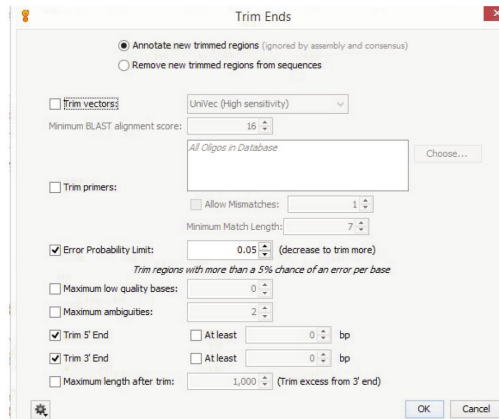
1.7 Sequence data cleanup.


In this part of the project, you will be cleaning up your sequences using Geneious 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

1.7.1 Trim the low-quality regions of your fragments. 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, 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. So, even if you can still see the trimmed bases, remember that they won’t interfere with the next steps in your workflow.

- 1.7.1.1** Sequences can be trimmed one at a time or in batch. Click to select one sequence (or multiple) from the document table. In the Sequence Viewer toolbar, click the Annotate & Predict button  and then choose **Trim Ends**. A new dialog box will appear:



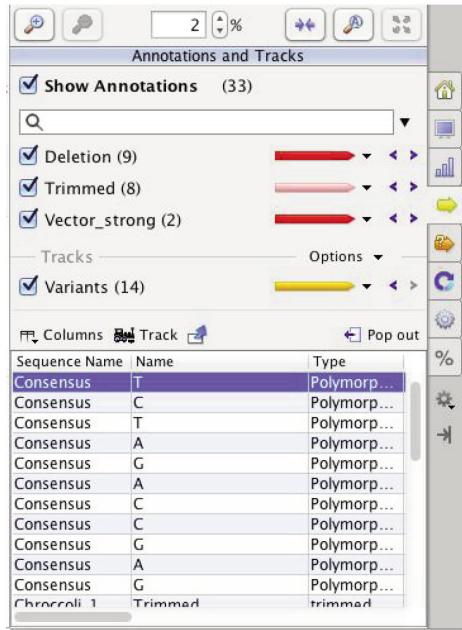
- 1.7.1.2** Select **Annotate new trimmed regions**.
- 1.7.1.3** Uncheck the **Trim vectors** and **Trim primers** for now (you'll go back and do this step later).
- 1.7.1.4** Click **OK**.
- 1.7.1.5** In the Sequence View toolbar, click **Save** to preserve the new trim regions. Geneious will now process the data and place 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. 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.

1.7.1.5.1 Annotation tools in Geneious.

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.

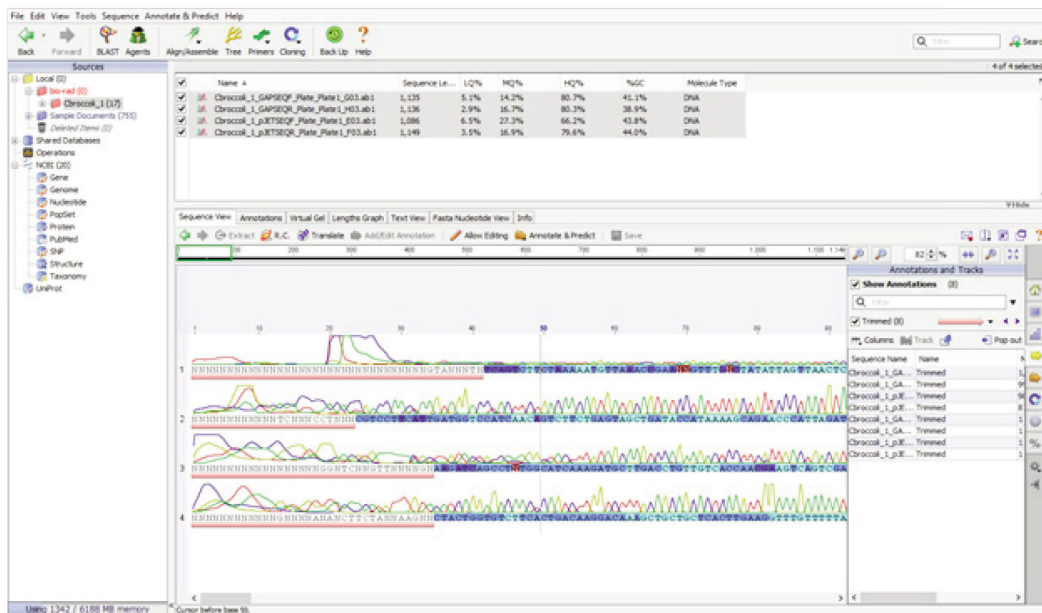
A track is a collection of one or more annotation types. Tracks are stacked vertically underneath the sequence, with a separate line for each track and its annotations. For example, when you trimmed your sequences for quality and vector masking in Section 1, these defined sections of your sequence became 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 yellow arrow icon (➡) in the options panel.



If you uncheck the box for Show Annotations, all of the annotations will be turned off and they will disappear from the Sequence View window.

- Directly beneath the top checkbox is a text field where you can search for annotations.
- Beneath the search box 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. Tracks with only one annotation type will show a single listing (for instance, CDS), while tracks with multiple annotations will show a listing of contained annotations, segregated by the annotation type (for instance, Trimmed).
- Clicking on the left or right arrowheads (◀ ▶) will take you to the part of the sequence where the annotation resides in the Sequence View window.



Quality-trimmed regions are annotated in light pink lines beneath the sequence (orange arrow).

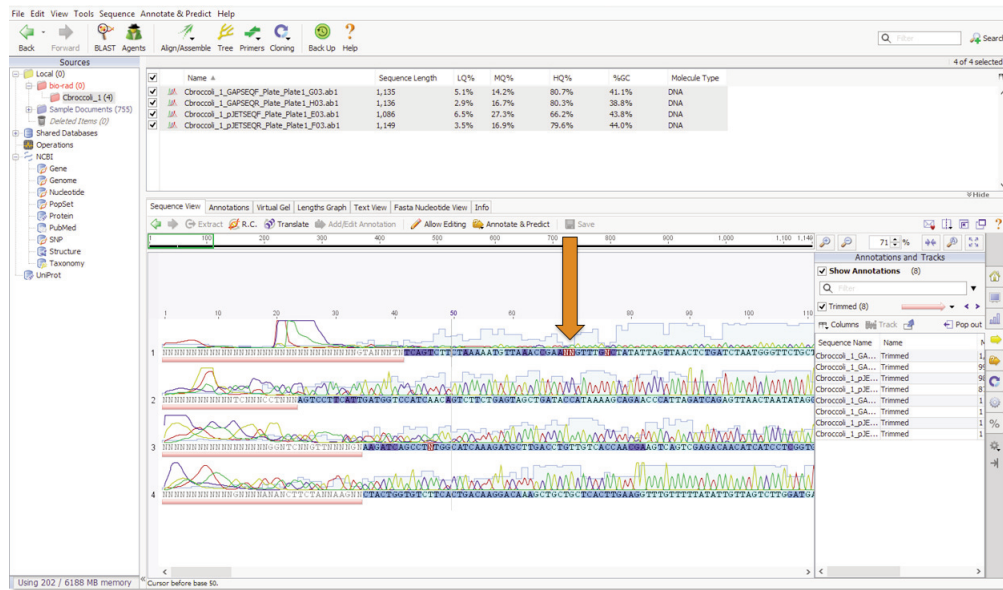
These regions are hidden from downstream processes and will not interfere with further bioinformatics steps. Sequences are zoomed in to 82%.

1.7.2 To manually trim your sequences:

Note: You will only want to trim the Ns 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 more than simply a low-quality error.

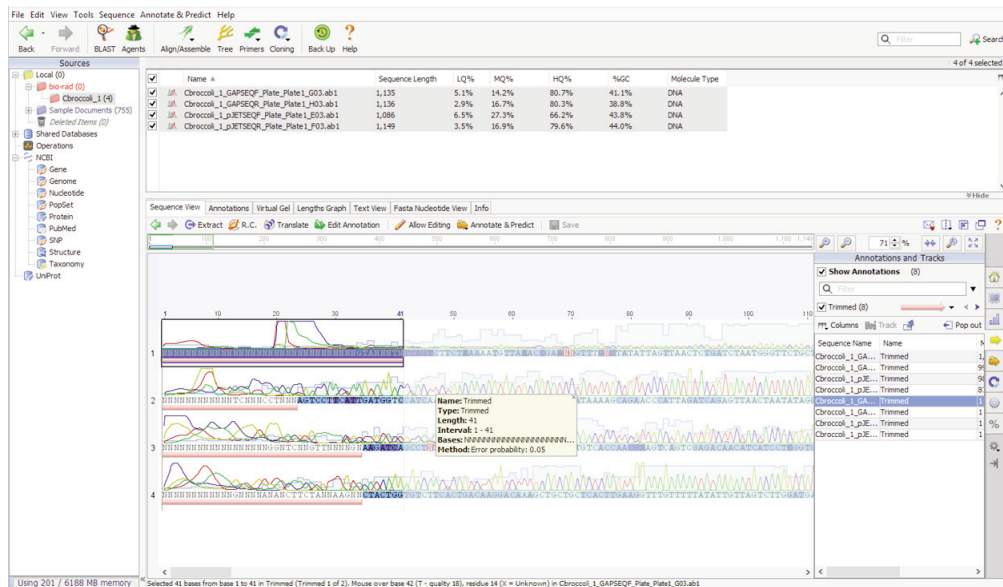
1.7.2.1 Click the single sequence you would like to work with and click the Allow Editing button in the Sequence View tab toolbar. Identify the region you would like to manually trim away. In the cbroccoli example below, the problematic bases are the two Ns in the first line of the top sequence.

1.7.2.2 Click the trim annotation. Hover your mouse over the end of the annotation and it will change to a vertical bar with arrows to the right and left:



Manual trimming of 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 trim annotation to cover them.

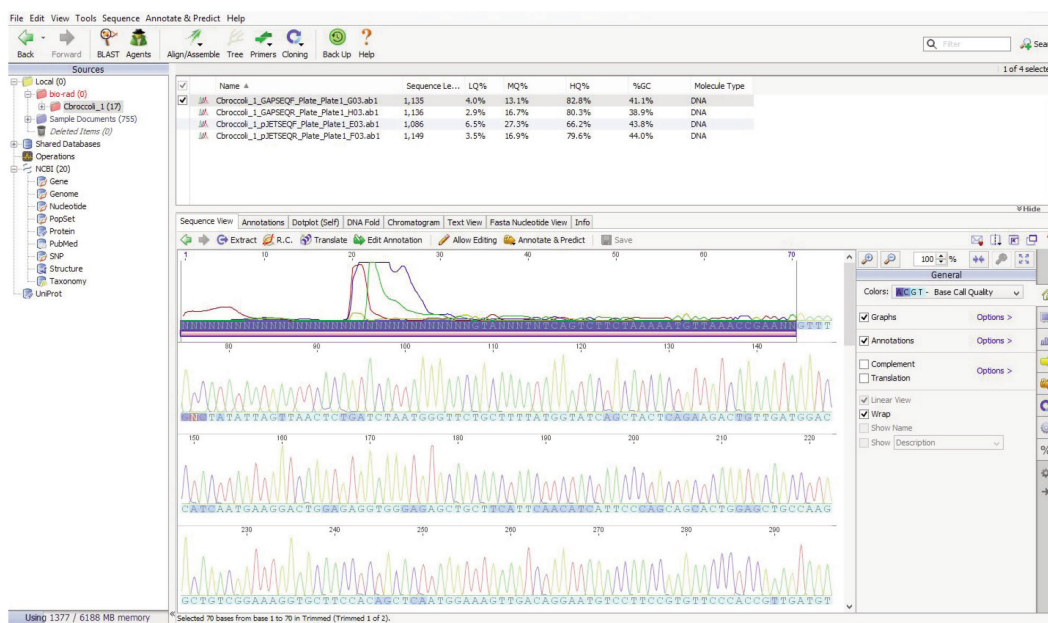
1.7.2.3 Click and drag the vertical bar to extend the light pink trim annotation and cover the Ns:



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.

1.7.2.4 In the Sequence View toolbar, click **Save** to preserve the new trim region. The trim annotation now covers the problematic Ns:

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 on in the Sequence View toolbar will be grayed out once the new trim region is saved. When the annotation is selected, the dark blue highlighting indicates the sequence covered by the trim region.


1.7.2.5 Uncheck Allow Editing when you are finished.

1.7.2.6 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.

1.7.3 Move low-quality 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

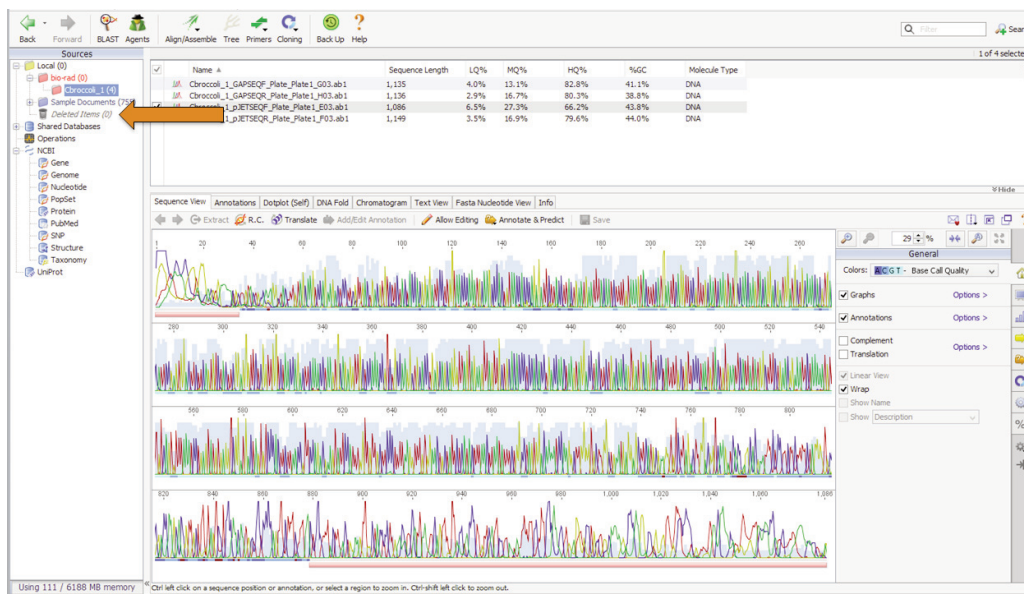
 Deleted Items (under the Local folder in the Sources panel) so as to not confuse any of your good sequences with poor sequences.

1.7.3.1 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 with a minimum Q20, Q30, Q40 scores listed below. In the example shown below, the sequence length is 1,086 bases with at least 93.5% of the trimmed sequence containing scores of at least Q20. Ninety-three and a half percent of 1,086 is 1,034 bases. 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.

1.7.3.2 Highlight the file names of any chromatogram files with fewer than 50 bases of at least Q20 and drag them to the Deleted Items folder.



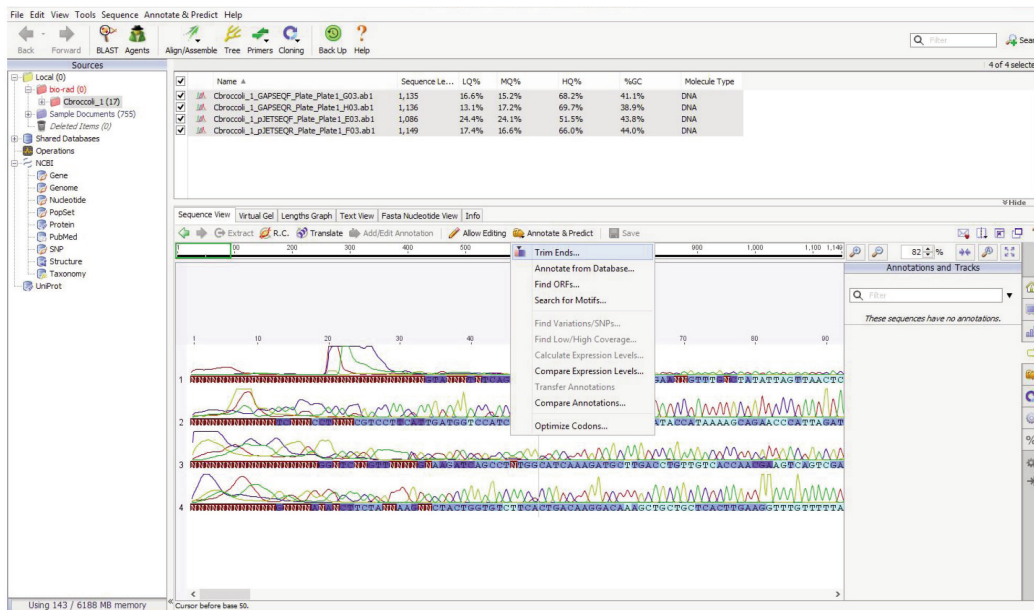
The Deleted Items folder (orange arrow) is a subfolder of the Local folder in the Sources panel.

1.8 Removing pJET1.2 vector sequence.

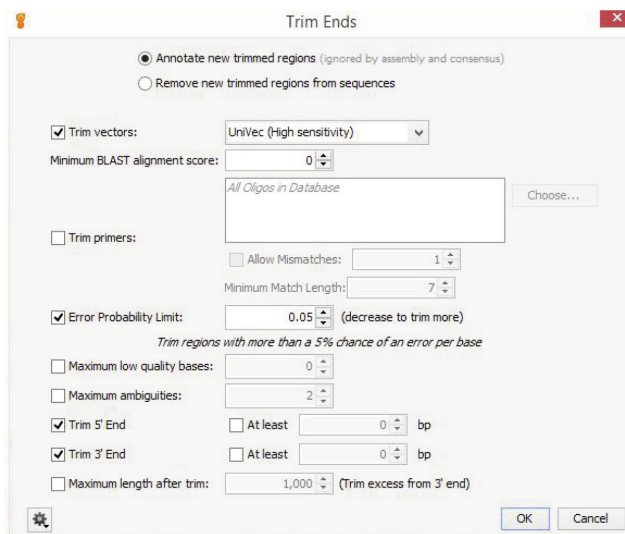
When using sequencing primers that anneal to the plasmid, such as the pJET SEQ F and pJET SEQ R sequencing primers, the resulting chromatogram may include sequence from the cloning vector. This sequence could immediately follow the sequencing primer or, if the sequenced clone is not particularly long, sequencing reactions could continue beyond the end of the gene and sequence the plasmid at the far end of the insert.

1.8.1 Trim vector sequences from your chromatogram files. Just as in quality trimming, vector trimming can be carried out on single sequences or in batch.

1.8.1.1 In the Sequence View toolbar, click the Annotate & Predict button and select **Trim Ends**:



1.8.1.2 A new dialog box will appear:



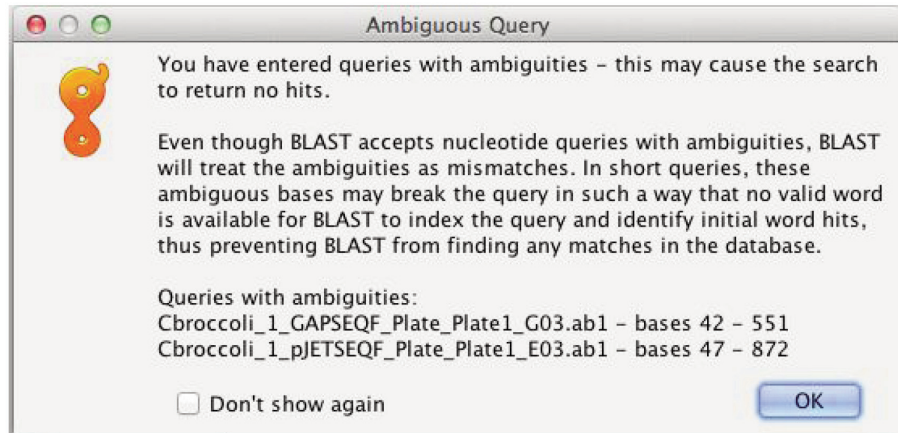
1.8.1.2.1 Select **Annotate new trimmed regions**.

1.8.1.2.2 Select **Trim vectors**, and in the dropdown menu select the default sensitivity setting, **UniVec (High sensitivity)**.

1.8.1.2.3 Set the minimum BLAST alignment score to **0**.

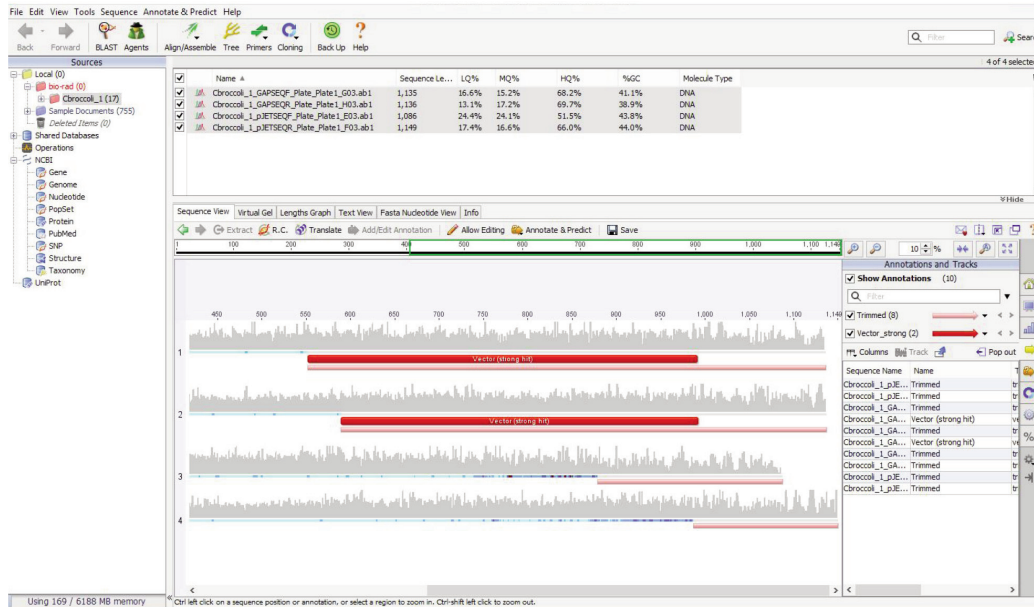
1.8.1.2.4 Click **OK**. This will trim vector from the ends of your sequence.

1.8.1.2.4.1 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 (the next step in the workflow). This will affect the returned results. If you see this window, 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. 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.

- 1.8.1.2.5** A new annotation may now appear in bright red beneath your sequence, indicating that a particular region has a strong hit to a vector sequence found in the UniVec database and is not likely to be part of your plant gene.

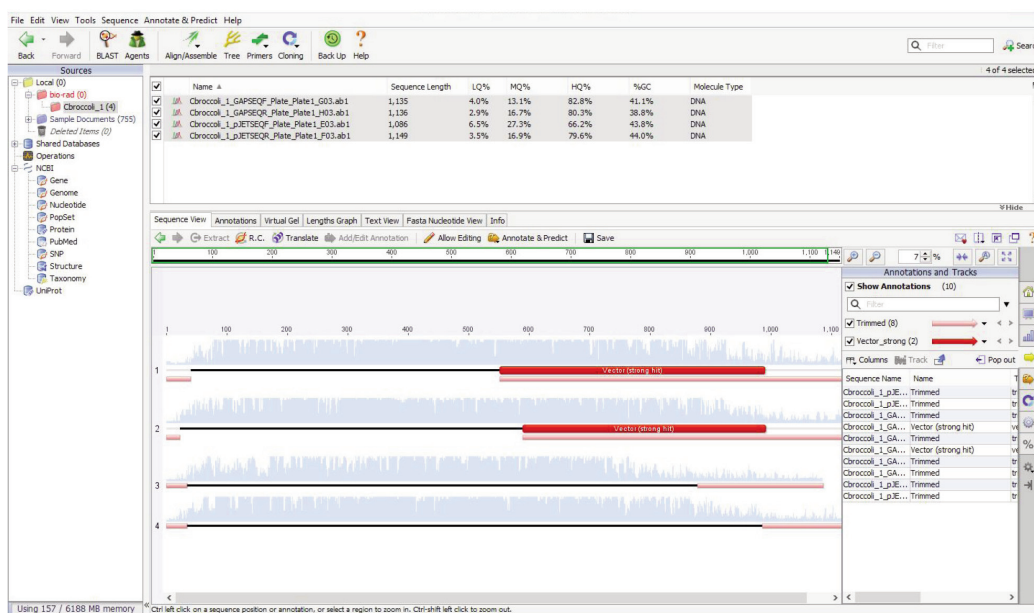


Vector-trimmed sequences. Bright red lines indicate strong hits to vector sequence, most likely from pJET1.2.

- 1.8.1.2.6** In the Sequence View toolbar, click **Save** to preserve the new trim regions.

- 1.8.1.2.7** Perform the vector trimming steps for all four of your sequences.

- 1.8.2 Move vector clones to the Deleted Items folder.** 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. Alternatively, you can select the poor sequence files and click the Delete key on your keyboard to move them 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 shown above from chroccol contain more than 50 bases of high-quality, non-vector-matching sequences so they will be kept for further analysis.

1.9 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.

1.10 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 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	

1.11 Results analysis of sequence data.

1. 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 3 when you will be assembling sequences.

1.12 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. Determine Sequence Identity Using BLAST

In this part of the procedure you will compare all the trimmed sequences to the sequences in a selected database. This is called a 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 (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 .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 Section 2.2 (Interpreting your blastn results).

You may easily run single or multiple BLAST searches (that is, a batch search) using NCBI's BLAST within the Geneious program. Geneious supports searching for RNA, DNA, and protein sequence.

IMPORTANT NOTE: Regarding BLAST searches using Geneious:

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 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 Section 2.1 using Geneious or using NCBI's BLAST website. Use Appendix I for protocol steps on how to export sequences as FASTA files for BLAST searching directly on the NCBI website.

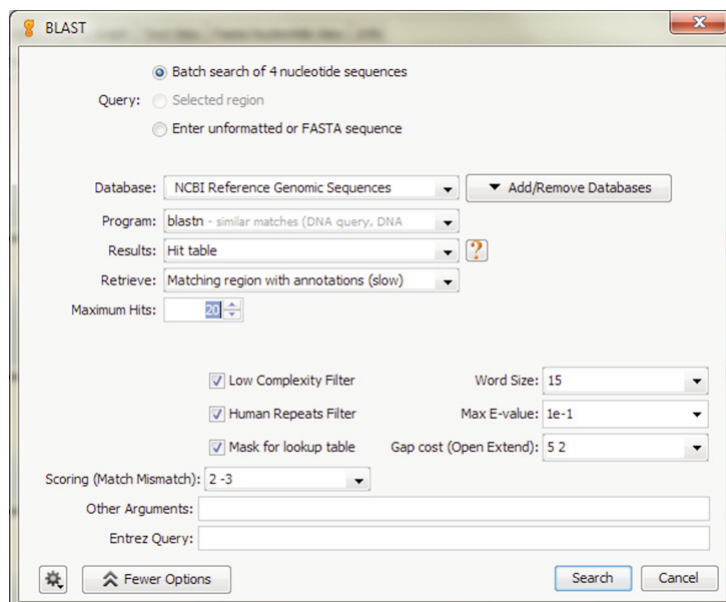
2.1 Using BLAST to perform a sequence search at NCBI using Geneious.

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. Be sure that your computer is connected to the Internet to perform BLAST searches.

2.1.1 Perform a BLAST search on multiple sequences against the NCBI sequence database.

2.1.1.1 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**.

2.1.1.2 Select the BLAST icon in the menu bar. A new dialog box will open:



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.

2.1.1.2.1 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**

2.1.2 Perform a BLAST search on a single sequence against the NCBI sequence database using Geneious.

2.1.2.1 Select a sequence in the document table.

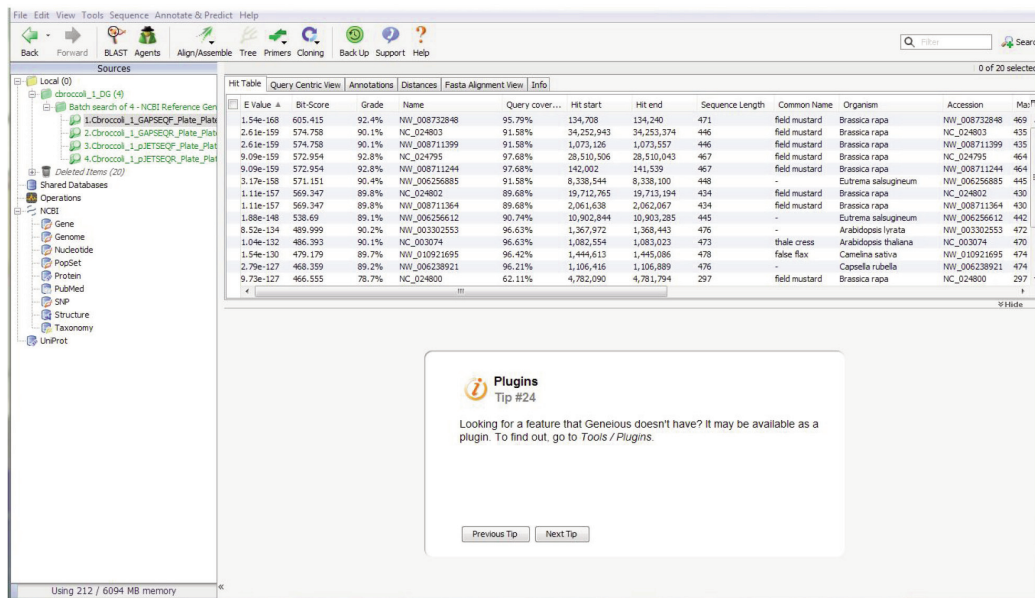
2.1.2.2 Select the BLAST icon in the menu bar. A new dialog box will open.

2.1.2.2.1 Follow step 2.1.1.2.1 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 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.

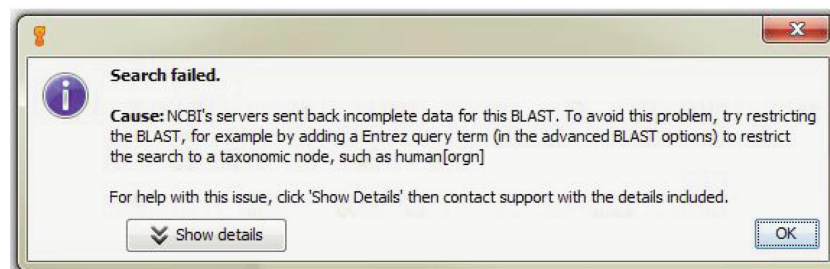
Tip: When results start downloading, a magnifying glass icon will appear on the folder. You can begin your analysis as soon as results start showing up in your folder.

2.1.3 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.

- 2.1.4 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:



- 2.1.4.1 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 2.1.2 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 if you are doing an experiment. To get relevant results, you need to optimize the experimental conditions settings.

2.2 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. By default, the search results are ordered by their **E-value**, which 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 is the same as 1.54e-168 and 1.54xe-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 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 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.

Plugins Tip #24
Looking for a feature that Geneious doesn't have? It may be available as a plugin. To find out, go to **Tools / Plugins**.


E Value	Bit-Score	Grade	Name	Query cover...	Hit start	Hit end	Sequence Length	Common Name	Organism	Accession	Max
1.54e-168	605.415	92.4%	NW_008732848	95.79%	134,708	134,240	471	field mustard	Brassica rapa	NW_008732848	469
2.61e-159	574.758	90.1%	NC_024803	91.58%	34,252,943	34,253,374	446	field mustard	Brassica rapa	NC_024803	435
2.61e-159	574.758	90.1%	NW_008711399	91.58%	1,073,126	1,073,557	446	field mustard	Brassica rapa	NW_008711399	435
9.09e-159	572.954	92.8%	NC_024795	97.68%	28,510,506	28,510,043	467	field mustard	Brassica rapa	NC_024795	464
9.09e-159	572.954	92.8%	NW_008711244	97.68%	142,002	141,539	467	field mustard	Brassica rapa	NW_008711244	464
3.17e-158	571.151	90.4%	NW_006256885	91.58%	8,338,544	8,338,100	448	-	Eutrema salisugneum	NW_006256885	445
1.11e-157	569.347	89.8%	NC_024802	89.68%	19,712,765	19,713,194	434	field mustard	Brassica rapa	NC_024802	430
1.11e-157	569.347	89.8%	NW_008711364	89.68%	2,061,638	2,062,067	434	field mustard	Brassica rapa	NW_008711364	430
1.88e-148	538.69	89.1%	NW_006256612	90.74%	10,902,844	10,903,285	445	-	Eutrema salisugneum	NW_006256612	442
8.52e-134	489.999	90.2%	NW_003302553	96.63%	1,367,972	1,368,443	476	-	Arabidopsis lyrata	NW_003302553	472
1.04e-132	486.393	90.1%	NC_003074	96.63%	1,082,554	1,083,023	473	thale cress	Arabidopsis thaliana	NC_003074	470
1.54e-120	479.179	89.7%	NW_010921695	96.42%	1,444,613	1,445,086	478	false flax	Camelina sativa	NW_010921695	474
2.79e-127	468.359	89.2%	NW_006238921	96.21%	1,106,416	1,106,889	476	-	Capella rubella	NW_006238921	474
9.73e-127	466.555	78.7%	NC_024800	62.11%	4,782,090	4,781,794	297	field mustard	Brassica rapa	NC_024800	297

Example of BLAST results in a Hit Table.

2.2.1 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 popup bubble will tell you this icon is called “Change the visible columns”). This will reveal all the column options that are available.

Columns

- ☒ E Value
- ☒ Bit-Score
- ☒ Grade
- ☒ Name
- ☒ Query cover...
- ☒ Hit start
- ☒ Hit end
- ☒ Sequence Length
- ☒ Common Name
- ☒ Organism
- ☒ Accession
- ☒ Max

Column options in the Hit Table. Not all the column options will be automatically displayed in the Hit Table. Use the small data table  on the upper right to reveal more 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.

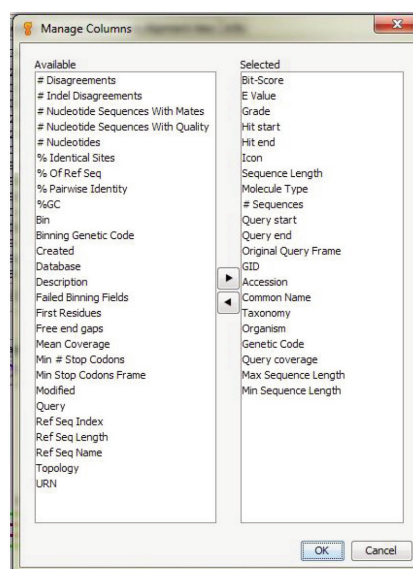
Generally, the closer the E-value is to zero, the more significant the match. An exception is that virtually identical short alignments have relatively high E-values because shorter sequences have a higher probability of occurring in a database purely by chance. **Tip:** The E-value can be a convenient way to create a significance threshold for reporting results. You can change the E-value threshold within Geneious 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 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.




Manage Columns lets you choose to view the data that will be most helpful to you.

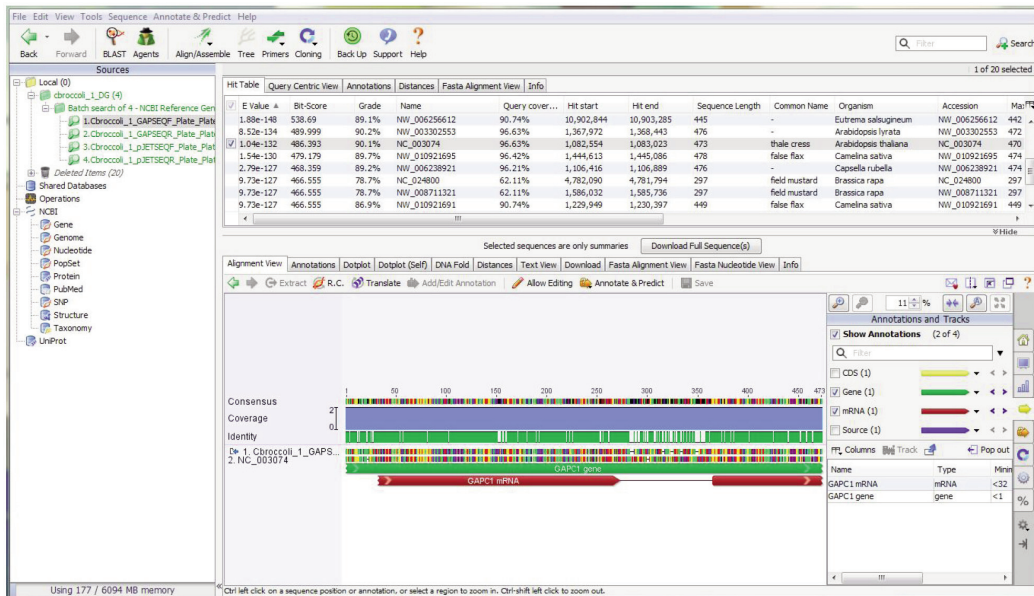
Click the small data table icon, then select Manage Columns from the list. A dialog box will open in which you can select your options.

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 will display the pairwise alignment for that hit in the document viewer window.


2.3.1 Click on one BLAST result from the Hit Table (lets say one with a description of *Arabidopsis thaliana*) 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 (see description in 2.3.1.2)
- The depth of coverage chart, displayed in blue (see description in 2.3.1.3). 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 software interface. The top menu bar includes File, Edit, View, Tools, Sequence, Annotate & Predict, and Help. The left sidebar shows the Sources panel with a tree view of local and shared databases. The main window is divided into several panels. The top panel is the Hit Table, which lists search results with columns for E Value, Bit-Score, Grade, Name, Query cover..., Hit start, Hit end, Sequence Length, Common Name, Organism, and Accession. The second panel is the Alignment View, which shows the consensus sequence, coverage chart, identity chart, and the query sequence (GAPC1 mRNA) aligned to the hit sequence (GAPC1 gene). The right panel shows the Annotations and Tracks for the selected hit, including CDS, Gene, mRNA, and Source. The bottom status bar indicates 'Using 177 / 6094 MB memory'.


Alignment view for one BLAST hit. An *Arabidopsis thaliana* query result is chosen for comparison with the GAP SEQ F read from cbroccoli. In Alignment View, you can see information (see 2.3.1.1 for more details) such as the consensus sequence, depth of coverage, identity chart, and known annotations from *Arabidopsis thaliana*.

2.3.1.1 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: Mousing over the annotations will display a popup with more information, including the gene name and gene product, if known. You can select text from within this yellow popup and copy the text into another file or an electronic lab notebook.

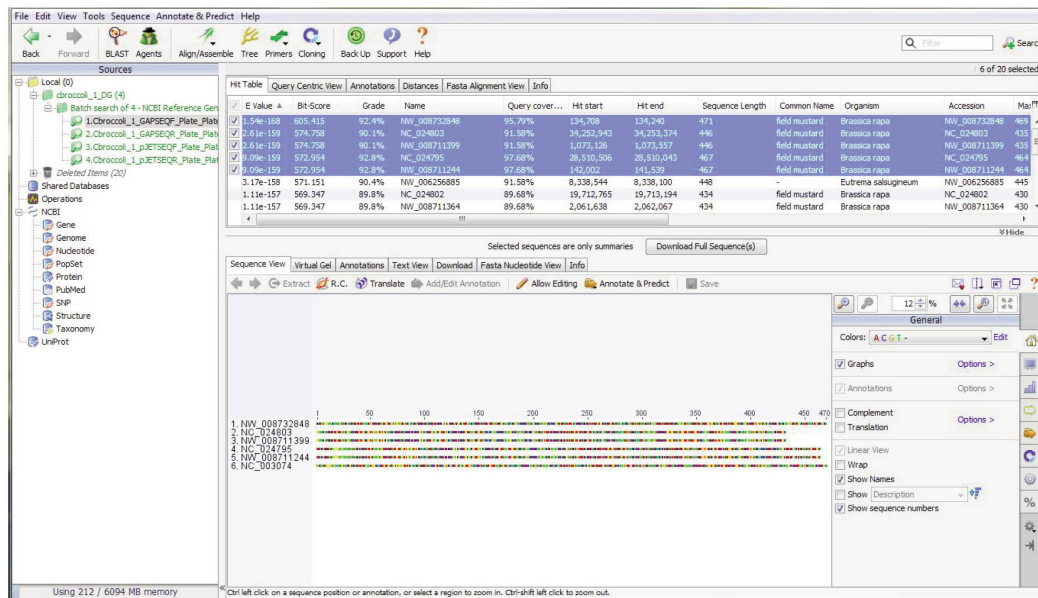
2.3.1.2 A consensus sequence is an alignment that occurs, with minor variations, across many genetic locations or organisms. It is constructed from the order of the nucleotides appearing most frequently at each position of a sequence alignment. The consensus is the same length as the contig (which includes 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).

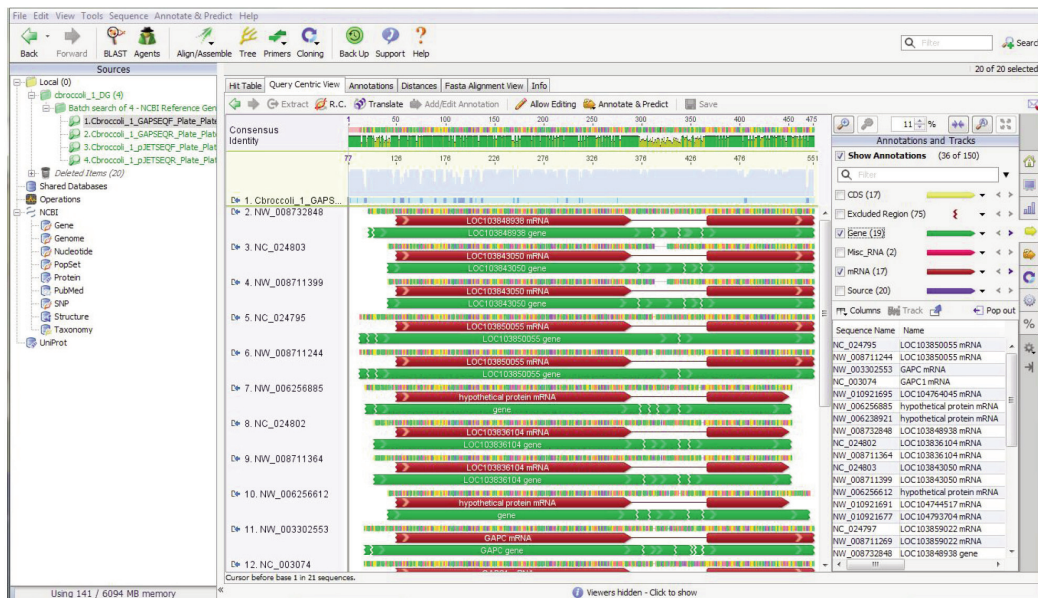
2.3.1.3 Depth of coverage represents the number (often an average) of nucleotides contributing to a portion of assembled sequences. On a whole-genome basis it means that each base has been sequenced, on average, a particular number of times (for example, 10x, 20x, etc.). For a specific nucleotide, it 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.

2.3.2 Now, select **multiple** BLAST results from the Hit Table, either by clicking the check boxes on the left-hand side or by holding down the shift key and clicking on multiple documents. When you look at Sequence View in the document viewer, you will now see only the sequences of the hits you selected and no information from the cloned sequence:



Viewing multiple BLAST query results in Sequence View.

2.3.2.1 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:



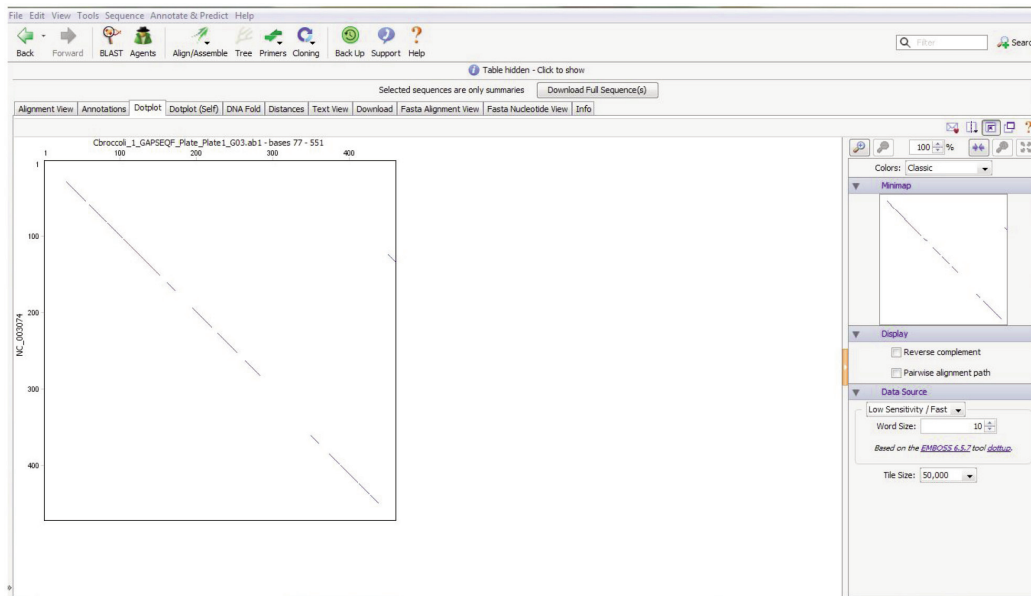
2.3.2.1.1 If a particular hit looks interesting to you, you can click on the icon next to the name of the BLAST hit to quickly navigate to the single sequence alignment in Alignment View.

2.3.3 Other useful tabs in the document viewer (when Hit Table is selected).


2.3.3.1 Dotplot

Selecting the query sequence and a single BLAST result brings up a new tab labeled Dotplot (third tab in the document viewer window). A dot plot is a visual comparison of two sequences in which each axis of the plot represents one of the two sequences being compared. Whenever the position of a sequence matches the position of the other sequence, a dot or short diagonal is drawn at the corresponding position in the array. When two sequences share similarity over their entire length, a line will extend from one corner of the dot plot to the diagonally opposite corner. If two sequences share only patches of similarity, this will be revealed by longer diagonal stretches, which may appear as an overall broken diagonal line. Dot plots are a powerful method for comparing two sequences. Dot plots do not bias the sequence analysis (different alignment algorithms have different scoring priorities, thus introduce bias into the alignment) and are an ideal first-pass method for visually comparing two sequences.

Based on the dot plot, you can decide whether this is a case of global, beginning-to-end similarity or local similarity. Local similarity between two sequences refers to the occurrence of similar regions embedded in the overall sequences that otherwise lack similarity. Sequences may also contain regions of self-similarity, which are frequently termed internal repeat regions. A dot plot comparison of the sequence will reveal internal repeats by displaying parallel diagonals.



A dot plot of the Cbroccoli_1_GAPSEQF sequence with the Arabidopsis thaliana BLAST query result.

The cbroccoli sequence lies on the x-axis while the Arabidopsis sequence (NC_003074) is on the y-axis. The diagonal line represents similarity between the two sequences at those base positions. (The Dotplot 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 window.)

2.3.3.2 Text View.

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

File Edit View Tools Sequence Annotate & Predict Help

Back

Forward

BLAST

Agents

Align/Assemble

Tree

Primers

Cloning

Back Up

Support

Help

Search

Table hidden - Click to show

Selected sequences are only summaries

Download Full Sequence(s)

Alignment View Annotations Dotplot Dotplot (Self) DNA Fold Distances Text View Download Fasta Alignment View Fasta Nucleotide View Info

>NC_003074 Arabidopsis thaliana chromosome 3, complete sequence
Length = 23459830

E-value = 1.04e-132, Score = 538, BitScore = 486.393, Identities = 395/473 (83%),
Positives = 395/473 (83%), Gaps = 17/473 (3%)
Frame = +1

Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 93 TGATCTAATGGGTTCTGCTTTTATGGTATCAGCTACTCAGAAGACTGTTGATGGACCATC 152
NC_003074 1082554 TGATTAATGGGTTCTGCTTTTATGGTATCAGCTACTCAGAAGACTGTTGATGGGCTTC 1082612


Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 153 AATGAAGGACTGGAGAGGTGGGAGAGCTGCTTCATTCAACATCATTCGCCAGCAGCACTGG 212
NC_003074 1082613 AATGAAGGACTGGAGAGGTGGGAGAGCTGCTTCATTCAACATCATTCGCCAGCAGCACTGG 1082672

Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 213 AGCTGCCAAGGCTGTGGAAAGGTGCTTCCACAGCTCAATGGAAAGTTGACAGGAATGTC 272
NC_003074 1082673 AGCTGCCAAGGCTGTGGAAAGGTGCTTCCACAGCTCTTAACGGAAGTTGACTGGAAATGTC 1082732

Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 273 CTTCCGTGTGCCACCGTTGATGTCCTCAGTTGTGACCTCAGGTTAGACTCGAGAAAGC 332
NC_003074 1082733 TTTCCGTGTGCCACCGTTGATGTCCTCAGTTGTGACCTTACTGTACACTCGAGAAAGC 1082792

Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 333 TGCTACCTACGACAGATCAAGAAGGCTATCAAGTAAGCTTT----CGGTTCCAGTTAAC 388
NC_003074 1082793 TGCTACCTACGA A ATCAA AAGGCTATCAAGTAAGCTTT C T CAG T A 1082851

Chroccoli_1_GAPSEOF_Plate_Platel_003.abl - bases 77 - 551 389 TAGTTTGATCAAAAT--CTTTG---TAGATT--TAAGTAAGTATT--GGATTGTAC 438
NC_003074 1082852 -AGTTTACTATATTTCAGTATGATCAAAATTACTCACCAAGTGTGTTTACCACCAATAC 1082910

Text view of the Cbroccoli_1_GAPSEOF sequence compared with the Arabidopsis thaliana BLAST query result. The alignment between the cbroccoli_1_GAPSEOF and the Arabidopsis sequence (NC_003074) is displayed in text view. See section 2.3.3.2 for a full description. 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 status table at the top of the Geneious window.

2.4 BLAST searching all of the trimmed sequences and recording matches.

For all of the individual trimmed and clipped sequences that have gone through a BLAST search for your clone, 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

2.5 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).

2.5.1 In the Hit Table, sort the results by clicking on the column headers for Bit-Score, E-value, and Grade.

2.5.2 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 3).

A number of scenarios can occur, depending on how your cloning reaction took place:

- 2.5.2.1** If the top match to your sequence IS NOT an *Arabidopsis* sequence, it is unlikely that you have cloned an *Arabidopsis* gene.
- 2.5.2.2** If your top match IS an *Arabidopsis* sequence, then look at both the dot plot and the sequence alignment of your novel sequence with the *Arabidopsis* sequence.
- 2.5.2.3** 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.
- 2.5.2.4** 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.
- 2.5.2.5** If the Identity value is below 90%, then it is unlikely to be an *Arabidopsis* GAPC gene.
- 2.5.2.6** If the Identity value is between 90 and 100%, do the same analysis on other sequences generated from the same miniprep plasmid with other sequencing primers and determine whether these also have a high identity with the *Arabidopsis* gene.
- 2.5.2.7** If some sequences from the same miniprep plasmid have low homology or have gaps that have no homology, it is unlikely to be an *Arabidopsis* gene.
- 2.5.2.8** If all sequences from the same miniprep plasmid have 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.

2.6 Results Analysis.

1. Record which GAPDH gene you predict has been cloned for each of your minipreps.

2. Are all the sequences for each individual miniprep most similar to the same GAPDH gene, or are each of the different sequences most similar to different GAPDH genes?

2.7 Section 2 Focus Questions

1. Why might sequencing reads from the same miniprep clone but with different sequencing primers be homologous to different GAPC genes?
2. Does an E-value of zero mean that your sequence matched the subject sequence well or poorly? Explain your answer.
3. What would it mean if you found a subject sequence with an E-value of 3?
4. Why did you search the reference genomic database?

3. Assemble the Sequences and Correct Mistakes in the Base Calls

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. Current 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 to assemble sequence reads from the clones.

The steps used by many assembly programs, including Geneious, 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, 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.



Contiguous sequence built from the sequence fragments

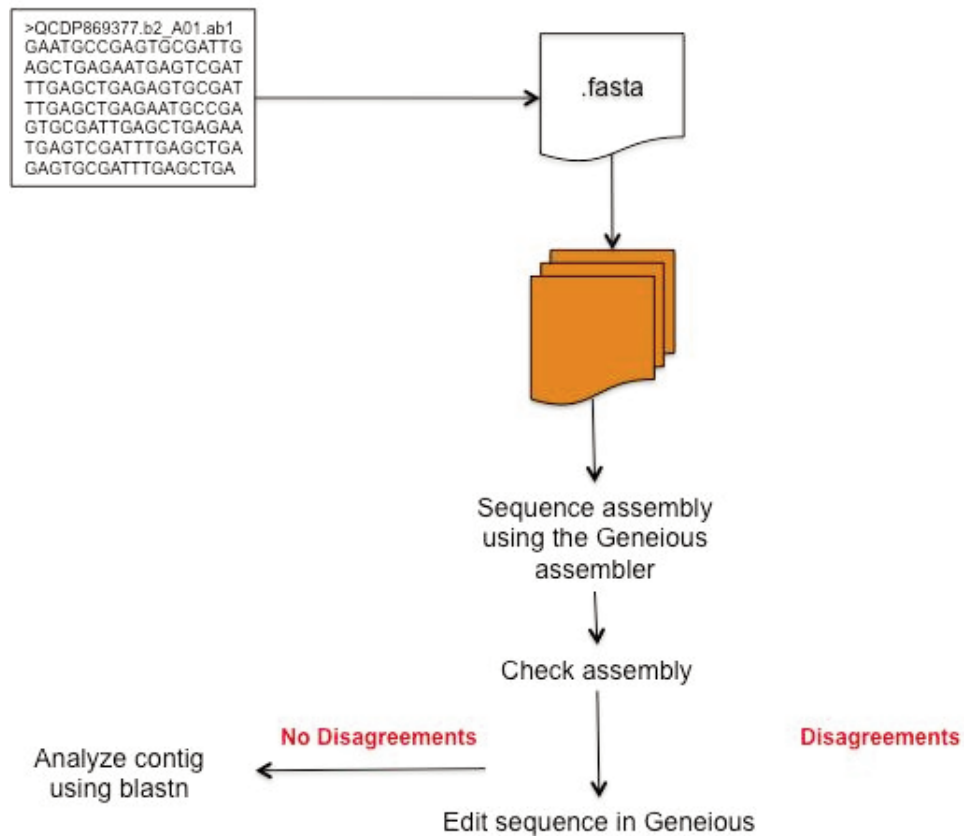
Generation of a contiguous sequence using Geneious. Individual shorter sequences are compared, aligned, and assembled by programs such as Geneious to generate longer contiguous consensus sequences. This methodology is used to continuous sequences that are longer than the current sequencing instruments 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. Finishing an entire genome can take years. In this lab, you will carry out the finishing step after you have assembled contigs.

3.1 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 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 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.

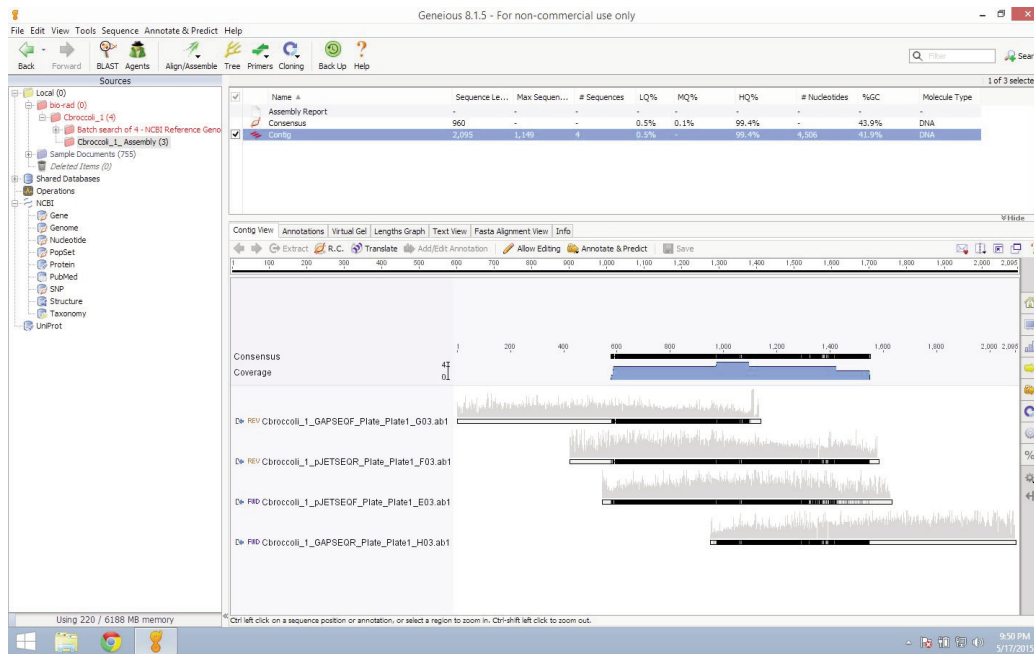
Sequence assembly workflow



Workflow for assembling sequences.

3.2 What is a contig and how is sequence assembly useful?

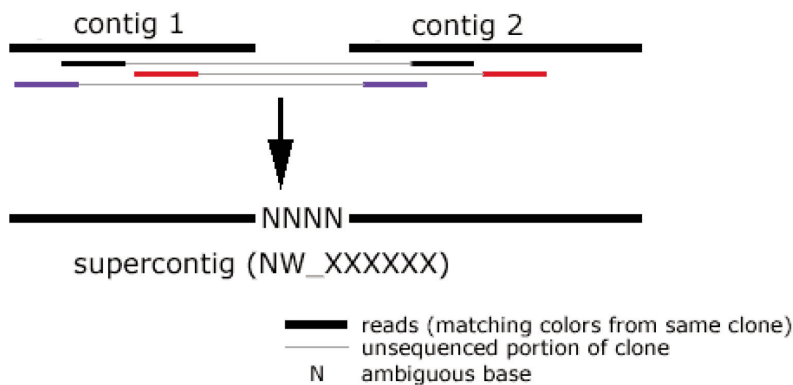
A **contig** (short for contiguous) refers to either a DNA segment or the reassembly of overlapping segments that form a continuous, extensive, and uninterrupted DNA sequence. By analyzing these segments, the researcher is able to discover the order of segments that make up various sequences. Contigs 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 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.



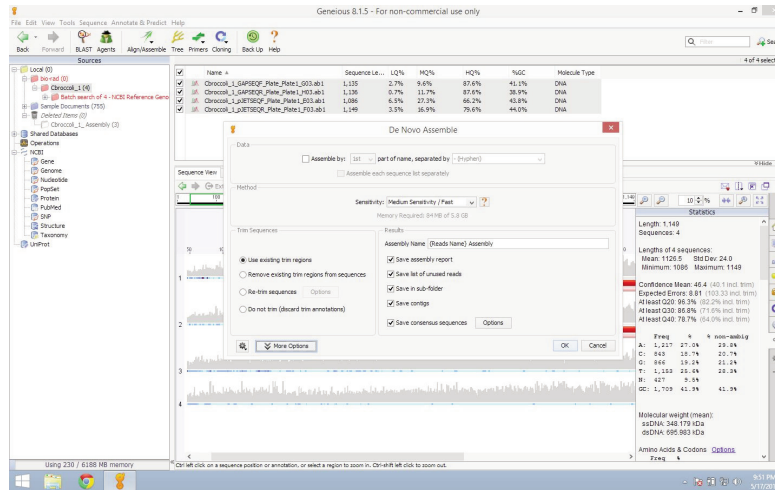
Since the lengths of the fragments are roughly known, the number of bases between contigs can be estimated.

3.3 Assembling your sequences in Geneious.

3.3.1 Select all of the documents in your folder that you would like to assemble.

3.3.2 Click the Align/Assemble icon in the menu bar.

3.3.3 Choose **De Novo Assemble** from the dropdown list. A new dialog box will appear:



3.3.3.1 Leave the “Assemble by” box in the Data section unchecked.

3.3.3.2 Select **Medium Sensitivity / Fast** in the Sensitivity dropdown menu of the Method section.

3.3.3.3 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 1.7.2.4, 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.

3.3.3.4 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.

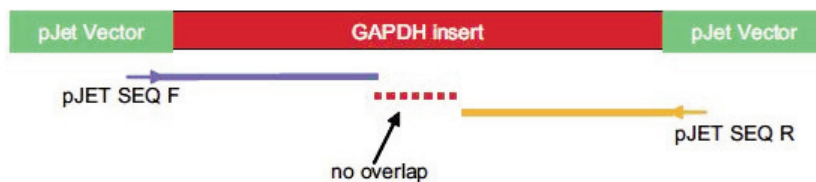
Save contigs — saves the assembly results as a contig.

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.

3.3.3.5 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.

3.3.4 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

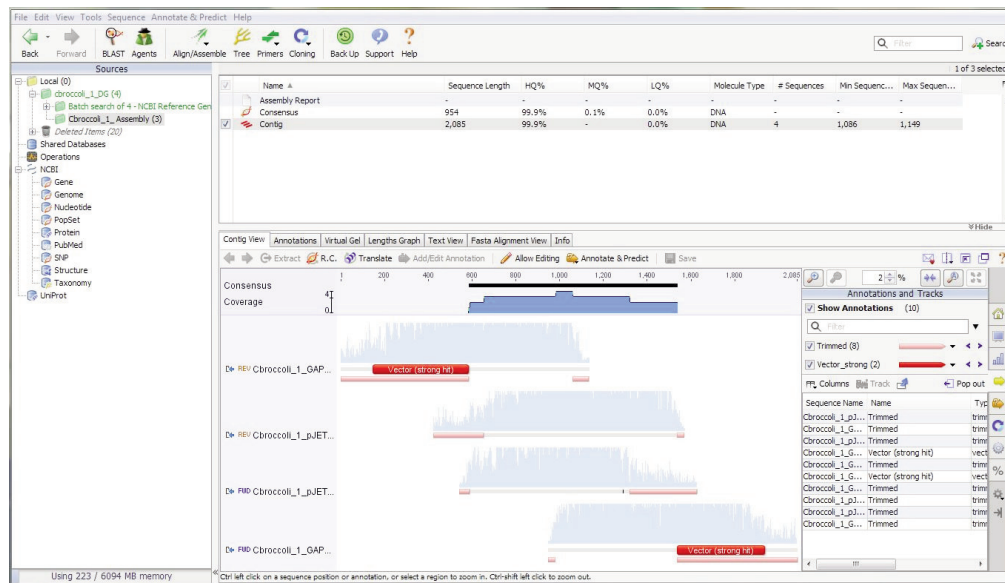
3.4 Viewing and understanding your contig.


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.

3.4.1 To view your contig:

3.4.1.1 Click the subfolder for your assembly, and then click Contig.

3.4.1.2 In the Contig View tab of the document viewer, you will see a zoomed out view of your contig:

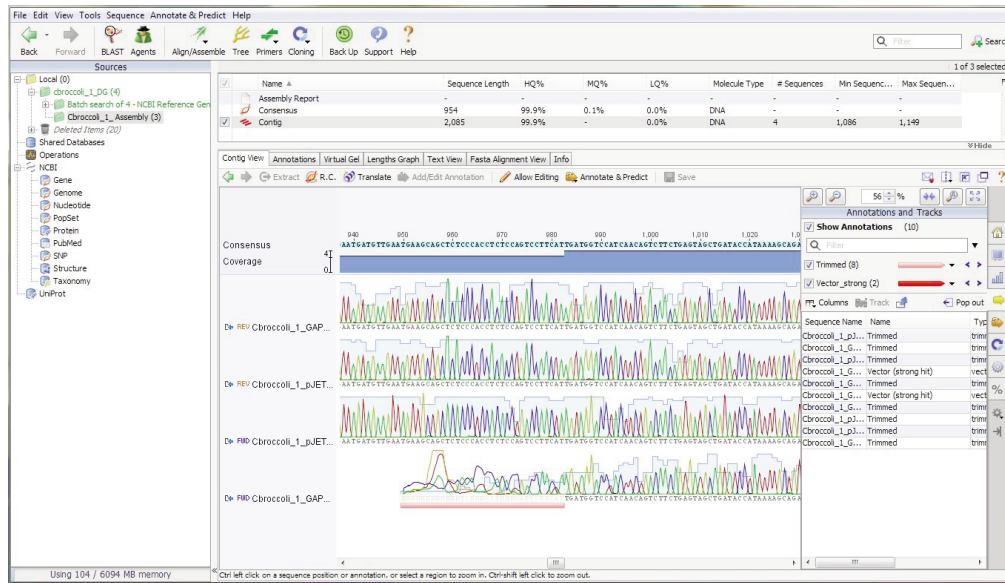


Viewing the cbroccoli contig in Geneious. Contig View allows you to see how your individual reads are assembled to form your contig. In this example, the cbroccoli contig, represented by the consensus sequence (black rectangle at the top), is generated from the four individual cbroccoli sequencing reads. 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).

3.4.1.3 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

3.4.1.4 If a read is in the same orientation as the consensus sequence, Geneious 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 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 56% at the start 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.

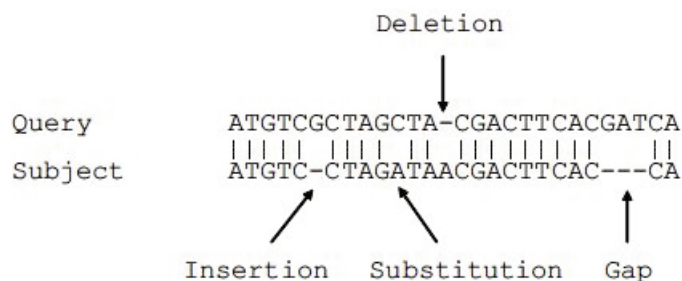
- 3.4.1.5** If you obtained good sequence from all four sequencing primers, then all your read sequences should be assembled into a single contig.
- 3.4.1.6** If you obtain multiple contigs, it is possible that the sequences you assembled did not really belong together or there wasn't significant enough 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 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.
- 3.4.1.7** If you do not see any contigs, these are sequences that could not be assembled, most likely for reasons described in section 3.3.4 and its accompanying figure. In this case, assemble your single sequences with the sequences from other student teams who worked on the same plant.

3.4.1.7.1 Create a new subfolder in the Local folder in Geneious (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 3.3. 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.

3.5 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, 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.

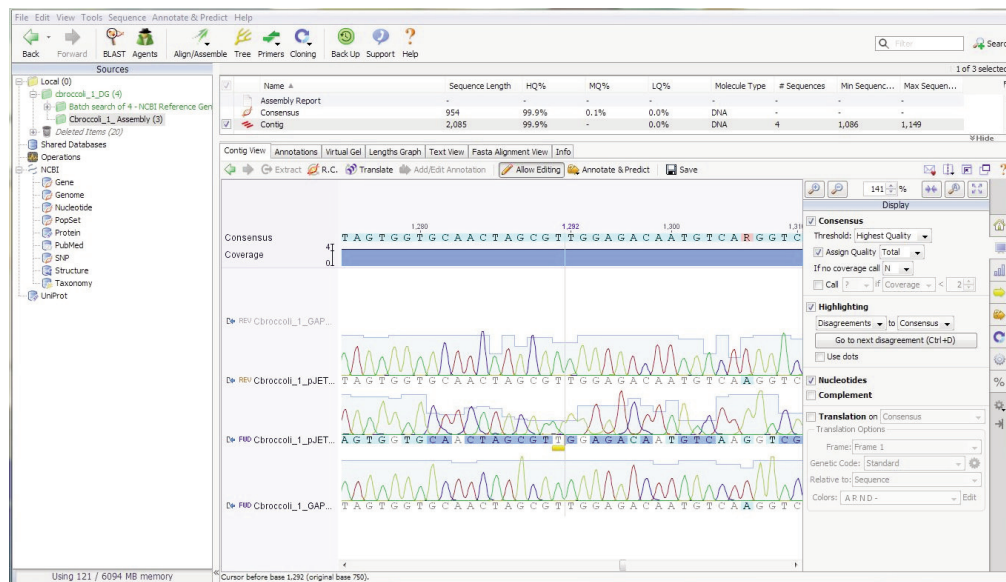


To determine which read is correct, you will review the chromatogram traces. In this part of the project, you will use Geneious to review the assembly results and resolve disagreements between reads. If you find a mistake in the consensus sequence, you will need to edit the reads and generate a new assembly and a new contig. This must be done manually and is often an iterative process.

3.5.1 Navigate through the disagreements.

3.5.1.1 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 .

3.5.1.2 In the Display tab  in the options panel, click the **Go to next disagreement** button. 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.

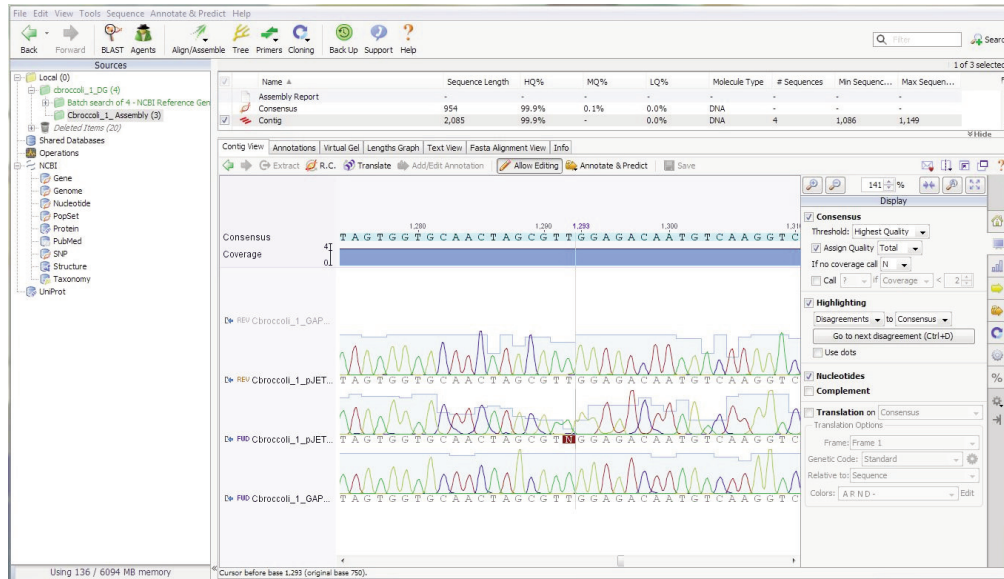
Looking at the trace above for the cbroccoli contig, it is pretty clear that one of the reads contained an error and the base that has an N is read as a T in the other three chromatograms, all of which have higher quality scores for base calling. Since the consensus sequence contained the base A and 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 apply your changes back to the source sequence documents.

3.5.1.3 To edit a base call:

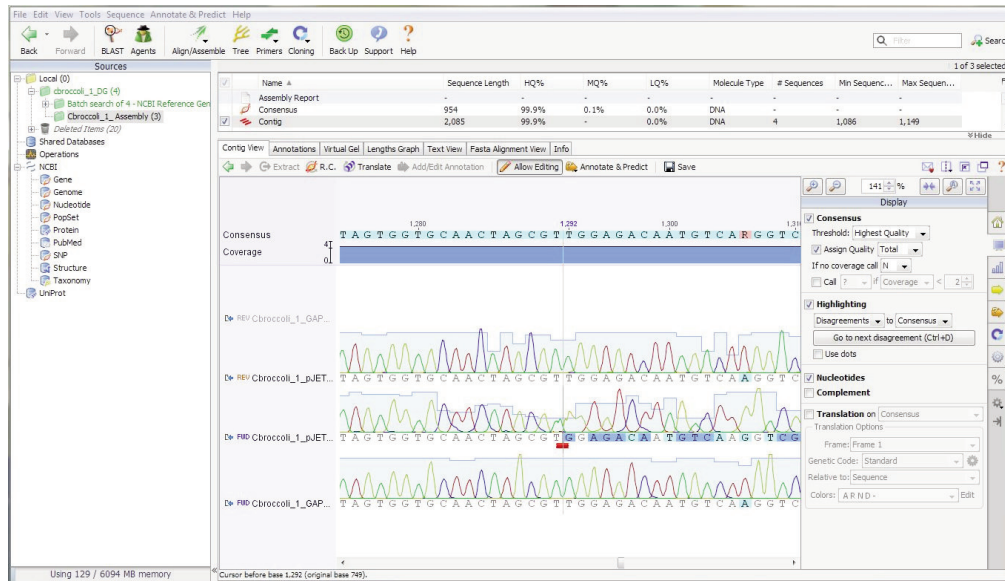
3.5.1.3.1 Identify the problematic base. In the example above, there appears to be an N in position 1,292 for the middle read whereas the base call is a T for the other two reads. Since the other two reads have higher quality scores for base-calling, it is most likely that this N should be a T.

3.5.1.3.2 In the Contig View toolbar, click the Allow Editing button. Place your cursor to the right of the target base.

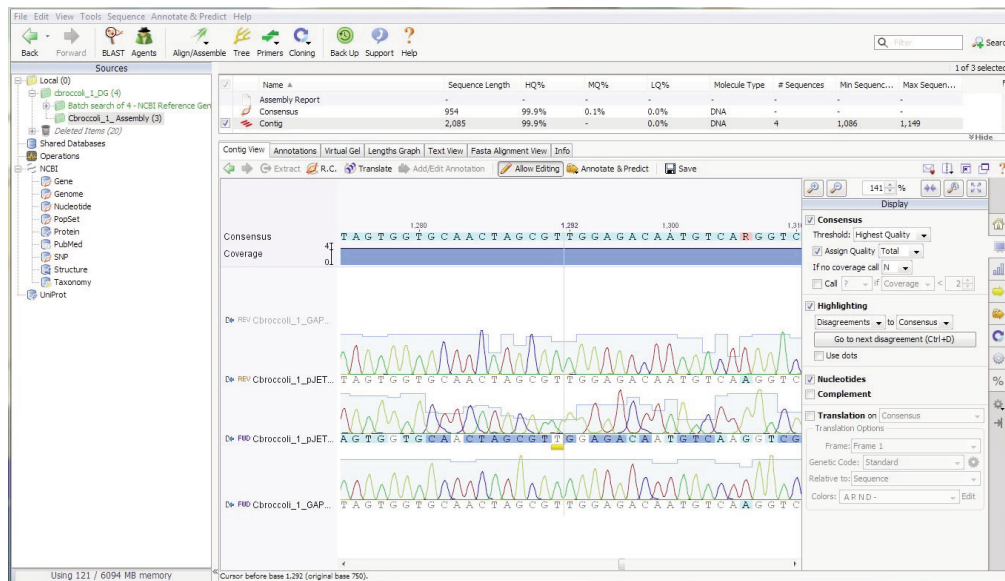


Editing a disagreement. To edit the N to a T in the middle read, click the Allow Editing button in the Contig View toolbar and place your cursor just to the right of the base.

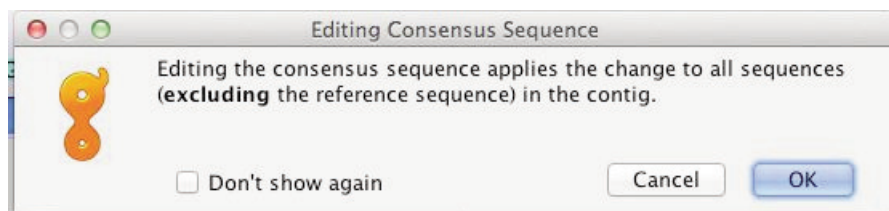
3.5.1.3.3 Click backspace or delete to remove the base. A small red rectangle will appear to mark where you have deleted the base.



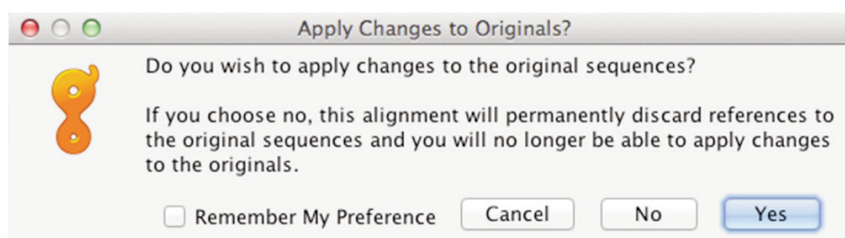
3.5.1.3.4 Type in the new base. A yellow line will appear under the new base call to indicate that has been edited.



If you edit a base in the consensus sequence rather than the single sequence, a dialog box will appear. You can click **OK** to accept the changes.



3.5.1.3.5 Click Save in the Contig View toolbar. A dialog box will appear. Click Yes to save and apply changes:



3.5.1.3.6 The change will also be added to the Annotations and Tracks tab ➡ in the options panel as a replacement.

3.5.2 Reassemble edited reads. If you edited bases in single sequences, you will need to carry out another assembly to get a corrected consensus sequence.

3.5.2.1 Repeat the following steps from Section 3.3 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.

3.5.2.2 Check your new assembly and determine whether there are any disagreements. If there are, repeat the steps for editing a base call (step 3.5.1.2), and continue to iterate until there are no more disagreements. This could take several rounds.

3.5.2.3 Record the name of your finished (fully corrected), final contig and its folder name.

3.5.3 Record the name of your final Assembly folder:

Record the name of your final contig file: _____

Plant name: _____

Clone number: _____

3.6 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?
4. What can you deduce about the orientation of the insert into pJET1.2 from the direction of the assembled fragments? Remember the PCR fragment can be inserted into pJET1.2 in either the forward or reverse orientation.
5. What was the maximum depth of coverage in your assembled sequence and how many bases had this depth of coverage?

3.7 Section 3 Focus Questions.

1. What is a read? How is this different from a sequence?
2. What is a contig?

4. Conduct a BLAST Search on the Contig Sequence to Verify Identity of the Cloned Gene

In Section 2, you did an individual BLAST search on all sequences of your clone to get an idea of which Arabidopsis GAPDH gene each of your sequences most closely resembles. 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 what sequence in the GenBank genomic DNA database most closely resembles the consensus sequence you generated.

IMPORTANT NOTE: Regarding BLAST searches using Geneious:

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 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 2.1 using Geneious or using NCBI's BLAST website. Use Appendix I for protocol steps to export sequences as FASTA files for BLAST searching directly on the NCBI website.

4.1 Use Geneious 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.

- 4.1.1 In your Assembly folder, click to select your final, corrected contig.
- 4.1.2 At the top of the menu bar, select the BLAST icon. A new dialog box will appear.

BLAST

☒ Contig consensus sequence Consensus Options

Query: ☐ Selected sequences (select several to batch search)
☐ Enter unformatted or FASTA sequence

Database: NCBI Reference Genomic Sequences Add/Remove Databases

Program: blastn - similar matches (DNA query, DNA)

Results: Hit table ?

Retrieve: Matching region with annotations (slow)

Maximum Hits: 50

☒ Low Complexity Filter Word Size: 15

☒ Human Repeats Filter Max E-value: 1e-1

☒ Mask for lookup table Gap cost (Open Extend): 5 2

Scoring (Match Mismatch): 2 -3

Other Arguments:

Entrez Query:

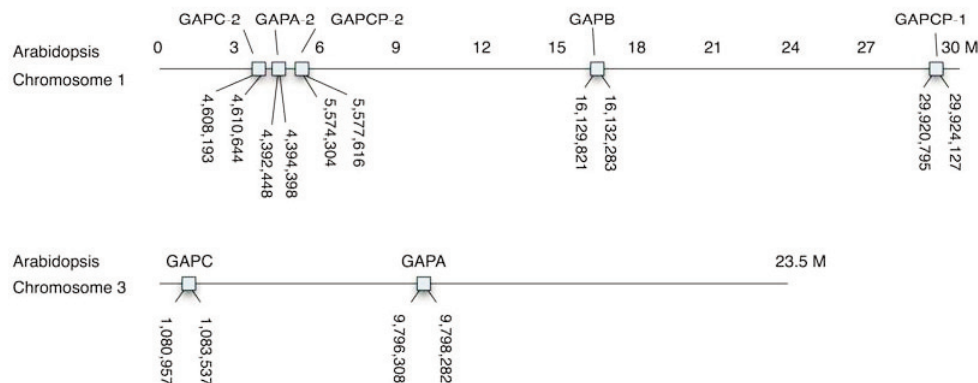
Fewer Options Search Cancel

4.1.2.1 The software will automatically select the Contig Consensus Sequence radio button for Query. Keep this selection.

- Select the **NCBI Reference Genomic Sequences** 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 **More Options**
- Select **15**, or the largest value listed, for Word Size
- Click **Search**. A new subfolder, named "Contig – NCBI Reference Genomic Sequences (refseq_genomic) blastn," will be created to contain your search results

4.2 Analyzing your blastn results.

In this section, you will examine the genes that your contig query sequence matches best and 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 provided below.

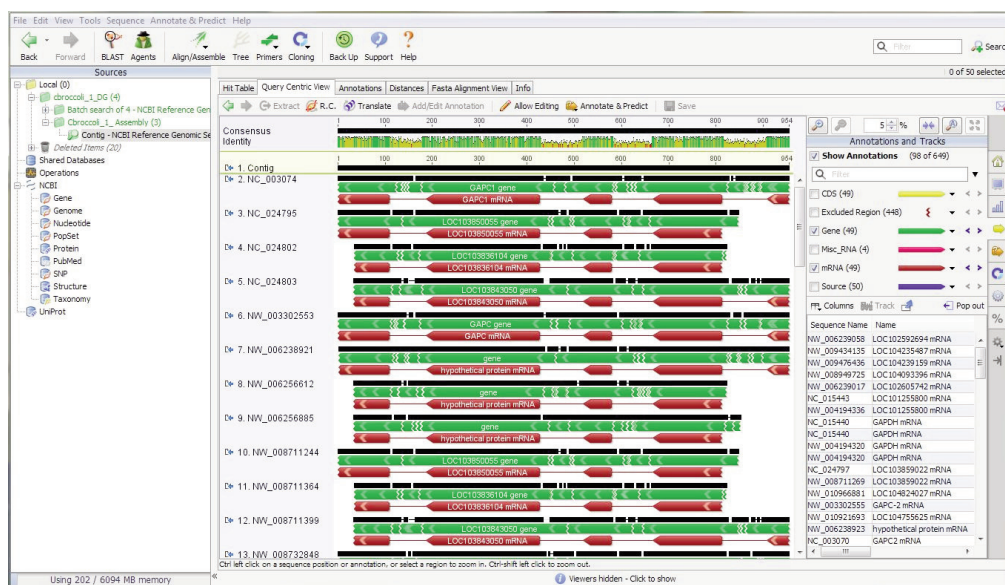


Chromosomal locations of *Arabidopsis* GAPDH genes. Note that GAPC is also known as GAPC-1.

4.2.1 Examine contig query sequence matches.

4.2.1.1 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**.

- **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



Example of BLAST query results using the cbroccoli contig as the query. BLAST results are viewed using Query Centric View.

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 the prepared tables show the four *GAPC* genes (*GAPC*, *GAPC-2*, *GAPC-1*, and *GAPC-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.

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.

- 4.2.1.2 Use the total score calculated from the table to identify the gene that your sequence matches best.
- 4.2.1.3 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.

4.2.1.3.1 The BLAST results from the contig (cbroccoli, in this example) are viewed using the Hit Table, sorted by E-value and Grade.


The screenshot shows the BLAST interface with the Hit Table sorted by E-value and Grade. The table contains the following data:

E Value	Bit-Score	Grade	Name	Query cover...	Hit start	Hit end	Sequence Length	Common Name	Organism	Accession	Max
0	717.224	88.9%	NC_003074	100.00%	1,083,023	1,082,032	1,003	thale cress	Arabidopsis thaliana	NC_003074	992
0	829.033	85.6%	NC_024795	88.89%	28,510,043	28,510,901	874	field mustard	Brassica rapa	NC_024795	859
0	677.55	81.3%	NC_024802	82.60%	19,713,183	19,712,375	821	field mustard	Brassica rapa	NC_024802	809
0	825.426	88.0%	NC_024803	96.54%	34,253,241	34,252,422	965	field mustard	Brassica rapa	NC_024803	921
0	720.83	88.7%	NW_003032553	100.00%	1,368,443	1,367,448	1,008	-	Arabidopsis lyrata	NW_003032553	996
0	713.617	88.3%	NW_006238921	99.90%	1,105,889	1,105,892	1,011	-	Capella rubella	NW_006238921	998
0	650.499	80.4%	NW_006256612	82.70%	10,903,256	10,902,412	852	-	Eutrema saulgneum	NW_006256612	845
0	829.033	84.2%	NW_006256885	95.85%	8,330,110	8,330,969	864	-	Eutrema saulgneum	NW_006256885	860
0	829.033	85.6%	NW_008711244	88.89%	141,539	142,397	874	field mustard	Brassica rapa	NW_008711244	859
0	677.55	81.3%	NW_008711364	82.60%	2,062,056	2,061,248	821	field mustard	Brassica rapa	NW_008711364	809
0	825.426	88.0%	NW_008711399	96.54%	1,073,524	1,072,605	965	field mustard	Brassica rapa	NW_008711399	921
0	888.544	90.0%	NW_008732948	100.00%	134,240	135,243	1,020	field mustard	Brassica rapa	NW_008732948	1,000
0	688.37	88.1%	NW_010921691	99.90%	1,230,397	1,229,405	1,015	false flax	Camelina sativa	NW_010921691	993
0	719.027	88.6%	NW_010921695	99.90%	1,445,086	1,444,093	1,010	false flax	Camelina sativa	NW_010921695	994
1.28e-179	643.286	87.4%	NW_010921677	99.90%	1,246,208	1,245,198	1,036	false flax	Camelina sativa	NW_010921677	1,011
3.68e-142	518.853	67.4%	NC_024797	92.41%	14,728,888	14,728,396	521	field mustard	Brassica rapa	NC_024797	903

A 'Quick document selection' tip is displayed in the center, stating: 'To quickly select several documents, click and drag in the column of tick-boxes on the left of the table.'

BLAST results using cbroccoli contig as the query. The Hit Table is sorted by E-value.

4.2.1.3.2 Using the Organism column, the BLAST hits can be re-sorted to find the *Arabidopsis* chromosome hits from the contig.

Tip: Alternatively, you can also 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.

BLAST results using **cbroccoli contig** as the query. The Hit Table is sorted by Organism in order to find the *Arabidopsis thaliana* hits.

E Value	Bit-Score	Grade	Name	Query cover...	Hit start	Hit end	Sequence Length	Common Name	Organism	Accession	Max
0	720.83	88.7%	NC_00302555	100.00%	1,367,442	1,367,442	1,008	-	Arabidopsis lyrata	NC_00302555	996
1.13e-91	351.14	69.6%	NC_00302555	65.62%	5,639,498	5,640,104	644	-	Arabidopsis lyrata	NC_00302555	626
1.06e-85	331.303	69.1%	NC_003070	65.62%	4,609,091	4,609,715	664	thale cress	Arabidopsis thaliana	NC_003070	626
0	717.224	88.9%	NC_003074	100.00%	1,083,023	1,082,032	1,003	thale cress	Arabidopsis thaliana	NC_003074	992
0	829.033	85.6%	NC_024795	88.89%	28,510,043	28,510,901	874	field mustard	Brassica rapa	NC_024795	899
3.68e-142	518.853	67.4%	NC_024797	52.41%	14,728,868	14,728,366	521	field mustard	Brassica rapa	NC_024797	503
1.13e-91	351.14	58.6%	NC_024797	33.54%	14,733,018	14,732,693	334	field mustard	Brassica rapa	NC_024797	326
8.68e-106	398.027	61.0%	NC_024800	30.92%	4,781,794	4,782,090	297	field mustard	Brassica rapa	NC_024800	297

Selected sequences are only summaries. Download Full Sequence(s).

Alignment View | Annotations | Dotplot | Dotplot (Self) | DNA Fold | Distances | Text View | Download | Fasta Alignment View | Fasta Nucleotide View | Info

Annotations and Tracks (3 of 6)

- ☒ Show Annotations (3 of 6)
- ☒ CDS (2)
- ☒ Gene (1)
- ☒ mRNA (2)
- ☒ Source (1)

Name	Type	Min
GAPC2 mRNA	mRNA	2
GAPC2 mRNA	mRNA	2
GAPC2 gene	gene	<1

Using 158 / 6094 MB memory

4.2.1.3.3

Click to select *Arabidopsis thaliana* chromosome 1. In Alignment View (document viewer), you can see that there are annotations for GAPC-2.

Click in the document table to select *Arabidopsis thaliana* chromosome 1 and click Alignment View in the document viewer. Here, you can see that there are annotations for GAPC-2.

E Value	Bit-Score	Grade	Description	Name	Query cover...	Hit start	Hit end	Sequence Length	Common Name
0	720.83	88.7%	Arabidopsis lyrata subsp. lyrata unplaced genomic scaffold...	NC_00302555	100.00%	1,367,442	1,367,442	1,008	-
1.13e-91	351.14	69.6%	Arabidopsis lyrata subsp. lyrata unplaced genomic scaffold...	NC_00302555	65.62%	5,639,498	5,640,104	644	-
1.06e-85	331.303	69.1%	Arabidopsis thaliana chromosome 1, complete sequence	NC_003070	65.62%	4,609,091	4,609,715	664	thale cr
0	717.224	88.9%	Arabidopsis thaliana chromosome 3, complete sequence	NC_003074	100.00%	1,083,023	1,082,032	1,003	thale cr
0	829.033	85.6%	Brassica rapa cultivar Chifu-401-42 chromosome A1, Brap...	NC_024795	88.89%	28,510,043	28,510,901	874	field mus
3.68e-142	518.853	67.4%	Brassica rapa cultivar Chifu-401-42 chromosome A3, Brap...	NC_024797	52.41%	14,728,868	14,728,366	521	field mus
1.13e-91	351.14	58.6%	Brassica rapa cultivar Chifu-401-42 chromosome A5, Brap...	NC_024797	33.54%	14,733,018	14,732,693	334	field mus
8.68e-106	398.027	61.0%	Brassica rapa cultivar Chifu-401-42 chromosome A6, Brap...	NC_024800	30.92%	4,781,794	4,782,090	297	field mus

Selected sequences are only summaries. Download Full Sequence(s).

Alignment View | Annotations | Dotplot | Dotplot (Self) | DNA Fold | Distances | Text View | Download | Fasta Alignment View | Fasta Nucleotide View | Info

Annotations and Tracks (3 of 6)

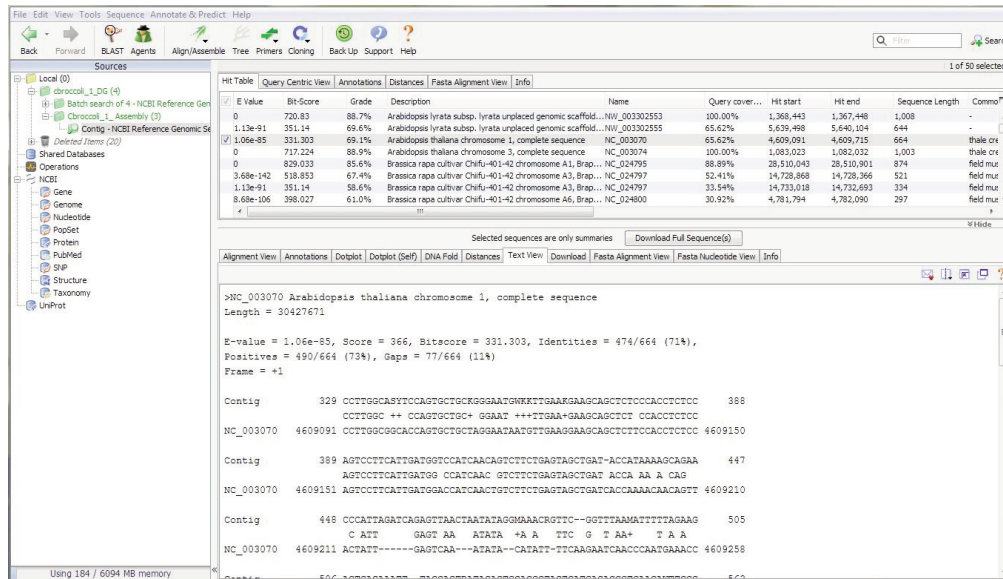
- ☒ Show Annotations (3 of 6)
- ☒ CDS (2)
- ☒ Gene (1)
- ☒ mRNA (2)
- ☒ Source (1)

Name	Type	Min
GAPC2 mRNA	mRNA	2
GAPC2 mRNA	mRNA	2
GAPC2 gene	gene	<1

Using 162 / 6094 MB memory

4.2.1.3.4 Switch to Text View in the document viewer. Use the information provided to fill out the gene tables. For a refresher on the Text View tab, see section 2.3.3.2.

Note: 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. For a refresher on the Text View tab, see section 2.3.3.2.

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-2</i>	1	4,608,193	4,610,644			
Query		4,609,091	4,609,715	329	954	331.303
Total score						331.303

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>	1	1,080,957	1,083,537			
Query		1,083,023	1,082,032	1	954	717.224
Total score						717.224

In the example shown in the table above, one fragment of our cbroccoli query sequence has homology within the *GAPC-2* gene on chromosome 1 and one fragment has homology within the *GAPC* gene (also known as *GAPC-1*) 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 717.224, which is higher than the next-best match (*GAPC-2*), with a score of 331.303.

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						

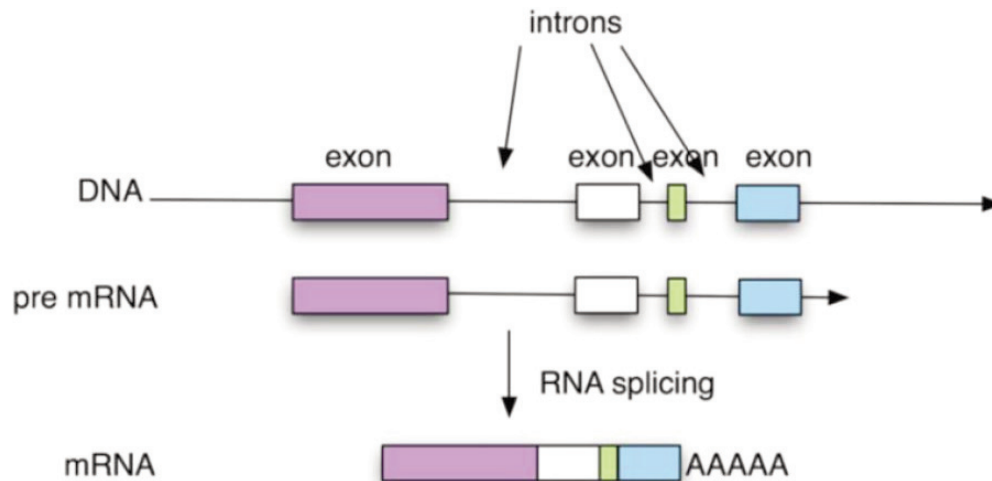
Gene	Arabidopsis Chromosome	Beginning of Arabidopsis Gene Location	End of Arabidopsis Gene Location	Query Begin	Query End	Bit-Score for Query Sequence Match
<i>GAPC</i>	3	1,080,957	1,080,957			
Total score						

4.3 Results Analysis.

Based on your total scores, which *Arabidopsis GAPC* gene does your contig show the highest homology to? Why?

5. Determine Gene Structure (Intron/Exon Boundaries) Using BLAST — Build a Gene Model

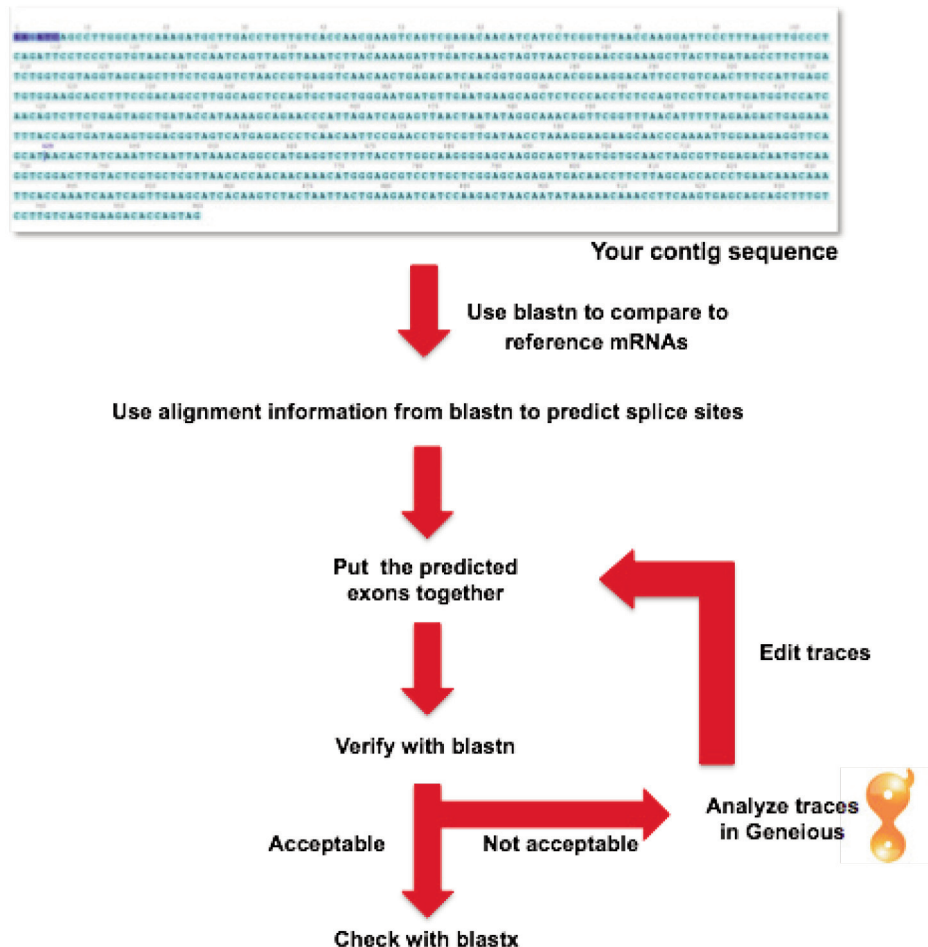
Researchers are often faced with having to build gene models to show where such features as exons and introns exist within the gene. Our goal at this stage of the project is to construct a gene model that shows where the exons are likely to be located within our contig. The process of identifying the 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, you will first compare your contig sequence to the reference mRNA 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.



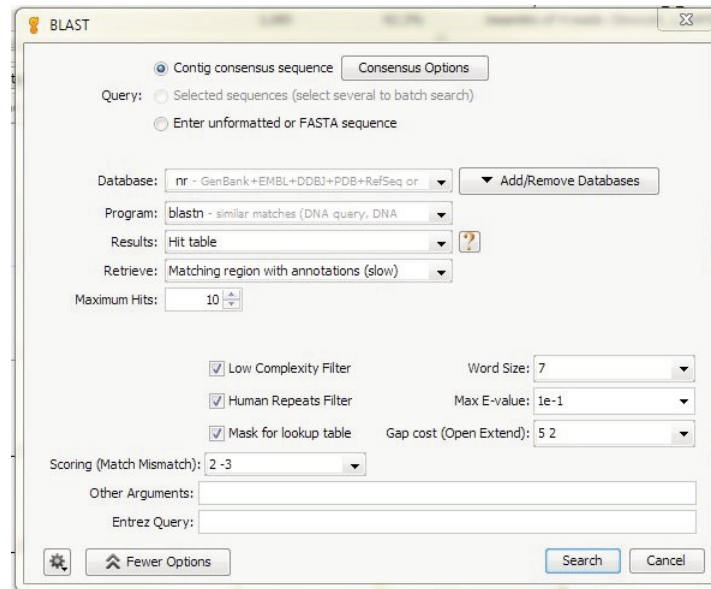
Workflow to determine intron/exon structure and the putative mRNA sequence.

5.1 Using blastn to align the contig to reference mRNA sequences.

For this section, retrieving results from a BLAST search using the Geneious 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 program.

5.1.1 Select your final, corrected contig sequence from your Assembly folder.

5.1.2 Select the BLAST icon from the menu bar. A new dialog box will appear:



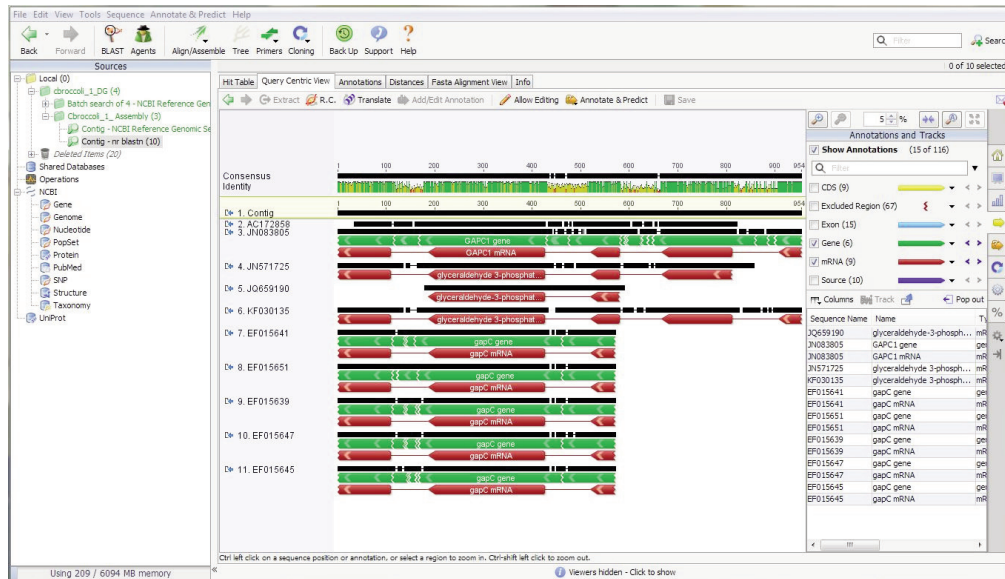
- Select nr as your database




- Keep the default selection **Contig consensus sequence** for Query
- 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**
 - o 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 Contig – nr blastn

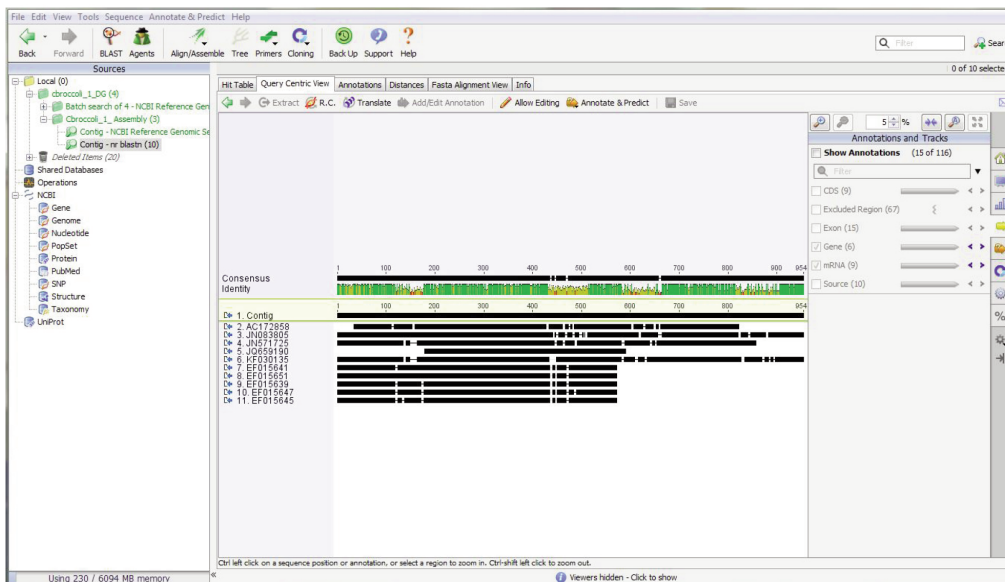
5.2 Interpreting the results and predicting the exon positions.

The Query Centric View will show your consensus and the nucleotide length near the top of the page. Below, you will see the BLAST results alignment showing where portions of the reference mRNA sequences align to your contig.



BLAST results for a cbroccoli contig using the nr database.

5.2.1 To make it easier to see the sequences, hide the annotations by unchecking the Show Annotations box in the Annotations and Tracks tab  in the options panel.



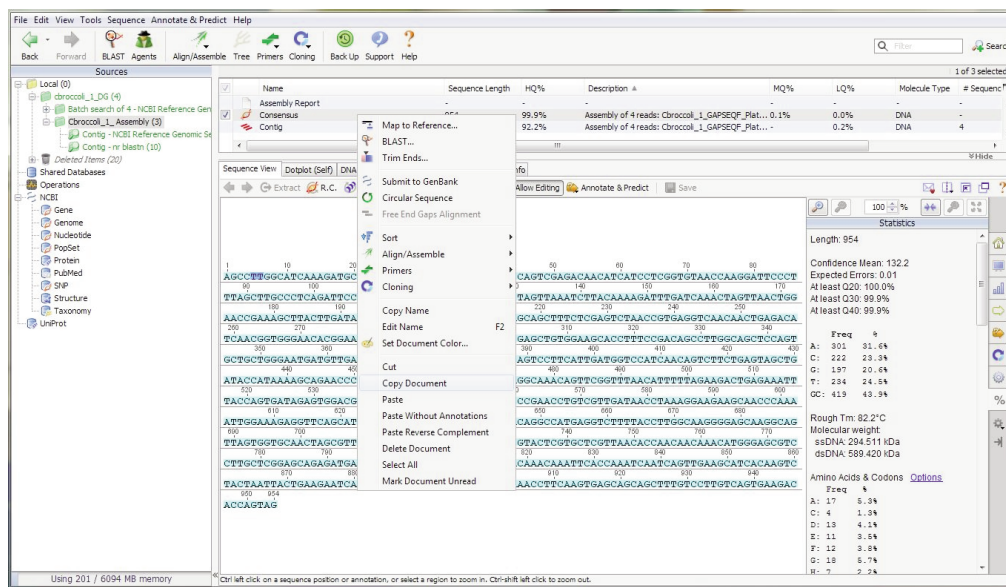
Query Centric View of cbroccoli contig BLAST results with annotations hidden.

5.2.2 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.

Tip: How should the statistics be prioritized to identify the best mRNA match? Recall from Section 2 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.

5.2.3 Align your contig to reference mRNA sequences. The next step is to perform a **multiple sequence alignment**. This way, you will be able to quickly see insertions, deletions, and frame shifts, which the Query Centric View does not reveal. For this step, you will be including your Consensus document with the multiple sequence alignment. This will help you keep track of the base numbering as you mark your intron/exon boundaries.

5.2.3.1 In your Assembly folder, right click your Consensus document to open the shortcut window. Select Copy Document:



Preparing your consensus document for multiple sequence alignment. In your Assembly folder, right click to select the consensus document and select Copy Document in order to paste it into your Contig – nr blastn folder.

5.2.3.2 Navigate to your Contig – nr blastn folder. In the Hit Table, click in a clear space to deselect any documents. Right click in the clear space to bring up a shortcut window. Select Paste. The Consensus document will now be in your Hit Table list (and will have no bit-score or E-value, etc.). The description will read “Assembly of 4 reads: . . .” and the name will read “Consensus: . . .”

The screenshot shows the Geneious software interface. The top menu bar includes File, Edit, View, Tools, Sequence, Annotate & Predict, and Help. Below the menu is a toolbar with icons for Back, Forward, BLAST, Agents, Align/Assemble, Tree, Primers, Cloning, Back Up, Support, and Help. The main window is divided into several panes. On the left is a 'Sources' pane showing a project tree with 'Local (0)', 'Batch search of 4 - NCBI Reference Genomes', 'Chroccoli_1_Assembly (3)', 'Contig - nr blastn (12)', and 'Deleted Items (20)'. The central pane displays a 'Hit Table' with columns for E Value, Bit-Score, Grade, Description, Name, Query cover..., Hit start, Hit end, Sequence Length, and Comment. The table lists several hits, including 'Assembly of 4 reads: Chroccoli_1_GAPSEQ_Plate1...', 'Brassica oleracea glyceroldehyde 3-phosphate dehydrogenase', and 'Pachyadon chesemani voucher CHR 559139 glyceroldehyde 3-phosphate dehydrogenase'. Below the Hit Table is an 'Alignment View' showing a consensus sequence alignment. The alignment is displayed as a grid of colored blocks representing different sequences. The bottom status bar indicates 'Using 292 / 6094 MB memory' and 'Cursor before base 114 in 11 sequences'.

The consensus document is now pasted into your Contig – nr blastn folder for multiple sequence alignment and intron/exon mapping.

5.2.3.3 Select all ten of the query results and your Consensus document in the Hit Table.

5.2.3.4 Click Align/Assemble in the menu bar, and then select Multiple Align.

The screenshot shows the Geneious software interface with a sequence alignment view. The top menu bar includes File, Edit, View, Sequence, Annotate & Predict, and Help. Below the menu is a toolbar with icons for Extract, R.C., Translate, Add Annotation, Allow Editing, Annotate & Predict, Primer Design, and Save. The main window displays a sequence alignment view with a consensus sequence at the top and multiple individual sequences below it. The sequences are aligned to a common reference sequence. The alignment is displayed as a grid of colored blocks representing different sequences. The bottom status bar indicates 'Selected 113 bases from base 1 to 113'.

Run a multiple sequence alignment. Select all ten BLAST results and the consensus document to run a multiple alignment using the Align/Assemble icon in the menu bar.

A new dialog box will appear:



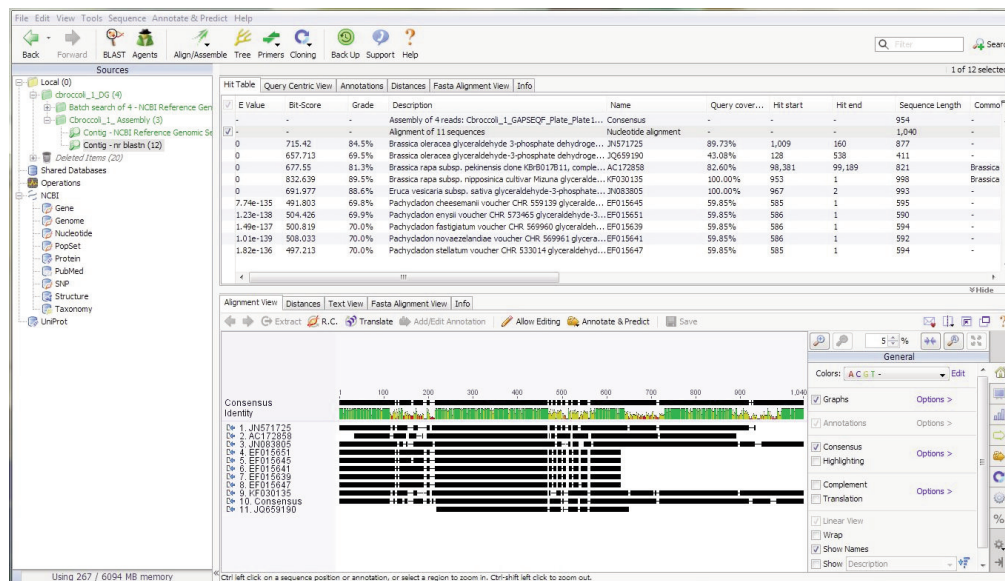
Tip: If you don't see the Multiple Align option, make sure your file is saved by clicking **Save** in the Sequence View toolbar.

5.2.3.5 Select **MUSCLE Alignment**.

5.2.3.6 Use the default **Maximum number of iterations** of 8.

5.2.3.7 Click **OK**. A new document named "Alignment of 11 sequences" will appear in the Hit Table containing your results.


5.2.3.8 Click to view the "Alignment of 11 sequences" document. In Alignment View, you will see how the ten query hits and your consensus align to your contig.




The multiple sequence alignment with your contig and BLAST hits are displayed in Alignment View.

NOTE:

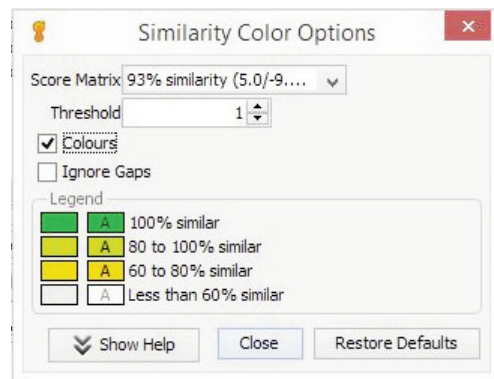
This is a convenient step at which to stop, if necessary.

Your data should already be saved on the local hard drive of your computer. If you wish, you may back up your data using the **Back Up** button  on the menu bar, and your data will be saved in a zipped folder in the location of your choice on your computer. Resume work during your next laboratory period.

5.2.3.9 To view the % similarity of the multiple sequence alignment, change the base coloring of the sequences.

5.2.3.9.1 In the General tab  in the options panel, click the dropdown list for Colors. Select the gray-scale icon for Similarity.

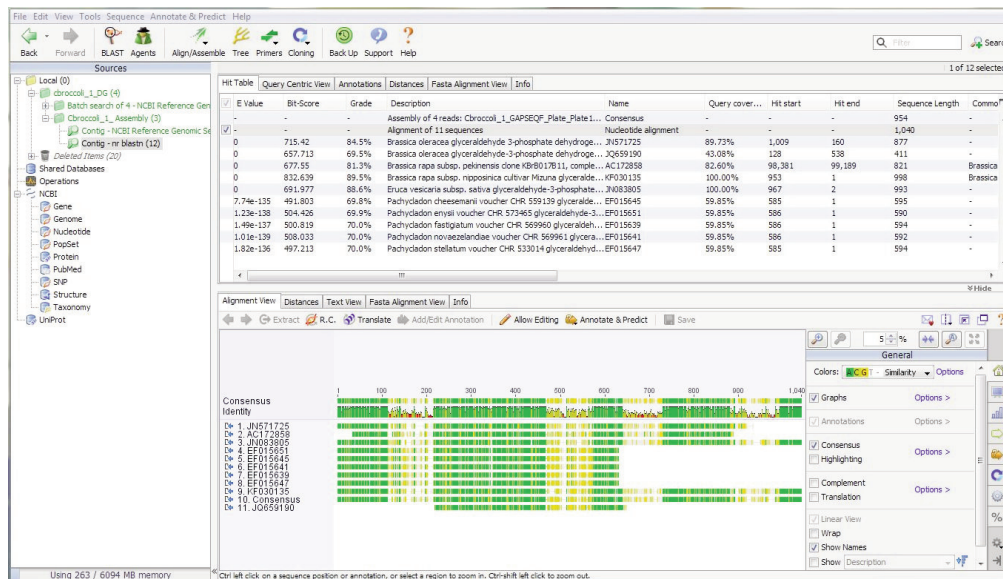
5.2.3.9.2 Clicking the blue Options link next to the Colors dropdown list will open a Similarity Color Options window with a legend for % similarity that corresponds to each gray level. Checking the Colours box will give you a color-scale rather than a gray-scale representation of similarity.



The similarity color options. Color options representing levels of similarity are accessed in the General tab of the options panel, where you can set the threshold level for viewing 93% similarity between alignments.

5.2.3.9.3 In the Similarity Color Options window, select 93% similarity for the Score Matrix.

With these settings, the ends of each 93% similarity alignment, which are green in the current color scheme, correspond approximately to the ends of each exon. You can predict from the example shown below that there are about five exons.



Viewing the % similarity of the multiple sequence alignment. With the color settings to represent similarity turned on, you can view the percentage of similarity across the sequence alignments. Green segments represent regions of at least 93% similarity.

5.3 Find the intron/exon boundaries and mark them.

Now you will look for more precise intron/exon boundaries and mark their positions in the contig. This will involve navigating through the Query Centric View of your contig and looking at query search results and your multiple sequence alignment at the same time, and then modifying your consensus sequence in a new window.

For these steps, you will find and work with each exon in turn. For each exon, you will locate its beginning to its 3' end in the consensus and mark it by adding annotations to the consensus sequence itself. When you are through, each exon will have an annotation, making it easier to pick out those regions of sequence.

5.3.1 In your Contig – nr blastn folder, double click to open your Consensus document in new window.

5.3.2 In your Contig – nr blastn folder, click to select the “Alignment of 11 sequences” document in the Hit Table. Click Alignment View to see the multiple sequence alignment.

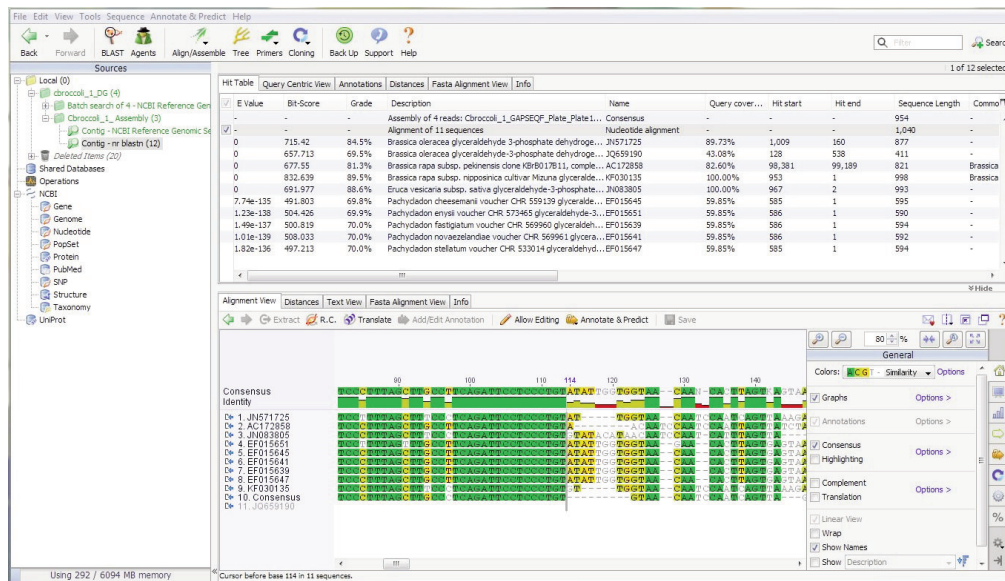
The screenshot displays the NCBI BLAST interface. The top panel shows the 'Hit Table' with columns for E Value, Bit-Score, Grade, Description, Name, Query cover..., Hit start, Hit end, Sequence Length, and Comment. The bottom panel shows the 'Alignment View' with a multiple sequence alignment of 11 sequences. The alignment is color-coded by similarity, with green indicating high similarity and yellow indicating lower similarity. The bottom of the alignment view shows the consensus sequence and its identity to the individual sequences.

E Value	Bit-Score	Grade	Description	Name	Query cover...	Hit start	Hit end	Sequence Length	Comment
-	-	-	Assembly of 4 reads: Chirocoli_1_GAPSEQF_Plate1...	Consensus	-	-	-	954	-
-	-	-	Alignment of 11 sequences	Nucleotide alignment	-	-	-	1,040	-
0	715.42	84.5%	Brassica oleracea glyceraldehyde 3-phosphate dehydroge...	J0571725	89.73%	1,009	160	877	-
0	657.713	69.5%	Brassica oleracea glyceraldehyde 3-phosphate dehydroge...	J0559190	43.08%	128	538	411	-
0	677.55	81.3%	Brassica rapa subsp. pekinensis clone BR601781.1, comple...	AC132858	82.66%	98,381	99,189	821	Brassica
0	832.639	89.5%	Brassica rapa subsp. nipposinica cultivar Mizuna glyceralde...	KF030135	100.00%	953	1	998	Brassica
0	691.977	88.6%	Eruca vesicaria subsp. sativa glyceraldehyde-3-phosphate...	J0603805	100.00%	967	2	993	-
7.74e-139	491.803	69.8%	Pachydodon chesemani voucher CHR 559129 glyceralde...	EF015645	59.85%	585	1	595	-
1.23e-138	504.426	69.9%	Pachydodon emys voucher CHR 573465 glyceraldehyde-3...	EF015651	59.85%	586	1	590	-
1.49e-137	500.819	70.0%	Pachydodon fastigiatum voucher CHR 569960 glyceraldeh...	EF015639	59.85%	586	1	594	-
1.01e-139	508.033	70.0%	Pachydodon novaezealandiae voucher CHR 569961 glyce...	EF015641	59.85%	586	1	592	-
1.82e-136	497.213	70.0%	Pachydodon stielum voucher CHR 533014 glyceraldehyd...	EF015647	59.85%	585	1	594	-

Viewing the multiple sequence alignment in Alignment View.

Note: You may have to reset the color similarity scheme and the threshold to 93% from the General tab in the options panel.

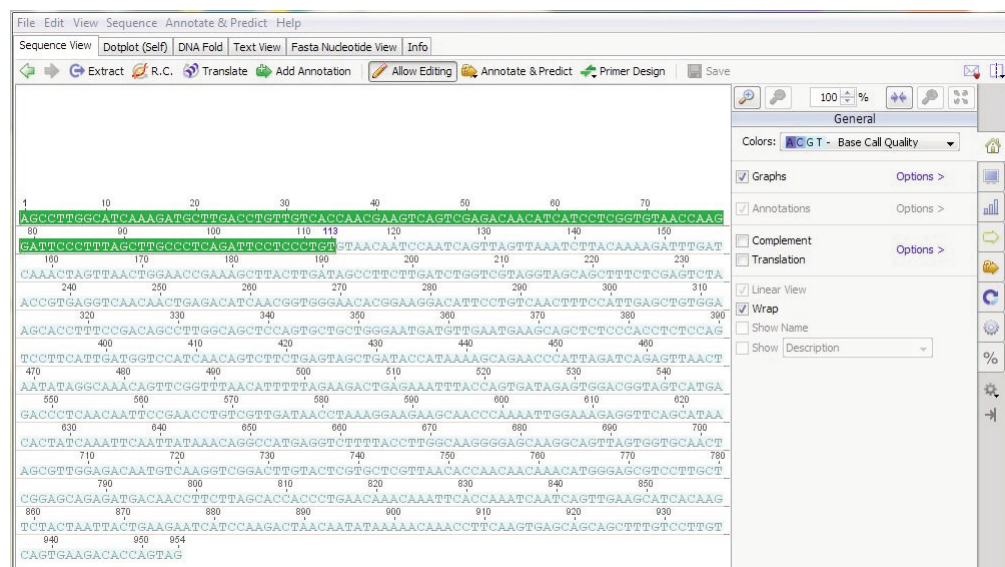
5.3.3 You are now looking for long stretches of green segments, which signify exons you need to mark. In Alignment View, start at the beginning of the sequences to identify the first green segment. Note the base number position in the bottom of the Alignment View window that marks the cursor position. In this example, the first segment spans from base 1 to 113. Zoom in from the options panel to see the exact base.



Mapping intron/exon boundaries. Using the cbroccoli contig as the example, the orange arrow points to a potential intron/exon boundary at base 113. Keep in mind that your contig is represented by the consensus sequence in the list of queries (#10), not the Consensus at the top of the Alignment View, above the Identity chart.

IMPORTANT NOTE: The Consensus in the list of queries in Alignment View (sequence #10 in the above example) is your actual consensus sequence. Don't be confused by the Consensus at the very top that is just above the Identity chart — that is the consensus of the alignment of the query results, not your contig!

5.3.4 In your Consensus window, click to highlight the bases corresponding to your first green segment. Double-check to make sure that the base numbering is in agreement from window to window.

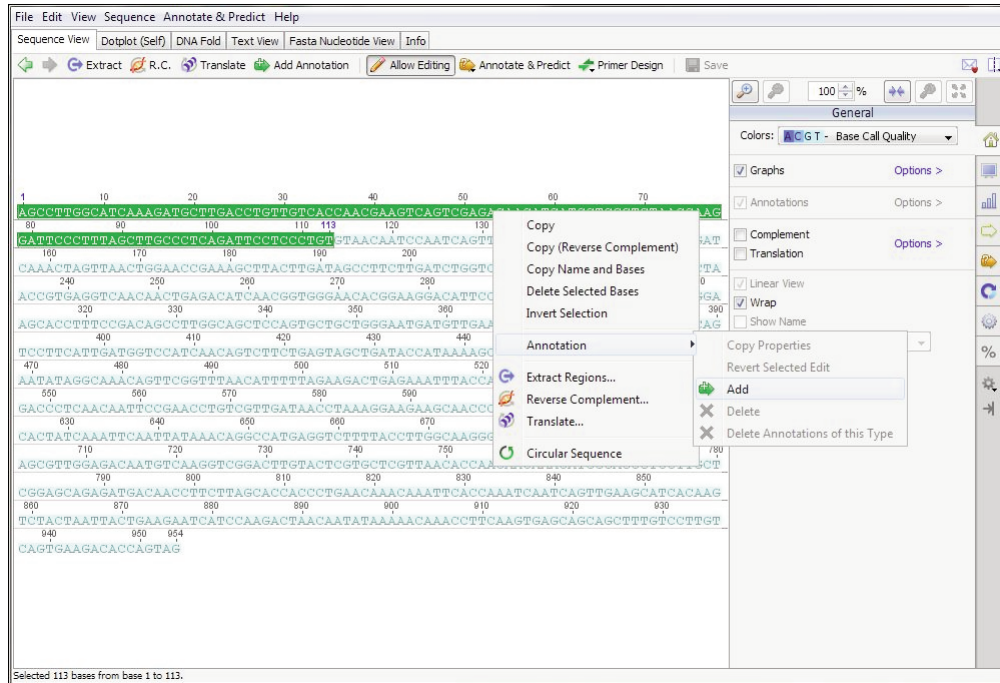


In your consensus document, highlight the bases that correspond to the first green segment of sequence from 5.3.4 (bases 1–113).

5.3.5 Annotate this segment of your consensus as an exon.

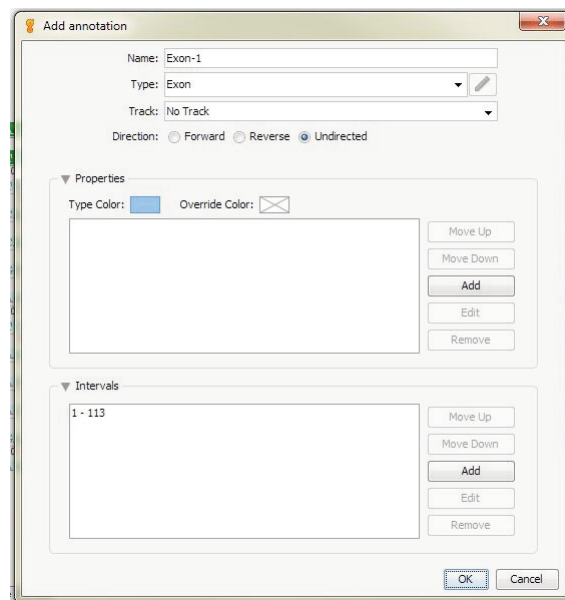
5.3.5.1 Right click the highlighted segment.

5.3.5.2 From the shortcut window, select **Annotation**, and then select **Add**.



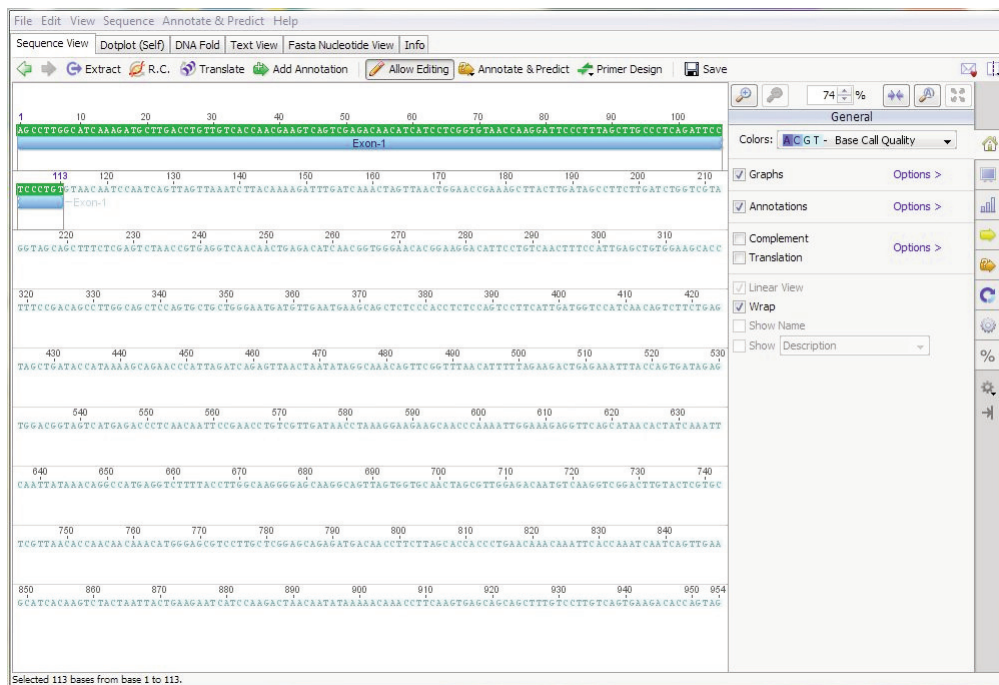
Annotating exons. Right click your highlighted sequence and select **Annotation**, then **Add** to open a new dialog box where you can enter annotation details.

5.3.5.3 A new dialog box will appear called “Add annotation.” Enter an annotation name(such as Exon-1) and select an existing type. Each annotation type will have a default color assigned to it.



Tip: Use a naming convention for your annotations to help you keep track of your work. For example, appending with numbers (that is, Exon-1, Exon-2, etc.) can be helpful to identify multiple segments within your consensus.

- 5.3.5.3.1 Track** — keep the default No Track to put the annotation directly onto the sequence.
- 5.3.5.3.2 Direction** — select Undirected. Currently, you do not know which is the correct direction for your sequence.
- 5.3.5.3.3 Properties** — expand the section to enter additional properties for that annotation. Here, you can choose the color of the annotation or click Add to type in extra information.
- 5.3.5.3.4 Intervals** — lists the bases that your annotations span. Adjust these numbers if they look incorrect. Or you can add an interval or mark the annotation as truncated at the 5' or 3' end.



Exon annotation. The first 113 bases of the consensus sequence have successfully been annotated as Exon-1.

- 5.3.5.4** Go along the sequences in Alignment View and keep creating annotations for your exons until you reach the end.

IMPORTANT NOTE: Keep in mind that the **base numbering will change** and will no longer match up with other sequences. This is because there are different numbers of bases in different sequences between your contig and the query results. See below for an example.

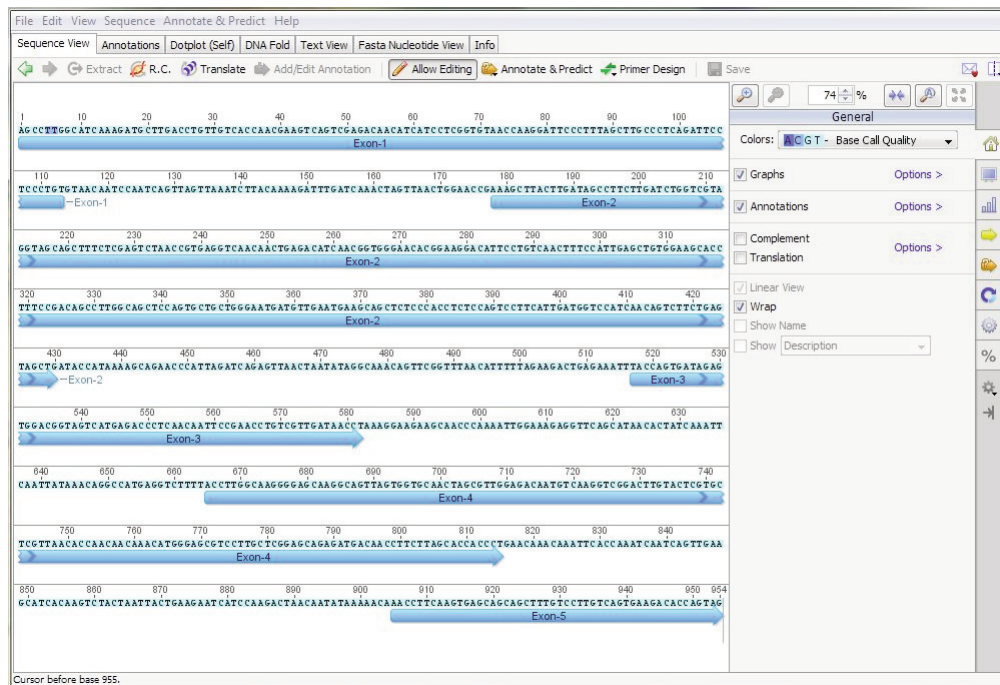
The screenshot shows the Geneious software interface. The top panel displays a 'Hit Table' with columns for E Value, Bit-Score, Grade, Description, Name, Query cover..., Hit start, Hit end, Sequence Length, and Comment. The bottom panel shows the 'Alignment View' with a consensus sequence at the top and individual sequences below it. The status bar at the bottom indicates 'Using 190 / 6094 MB memory' and 'Selected base 468 (original base 430)'.

Hit Table	Query Centric View	Annotations	Distances	Fasta Alignment View	Info				
E Value	Bit-Score	Grade	Description	Name	Query cover...	Hit start	Hit end	Sequence Length	Comment
-	-	-	Assembly of 4 reads: Cbroccoli_1_GAPSEQ_Plate_Plate1...	Consensus	-	-	-	954	-
-	-	-	Alignment of 11 sequences	Nucleotide alignment	-	-	-	1,040	-
0	715.42	84.5%	Brassica oleracea glyceroldehyde 3-phosphate dehydroge...	J0571725	89.73%	1,009	160	877	-
0	657.713	69.5%	Brassica oleracea glyceroldehyde 3-phosphate dehydroge...	J0559190	43.08%	128	538	411	-
0	677.55	81.3%	Brassica rapa subsp. pekinensis clone IFR6017811, comple...	AC172858	82.60%	96,381	99,189	821	Brassica
0	832.639	89.5%	Brassica rapa subsp. nipponensis cultivar Mauna glycerolde...	EF030135	100.00%	953	1	998	Brassica
0	691.977	88.6%	Eruca vesicaria subsp. sativa glyceroldehyde 3-phosphate...	J0683805	100.00%	967	2	993	-
7.74e-135	491.803	69.8%	Pachyadon chesemani voucher CHR 559139 glycerolalde...	EF015645	59.85%	585	1	595	-
1.23e-136	504.426	69.9%	Pachyadon ernst voucher CHR 573465 glyceroldehyde-3...	EF015651	59.85%	586	1	590	-
1.49e-137	500.819	70.0%	Pachyadon festigatum voucher CHR 569960 glycerolalde...	EF015639	59.85%	586	1	594	-
1.01e-139	508.033	70.0%	Pachyadon novaezealandiae voucher CHR 569961 glycerol...	EF015641	59.85%	586	1	592	-
1.82e-136	497.213	70.0%	Pachyadon stellatum voucher CHR 533014 glyceroldehyd...	EF015647	59.85%	585	1	594	-

Keep track of the base numbering. Since there are different numbers of bases in the sequences of your contig and those of the query results, the base numbering between different sequences may change. Keep your eye on the text at the lower left of your screen. In this example it says “Selected base 468,” which refers to overall consensus of the multiple alignment (listed at the top, above the Identity chart). The “(original base 430)” refers to the base in your contig/consensus.

Rely on the base numbering at the bottom of the window (circled in orange). In this example it tells you the cursor has “Selected Base 468,” which corresponds to the base numbering for the overall consensus of the alignment. The text in parentheses, “(original base is 430),” tells you that base 468 in this multiple alignment refers to base 430 in your contig/consensus. Therefore, mark the end of this exon segment as base 430 in your consensus window.

5.3.5.5 When you are through annotating your Consensus, save your changes by navigating to the File tab in the menu bar of the Geneious window and select **Save As** in the dropdown list. Choose a name to distinguish your document from those of other student groups. For example, “cbroccoli_1_consensus_DG.” Your sequence will look something like this:



Annotating the consensus sequence. All the green segments of high-identity sequence from the multiple sequence alignment for step 5.3.6.4 have been annotated on the cbroccoli consensus sequence.

5.4 Identifying and correcting errors in your reads and contig sequence based on mRNA alignment.

Look through your sequence. Are there any insertions or deletions relative to all the matched sequences? In examining the alignments, you may notice some differences between your contig sequence and the mRNAs that you retrieved from the BLAST results. These differences could result from sequencing, assembly, or base-calling errors.

Before going further, you will need to review the traces that contain these sequences and determine whether there are errors in the contig sequence. This involves identifying the reads and using Geneious to examine the trace. You should also be concerned with indels (insertion/deletions), since these can affect the reading frame of the gene model.

- 5.4.1 Identify the regions that display variations from your contig in Alignment View for your “Alignment of 11 sequences” document.
- 5.4.2 Navigate to your contig in your Assembly folder. Double check these base regions by looking at the traces of the single-sequence chromatograms.
 - 5.4.2.1 Determine whether any of your regions are indels and need editing. If so, click **Allow Editing** in the Sequence View toolbar and edit the bases as described earlier in section 3.5.

- 5.4.3 After checking through your contig, if you have made changes to the consensus, you can make the same changes in your annotated consensus document. If you wish, you can save this corrected annotated consensus document with a new name. Record the file name if necessary:

_____.

5.5 Check initial gene model (proposed 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.

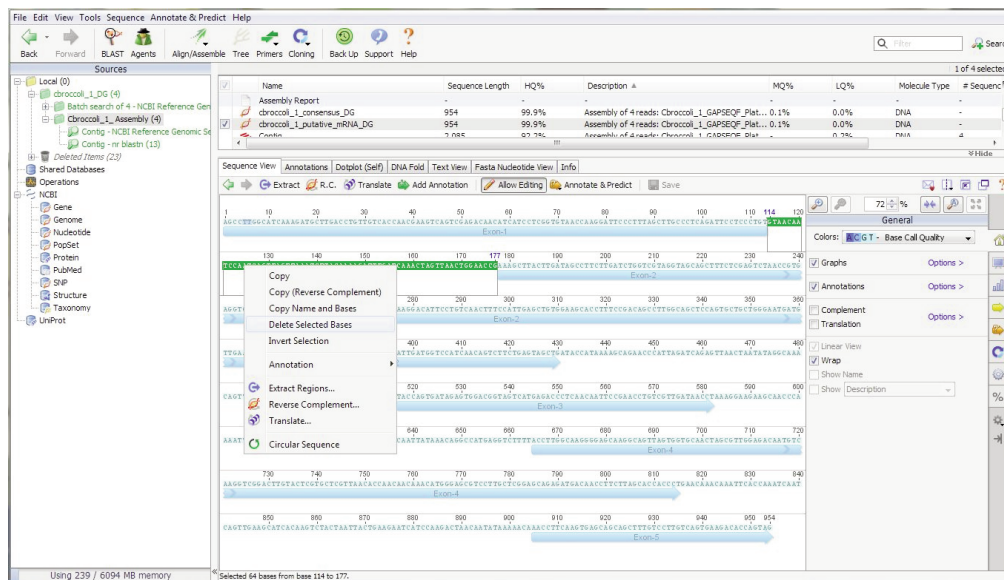
5.5.1 To create a document with only your exon sequences as putative mRNA, go to your Assembly folder and select your corrected annotated consensus sequence document from step 5.3.6.5.

5.5.2 Navigate to the File tab at the top of the Geneious window and select **Save As**. Create a new name for this file. For example, cbroccoli_1_putative_mRNA_DG, and click **OK**.

5.5.3 To delete the non-exon bases, click Allow Editing in the Sequence View toolbar.

5.5.3.1 Select each region of the non-exon sequences. Be sure to click on the sequence itself rather than the exon annotation. Otherwise, you will drag and extend the annotation rather than highlight non-exon regions. If you accidentally drag the exon annotation, go to **Edit** in the main toolbar and select **Undo**.

5.5.3.2 Once the sequences are highlighted, right click and select **Delete Selected Bases**.

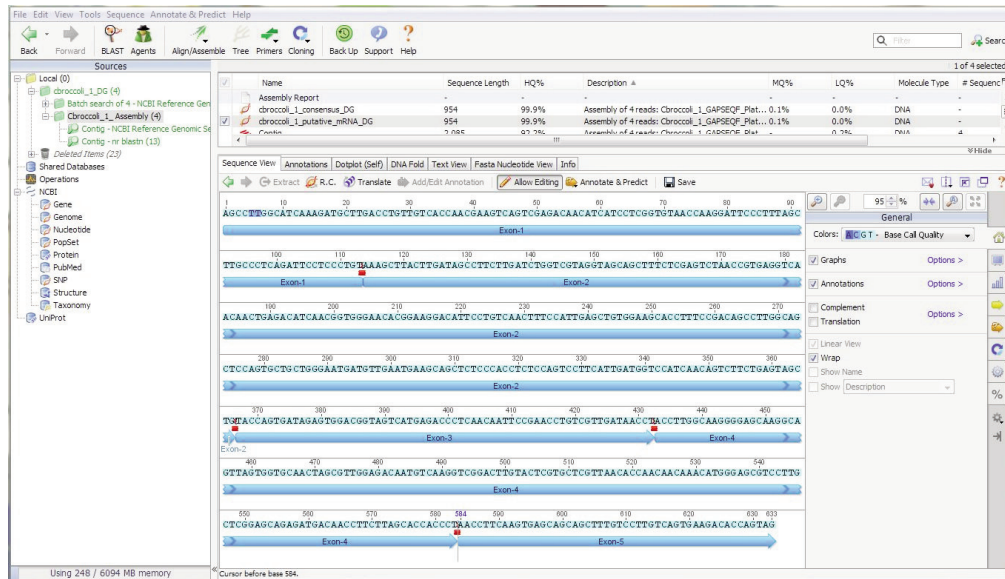


Creating a putative mRNA sequence document. Non-exon bases are deleted from the annotated consensus sequence. Here, bases 114–177 are highlighted for deletion.

5.5.3.3 The bases will now be gone, and a placeholder will appear in the sequence as a red, squiggly line with a red rectangle below to indicate the region where bases were removed:



5.5.3.5 When you are done, click **Allow Editing** again to turn it off, and then click **Save**. Your final sequence should look something like this:



5.5.4 Run a blastn search on your putative mRNA sequence.

5.5.4.2 Click the BLAST icon in the menu bar. Keep the defaults from the previous search, namely:

- Database should be **nr**
- 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 “- nr blastn” appended to the end as the name of the folder. For example, `cbroccoli_1_putative_mRNA_DG - nr blastn`.

Running a BLAST search on your putative mRNA sequence using the nr database. For your own BLAST search, fill in the BLAST search parameters as shown in the dialog box above.

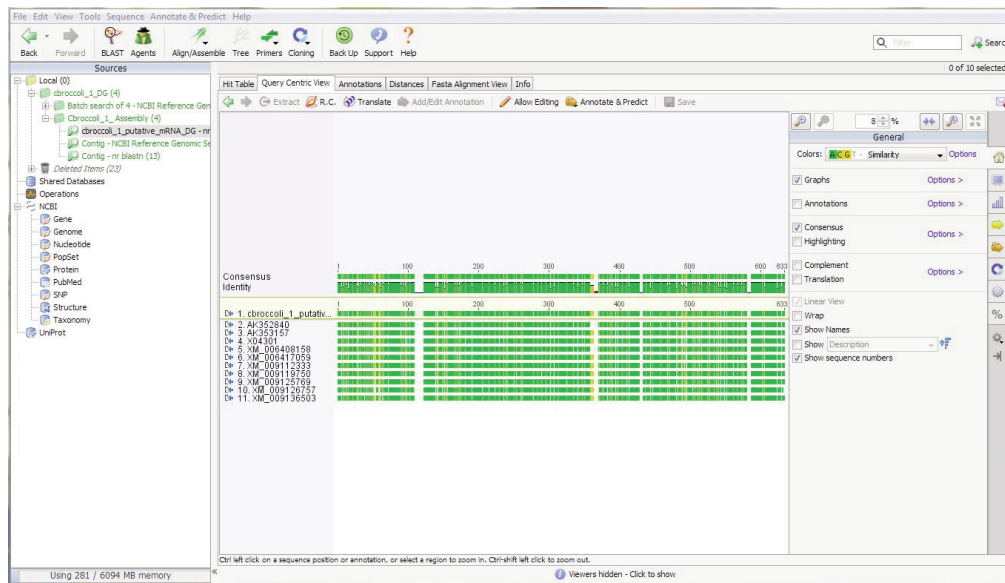
5.6 Interpreting the results and finding and resolving errors.

You should now see your sequence and the ten best fits from the blastn search aligned. 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. Ideally, there should be a high level of query coverage and a low E-value; the % 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 is still a reasonable amount of variation between plants in the coding regions.

When you did a blastn search in Section 4.1 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 proposed mRNA matches several subject sequences along the entire length with very few gaps.

Your results may be similar, or you may find breaks in the sequence where a portion of sequence is missing or differs between plant species. You will now need to examine your results in further detail, since it is likely that the regions where two exons are joined have errors. There may also be errors that were missed on the first run-through.

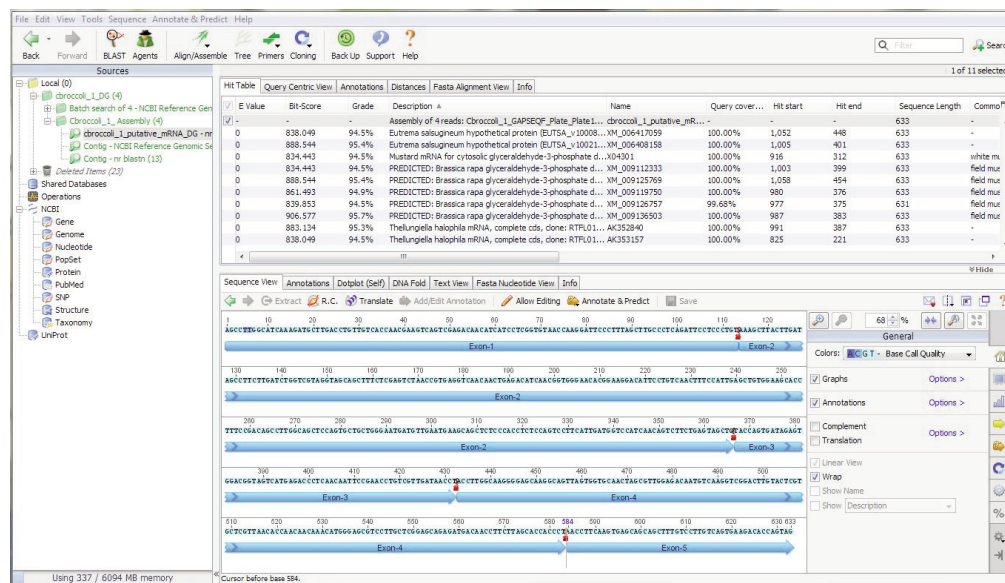
- 5.6.1** Look through your sequence. Are there any indels relative to all the matched sequences? If so, there may be an issue with your assigned splice locations. View your putative mRNA and blastn results in the Hit Table and Query Centric View to see if there are any gaps.



Query Centric View of query results from a blastn search using the cbroccoli putative mRNA sequence document.

5.6.2 Perform another multiple sequence alignment. This will help you find errors that have been introduced in your alignment. For example, the alignment may have incorporated gaps where none existed in the original sequence. If so, you will need to manually correct the alignment.

5.6.2.1 Copy and paste your putative mRNA document from your Assembly folder into your putative mRNA blastn folder.



Copy and paste your putative mRNA sequence document into your putative mRNA - nr blast folder.

5.6.2.2 In the Hit Table, select all ten of the query results and your putative mRNA sequence. From the menu bar, click **Align/Assemble** and select **Multiple Align**. A new dialog box will appear.

5.6.2.2.1 Select the **MUSCLE Alignment** option.

5.6.2.2.2 Set the Maximum number of Iterations to **8**.

5.6.2.2.3 Click **OK**.

A new document called “Alignment of 11 sequences” will contain your results and appear in the Hit Table. Select this document to view the alignments.

The screenshot shows the Geneious software interface. The top panel is the Hit Table, which lists search results. The bottom panel is the Alignment View, showing a multiple sequence alignment of 11 sequences. The alignment is color-coded by base (A, C, G, T) and includes a gene model with exons and introns. The right sidebar contains options for General, Graphs, Annotations, Consensus, Highlighting, Complement, Translation, Linear View, Wrap, Show Names, Show Description, and Show sequence numbers.

Multiple sequence alignment of putative mRNA sequence with BLAST results. Since there are still some gaps in the alignment, the gene model will need further refinement to eliminate any erroneous gaps or indels.

5.6.2.3 Scroll through the aligned sequences and look for possible errors. The errors we are concerned with are indels or gaps in the alignment. Notice where these bases are located; most likely they are located at the intersection of two hypothetical exon sequences (marked with the red rectangle where bases were deleted). These can result from improperly joined exons, from sequencing errors, or from errors introduced by the alignment algorithm.

5.6.2.4 Fix these errors by editing the sequence itself. You can edit the putative mRNA directly in this “Alignment of 11 sequences” document.

5.6.2.4.1 Select Allow Editing, and then click on the area where you need to make the correction. For example, if there are bases that need to be removed, you can edit the exon annotations to match the boundaries and delete the extra bases. See the example below:

The screenshot shows the Geneious software interface. The top menu bar includes 'File', 'Edit', 'View', 'Tools', 'Sequence', 'Annotate & Predict', and 'Help'. The 'Edit' menu is open, and the 'Allow Editing' option is highlighted. The main window displays the 'Hit Table' and 'Alignment View'. The 'Hit Table' shows a list of sequences with columns for 'E Value', 'Bit-Score', 'Grade', 'Description', 'Name', 'Query cover...', 'Hit start', 'Hit end', 'Sequence Length', and 'Comment'. The 'Alignment View' shows a consensus sequence with annotations for Exon-1 and Exon-2. The 'Allow Editing' button is highlighted in the top menu bar.

Refining the gene model. There are extra bases within the exon that were not removed during the first intron/exon mapping between the hypothetical Exon 2 and Exon 3. To delete these bases from your putative mRNA sequence, click to highlight your sequence file and enable editing by clicking the Allow Editing button on the Alignment View menu bar. Highlight then delete the bases that need to be removed.

5.6.2.4.2 Click the exon annotations and drag them to match the ends from the alignment.

The screenshot shows the Geneious software interface. The top menu bar includes 'File', 'Edit', 'View', 'Tools', 'Sequence', 'Annotate & Predict', and 'Help'. The 'Edit' menu is open, and the 'Allow Editing' option is highlighted. The main window displays the 'Hit Table' and 'Alignment View'. The 'Hit Table' shows a list of sequences with columns for 'E Value', 'Bit-Score', 'Grade', 'Description', 'Name', 'Query cover...', 'Hit start', 'Hit end', 'Sequence Length', and 'Comment'. The 'Alignment View' shows a consensus sequence with annotations for Exon-1 and Exon-2. A tooltip is visible over the Exon-2 annotation, showing details like 'Name: Exon-2', 'Type: Exon (Created by User)', 'Length: 253', 'Interval: 114 - 366', and 'Bases: AAAGCTTACTGATAGCTTC...'. The 'Allow Editing' button is highlighted in the top menu bar.

Refining the gene model. Modify the exon annotations by clicking on the annotations themselves and dragging them to match the ends of the sequence from the alignments.

5.6.2.4.3 Next, highlight the bases to be removed from your putative mRNA sequence.

The screenshot shows the Geneious software interface. The 'Alignment View' is active, displaying multiple sequence alignments. A region of the sequence is highlighted in red, indicating bases to be removed. The status bar at the bottom indicates 'Selected 13 bases from base 109 to 121 (13 unpaired bases from base 109 to 121, Mouse over base 109 (C - quality 125), residue 37 (P/Pro/Prolina) in cbroccol_1_putative_mRNA_DG'.

Refining the gene model. From the putative mRNA sequence, highlight the sequences to be removed for deletion.

5.6.2.4.4 Right click and select Delete selected bases. Or click the Delete or Backspace key on your keyboard. The bases will now turn into dashes, indicating a gap.

The screenshot shows the Geneious software interface after deleting the selected bases. The alignment view displays multiple sequence alignments with a consensus sequence at the top. The region that was previously highlighted is now shown as dashes, indicating a gap. The status bar at the bottom indicates 'Selected 13 bases from base 109 to 121 (0 unpaired bases before base 109)'.

Refining the gene model. Once the extraneous bases are deleted from the putative mRNA sequence, the highlighted bases turn into dashes, indicating gaps.

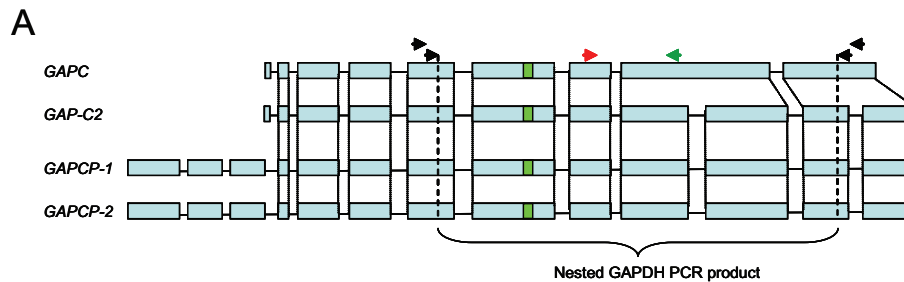
5.6.2.5 Continue scanning the alignments until you have examined and resolved all possible errors. When you have finished fixing errors, click **Save** in the Alignment View toolbar to save your work.

5.6.2.6 Carry out another blastn search with your refined mRNA model. Your new folder containing the second round of query results will be a subfolder within your original “putative_mRNA – nr blastn” folder. If there are indels or gaps, resolve them as before. It may take two to three rounds of BLAST searching and fixing before you have found and resolved all the errors. When your putative mRNA sequence has no more errors relative to the majority of the blastn results aligned with your query sequence, you can move on to the next steps.

5.7 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.



B

```

CTACTGGTGTCTTCACTGACAAAGACAAGGCTGCAGCTCACTTGAAGTTGTCTTATTTGAATTGGTTATTTTGT
CTTGTATGATATAAATAGTTTATGTCTAGAAATTGCTTAGTATCATTCAACTAAATTTGTGACTTGTGTATTTT
CAGGGTGGTGCCAAAGAAGTTGTTATCTCTGCCCCAGCAAGAGCTCCAATGTTTGTGTTGGTGTCAACGAGCA
CGAATACAAGTCCGACCTTGACATTGTCTCCAACGCTAGCTGCACCACTAACTGCCTTGCTCCCTTGCCAAAGTAA
AATATCTGATATCTATATGATCAAAATTTGACTTTGTATTTCAGTTGAACGACTAATTCATTAAAGCTTCTTTG
AATTTATTTGTAGGTATCAATGACAGATTTCGAATGTTGAGGGTCTTATGACACAGTCCACTCAATCACTGGT
AAATTTATCAATCAGTTAGAAATTTATTACAAACTGCTTGCTTATAGGTGAAAATTTGTGATTAAATGGGGTTG
CTTTATCAATTTTCAGTACTCAGAAAGCTGTTGATGGGCTTCAATGAAGGACTGGAGAGGTGGAAGAGCTGCTTCAT
TCGACATTATTTCCAGCAGCACTGGAGCTGCCAAGGCTGTCGGAAGGTTGCTTCAGCTCTTAACGGAAGTTGACT
GGAATGATTTTCGTGTCCCAACCGTTGATGCTCAGTTGTTGACCTTACTGTGAGACTCGAGAAAGCTGCTACCTA
CGATGAAATCAAAAGGCTATCAAGTAAGCTTTGAGCAATGACAGATTAAAGTTACTTATATCCAGTAGTGATCA
AATTACTACCAAGTGTTTTACCACCAATACATAGGGAGGAATCCGAAGGCAAACTCAAGGGAATCCTTGGATACA
CCGAGGATGATGTTGTCTCAACTGACTTCGTTGGCGACAACAGGTCGAGCATTTTGACGCCAAGGCTGG

```

Intron/exon structure of *Arabidopsis GAPC* gene family. 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 bar shows the sequence coding for the active site of the *GAPDH* enzyme. B, the r 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.

5.8 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?

5.9 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?

6. Predict an Amino Acid Sequence from the Cloned Gene (blastx)

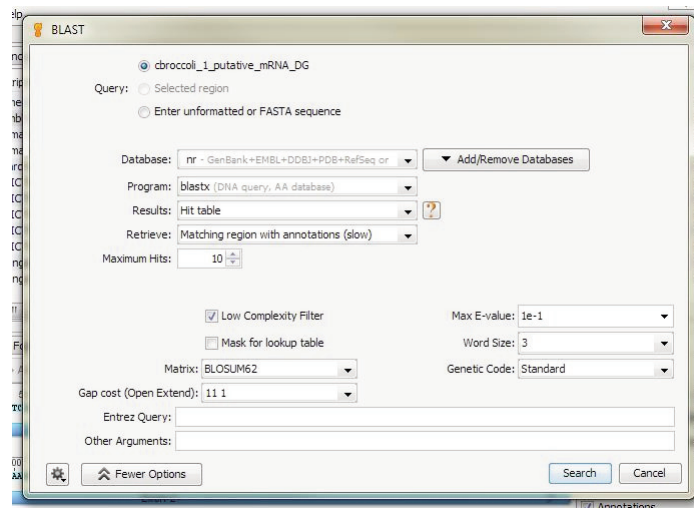
mRNA is translated into proteins by ribosomes and tRNA. Each amino acid is encoded by a group of three RNA bases called codons. A DNA sequence encodes six potential protein sequences, three for each strand, which are referred to as reading frames. The first codon in a sequence can start at base position 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.

6.1 Checking the mRNA prediction with a blastx search.

6.1.1 In your Assembly folder, select your putative mRNA sequence from step 5.6.2.6.

6.1.2 Click the BLAST icon in the menu bar. A new dialog box will appear.



6.1.2.1 Select **nr** as the Database.

6.1.2.2 Select **blastx** as the Program.

6.1.2.3 Select **Hit table** for Results.

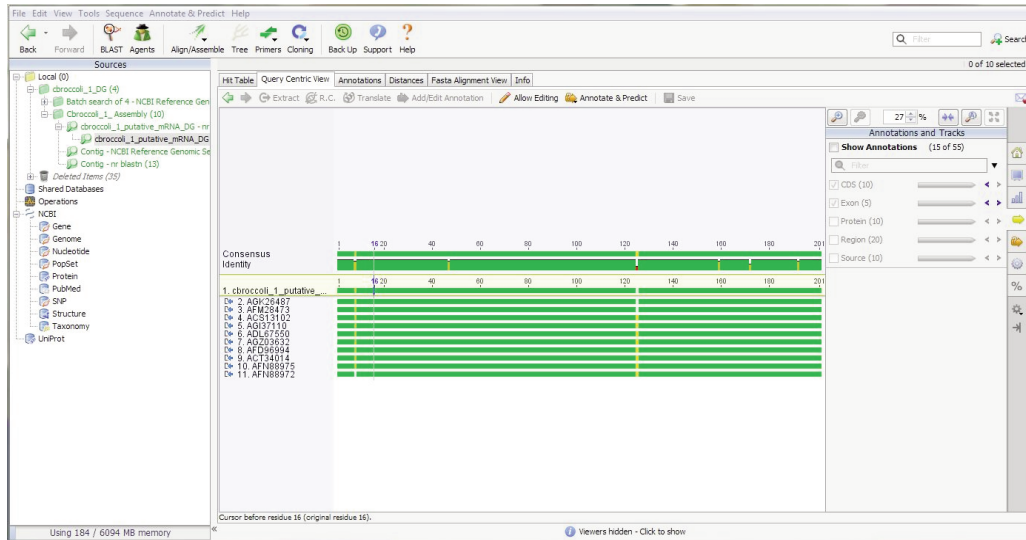
6.1.2.4 Select **Matching region with annotations (slow)** for Retrieve.

6.1.2.5 Set Maximum Hits to **10**.

6.1.2.6 Click **Search**. A new folder containing your results will appear. The name of the results folder will be the name of your document with “- nr blastx” appended to the end. For example, the cbroccoli contig would be: cbroccoli_1_putative mRNA_DG - nr blastx.”

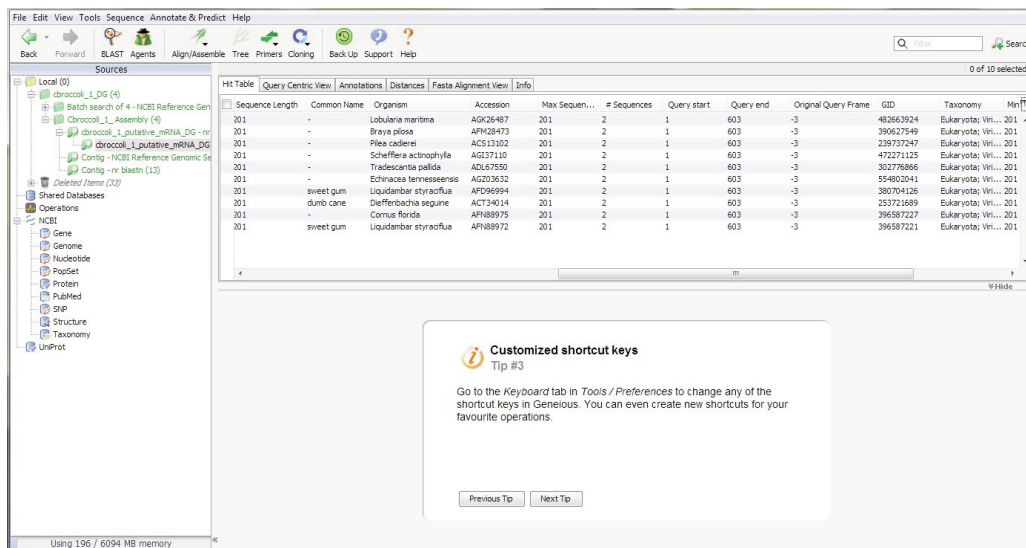
6.1.3 If your mRNA model is correct, you should see the following results:

6.1.3.1 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.

6.1.3.2 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.



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 reverse.” This means that amino acid coding begins on the third base of the reverse 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.

The screenshot shows the NCBI BLAST interface. The top panel is the Hit Table, which lists the following hits:

E Value	Bit-Score	Grade	Description	Name	Query cover...	Hit start	Hit end	Sequence Length	Comment
0	838.049	94.5%	Eutrema selaginum hypothetical protein (EUTSA_y10008...)	XM_006417059	100.00%	1,052	448	633	-
0	834.443	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009112333	100.00%	1,003	399	633	field mus
0	861.493	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009119750	100.00%	980	376	633	field mus
0	888.544	95.4%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009125789	100.00%	1,058	454	633	field mus
0	839.853	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009126757	99.68%	977	375	631	field mus
0	906.577	95.7%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009136503	100.00%	987	383	633	field mus
-	-	-	Assembly of 4 reads: Cbroccoli_L_GAPSEOF_Plate_L...	cbroccoli_L_putative_mRNA_DG	-	-	-	605	-
-	-	-	Alignment of 11 sequences	Nucleotide alignment (modified)	-	-	-	633	-

The bottom panel shows the Sequence View, which displays the query sequence (cbroccoli_L_putative_mRNA_DG) and the hit sequence (XM_006417059). The query sequence is highlighted in blue, and the hit sequence is highlighted in green. The dotplot shows a strong match between the query and the hit, indicating a high degree of similarity.

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 reverse, the first codon in the sequence is predicted to be ACU, which codes for threonine.

The screenshot shows the NCBI BLAST interface. The top panel is the Hit Table, which lists the following hits:

E Value	Bit-Score	Grade	Description	Name	Query cover...	Hit start	Hit end	Sequence Length	Comment
0	838.049	94.5%	Eutrema selaginum hypothetical protein (EUTSA_y10008...)	XM_006417059	100.00%	1,052	448	633	-
0	834.443	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009112333	100.00%	1,003	399	633	field mus
0	861.493	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009119750	100.00%	980	376	633	field mus
0	888.544	95.4%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009125789	100.00%	1,058	454	633	field mus
0	839.853	94.5%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009126757	99.68%	977	375	631	field mus
0	906.577	95.7%	PREDICTED: Brassica rapa glyceraldehyde-3-phosphate d...	XM_009136503	100.00%	987	383	633	field mus
-	-	-	Assembly of 4 reads: Cbroccoli_L_GAPSEOF_Plate_L...	cbroccoli_L_putative_mRNA_DG	-	-	-	605	-
-	-	-	Alignment of 11 sequences	Nucleotide alignment (modified)	-	-	-	633	-

The bottom panel shows the Sequence View, which displays the query sequence (cbroccoli_L_putative_mRNA_DG) and the hit sequence (XM_006417059). The query sequence is highlighted in blue, and the hit sequence is highlighted in green. The dotplot shows a strong match between the query and the hit, indicating a high degree of similarity.

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.


6.1.3.3 Record the reading frame of the best GAPDH protein sequence match for your clone: _____

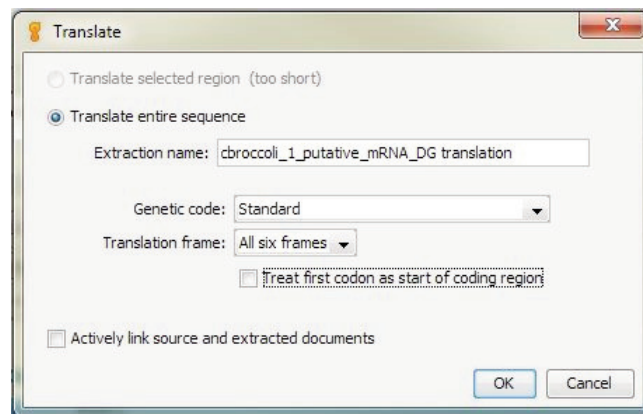
6.1.3.4 If your translated sequence matches with sequences in the database but requires different reading frames for the entire sequence to match, it is possible that there is still an error in the mRNA sequence. You can go back to section 5.7 to check your mRNA sequence. Any insertions or deletions 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).

6.2 Translate your putative mRNA sequence to obtain the predicted sequence of the protein.

By translating your sequence into all six frames, you will be able to identify the correct reading frame for your predicted protein sequence.

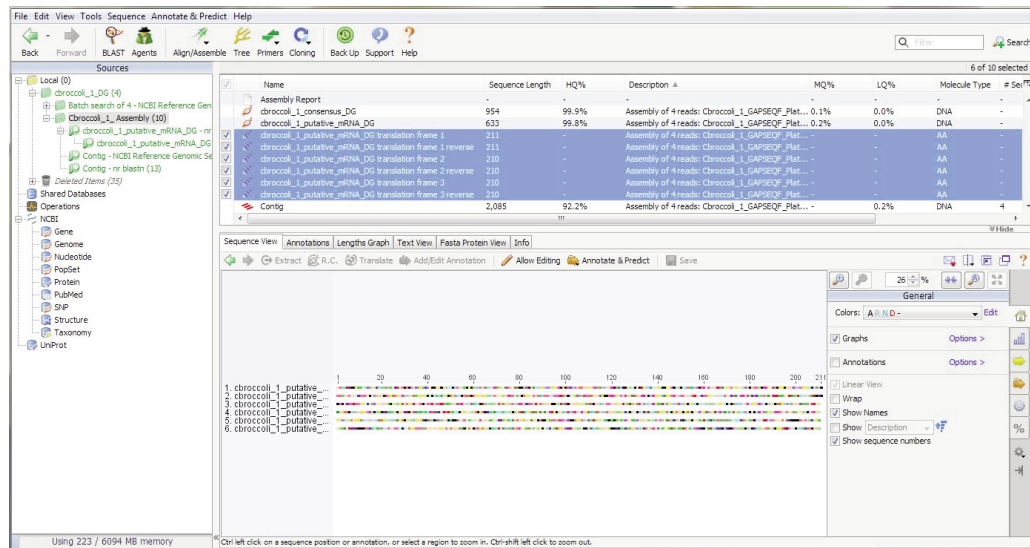
6.2.1 In your Assembly folder, select your putative mRNA document file.

6.2.2 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 **All six frames** as the Translation frame
- 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**. Six new documents will appear in your Assembly folder, each of which represents a translation frame and direction.

6.2.3 Look at each document in Sequence View. Determine the reading frame that matches the protein sequence you saw in your blastx search.

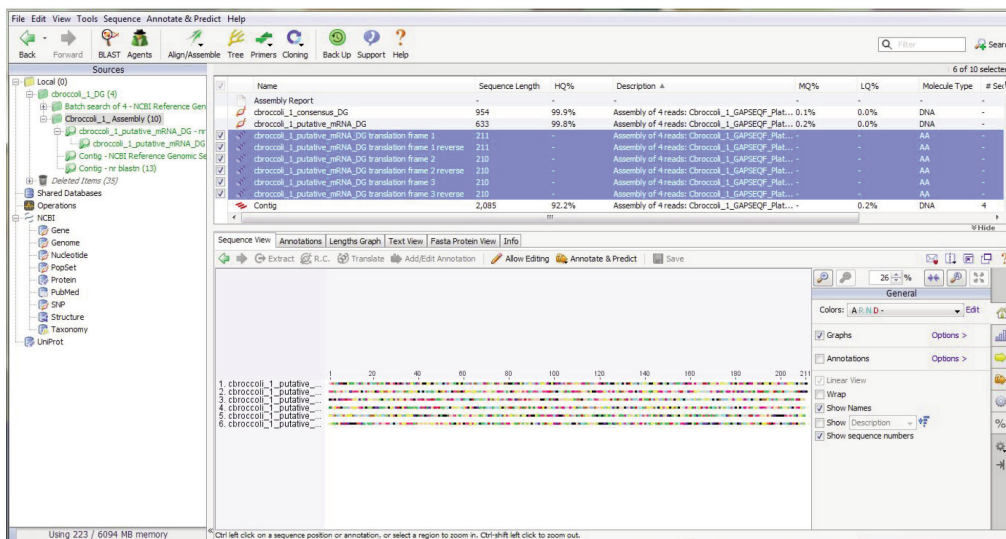


Six frame translation of the cbroccoli putative mRNA sequence. Using the Translate button in Sequence View allows you to create a separate document of the predicted protein sequence for each reading frame.

Optional: You can also view the translation of your putative mRNA sequence on the document itself. However, this will not generate new amino acid documents that you will need for the last step in this workflow.

6.2.3.1 Select the putative mRNA sequence document from your Assembly folder and go to Sequence View.

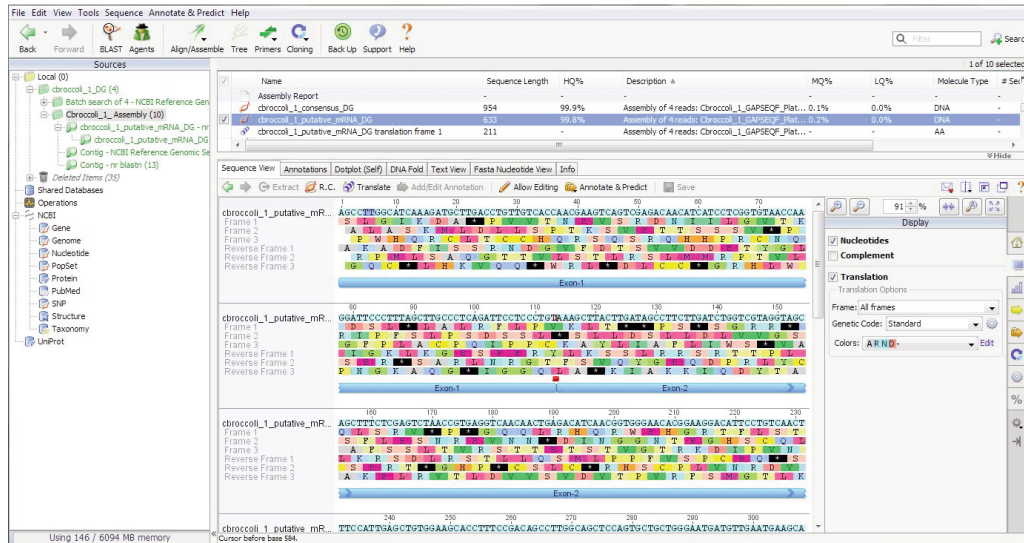
6.2.3.2 In the options panel, go to the Display tab and click the checkbox for Translation. The amino acids will now appear as colored rectangles beneath the nucleotide sequence:



Viewing the protein translation of the cbroccoli putative mRNA sequence from the options panel.

You can view the protein translation of your putative mRNA by selecting the Translation option in the Display tab of the options panel.

6.2.3.3 The default setting displays only one reading frame. In the Translation Options section, select “All frames” in the Frame dropdown list and use the standard genetic code to display all six reading frames. You should now see all six reading frames beneath your sequence:



Display all six possible reading frames for your putative mRNA sequence. Select “All frames” in the Translation Options section of the Display tab (circled in orange) to reveal all six possible reading frames for your putative mRNA.

6.2.3.4 Look at each of the reading frames. Save the corrected translated putative mRNA sequence file that matches the protein sequence you found in your blastx search. You can rename the file if you wish.

Record the name of your putative mRNA document containing the correct reading frame here:

6.3 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.

- 6.3.1** Select the document in your Assembly folder for your translated putative protein.
- 6.3.2** Click the BLAST icon in the menu bar. A new dialog box will appear.
 - 6.3.2.1** Select **nr** as the Database.
 - 6.3.2.2** Select **blastp** as the Program.
 - 6.3.2.3** Select **Hit table** for Results.
 - 6.3.2.4** Select **Matching region with annotations (slow)** for Retrieve.
 - 6.3.2.5** Set Maximum Hits to **10**.
 - 6.3.2.6** Click **Search**. A new folder that contains your results will appear, named “putative mRNA from consensus - nr blastp.”
- 6.3.3** Your results can be examined under both the Hit Table and Query Centric View tabs.
- 6.4** 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, or submit your sequence directly into GenBank using BankIt.

6.5 Section 6 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.
2. Export the contig sequences from each student group as Geneious 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.
4. Highlight all the contig files and run a sequence assembly as in Section 3.3.
5. Examine the assembly, check for any errors, and make corrections to the original reads and contigs using the techniques described in Section 3.
6. Review the locations of introns and exons and make corrections to the original reads using the techniques described in Section 5.

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 license expires, you will still have access to the software, but some functions will be restricted.

Appendix A

GenBank: Searching and Submitting Sequences

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., 2008). 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., 2008). 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 new submissions. All of these databases can be accessed by the public, free of charge, via the Internet.

Any sequences from this project may be submitted for publication in the NCBI GenBank. This information could be useful to researchers interested in the evolution or function of the gene or gene product. Sequences of genes from a species that has already been published in the GenBank should be submitted again as confirmatory data, especially if it is discovered to contain a polymorphism (a difference to the published gene sequence). Changes as minor as a single nucleotide base can have a dramatic effect on something like the function of an enzyme. Alternatively, the published sequence (if not referenced) could be incorrect.

Searching GenBank for Existing GAPDH Sequences

1. To examine the GAPDH sequences that have already been published in GenBank, go to <http://www.ncbi.nlm.nih.gov>.
2. Look for the box (near the top) titled "Search" and use the pull-down feature to select "Nucleotide" (instead of the default "All Databases").
3. In the query box, enter search terms appropriate for your gene and organism, for example "GAPC Viridiplantae," and hit the "GO" button. On the results page, there may be hundreds of gene accessions. They are listed in chronological order with the most recently submitted (or updated) accessions listed first.
4. You can narrow the list using the "Limits" tab near the top left of the results page and change fields related to the databases relevant to your sequence. For example, if you have cloned a genomic DNA fragment choose "Genomic DNA/RNA" from the "Molecule" field.

Submitting a Sequence to GenBank

Note: The NCBI frequently updates GenBank. The information presented here may not match the website. Use the help resources on the NCBI website for assistance with tasks if the information presented here is unclear.

1. Go to the NCBI account sign in page: www.ncbi.nlm.nih.gov/account/.
2. Create an account or log in to My NCBI account.
3. In the upper left corner of your My NCBI page, go to the Resources dropdown menu, and hover your mouse over **DNA & RNA**.
4. Click **GenBank: BankIt**. This will redirect you to the BankIt site.

5. In the upper right corner of the BankIt site, click on **Sign in to use BankIt**.
6. Click on the **New Submission** button.
7. Provide your contact information and click **Continue**.
8. You will be assigned a BankIt submission number. You will use this number whenever you communicate with the NCBI about the submission. It can also be used in any literature citations or research papers you might prepare related to this gene prior to publication in the GenBank. The submission number will continue to be displayed throughout the sequence submission process.

Record your BankIt number: For example, bankit1111279.

9. Add the names of all of the authors you wish to list in the GenBank accession. Generally, the first names are students and technicians, while the latter names are the instructors or principal investigators.
10. Publication Status will be Unpublished unless you have a research paper about this accession in the process of being published. Fill in the box that says Reference Title. Whatever is entered here will show up as the Title in the GenBank. Write this so that it is fairly general. If you are planning on submitting more than one sequence, for instance, you could use the same title on each submission. Click **Continue**.
11. Choose which sequencing technology was used to prepare sequences. In the case of the Cloning and Sequencing Explorer Series, choose **Sanger deoxy sequencing**.
12. Choose whether the sequence is unassembled or assembled. If you used only one primer for sequencing, then choose **unassembled sequence reads**. If you used more than one primer and assembled the resulting sequence into a longer fragment, then choose **assembled sequences (consisting of two or more sequence reads)**. Click **Continue**.
13. Choose a release date for the sequence or use the default **Immediately After Processing**. Choose No for 16S rRNA submissions. Choose the molecule type (**genomic DNA**), Topology (**Linear**) and Genomic completeness (**Unknown**).
14. Define the number of nucleotide sequences you intend to submit and paste the sequence, in FASTA format, into the frame. Follow the Example FASTA nucleotide format to properly format your sequence. You may also upload a text file in FASTA format by clicking the **Choose File** button and selecting your sequence file. Click **Continue**.
15. If you did not include an organism name in your FASTA file, you will be prompted to enter one. The organism name may be searched on the NCBI Taxonomy browser. (www.ncbi.nlm.nih.gov/taxonomy). Enter the organism name and click **Continue**.
16. Define the submission category as **Original**. Click **Continue**.
17. The Source Modifiers section is about the organism from which the genomic DNA was extracted. Although many model species are listed in the pull-down section, chances are the plant species you are working with will not be on that list. You will want to use an approved scientific name for the organism you are describing. Go to the NCBI Taxonomy Browser at www.ncbi.nlm.nih.gov/taxonomy and do a search using either the plant's scientific or common name. Once you have the binomial name (genus and species) of the plant, enter it in the appropriate box. You should use the Source Modifiers section if you know the subspecies, cultivar, variety, etc. Remember that this GenBank accession might be available to researchers for decades to come so you will want to be as specific as possible. Click **Continue**.

18. Click **Add features by completing input forms**. Select **Coding Region** and click **Add CDS by providing intervals**. Click **Add**.
19. Provide the **Features (Detail)**. This section asks for more details about your sequence. If you have correctly translated the DNA sequence into an appropriate protein sequence that appears fairly similar to other published proteins, then you are looking at the positive (+) strand. The positive strand starts with the 5' end of the DNA on the left.
20. Check the boxes for both 5' partial and 3' partial. The gene fragment cloned using this kit starts and ends in the middle of an exon, and is therefore both a 5' and 3' partial sequence.
21. Indicate the reading frame by checking one of the three boxes. If the very first DNA base on the 5' end (furthest left) is part of the first codon, then check box **1**. If the first codon does not start until the second DNA base, check box **2**. If the first codon does not begin until the third base, check box **3**.
22. Leave Pseudogene and intronless gene as default **No**.
23. Coding interval refers to the DNA sequence that is actually translated into protein (the exon regions). All of the GAPDH sequences have from one to several exons each, so will have to be described appropriately. The sequence between the "Start" and "Stop" is exon only. Each exon region will have to be entered separately. Define nucleotide interval and provide nucleotide coordinates for each exon region.
24. Enter the protein name as glyceraldehyde-3-phosphate dehydrogenase. Protein Description and EC Number can be left blank.
25. Leave the Gene and mRNA information blank. Click **Accept**.
26. Review the **Features (Overview)**. Based on the information provided, BankIt will automatically translate and define the features of your sequence. If it finds errors, it will direct you to correct them. A preview of the output will appear. Review and ensure all the information is correct and nothing is missing. If anything is missing or incorrect, you can edit the features by clicking on the edit button to the left of the Feature name. Click **Continue**.
27. The next page will provide an opportunity to review all the information and see a preview of your submission containing all the information you provided. If everything is complete and correct, click **Finish Submission**.
28. After submitting your sequence, you will receive a fresh page thanking you for using BankIt.
29. Congratulations you have now successfully submitted your DNA sequence! 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.

Appendix B

Instructor Answer Guide

Sequencing Focus Questions

1. What is DNA sequencing?

DNA sequencing is a method to determine the exact order (or sequence) of nucleotides in a DNA molecule.

2. Briefly explain the role of dideoxynucleotides in the traditional Sanger method of DNA sequencing.

Each of the four dideoxynucleotides (ddNTPs) is added at a low concentration to separate tubes containing regular radiolabeled deoxynucleotides (dNTPs). When the dideoxynucleotides are incorporated into a DNA molecule they terminate the nucleotide chain, resulting in each tube containing a panel of DNA molecules of different lengths each ending in a known base (the dideoxynucleotide added to that tube). When analyzed by gel electrophoresis, side by side on extremely thin vertical polyacrylamide gels, these DNA molecules are sorted out by size. The gel is exposed to X-ray film and the sequence of the DNA is read from the film.

3. How does automated sequencing that uses Sanger principles differ from traditional Sanger sequencing?

No radioactivity is used in automated sequencing because the ddNTPs are fluorescently labeled with different colored fluorophores. The sequencing reactions for all four ddNTPs are performed in the same tube. The panel of DNA molecules is sorted by capillary electrophoresis. As the DNA elutes from the capillary, the labeled molecules are excited with a laser and the fluorescence is measured. Software in the sequencer performs the base calling, rather than in manual sequencing in which a person reads an X-ray film. Longer DNA molecules are sequenced this way.

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 used to acquire more sequence data?

DNA sequencing is performed in only one direction, so full sequence coverage of a gene requires multiple primers. Oligonucleotide primers are synthesized to bind to sites at either end of the gene of interest (called forward and reverse primers). Sequence data is obtained working from either end of your gene of interest. Once you have sequence data, you can design internal forward and reverse primers (internal to the original primers) and generate interior sequences.

Bioinformatics Focus Questions

Section 1.9 Focus Questions

1. If a base has a quality value of 20, what is the chance that the base has been mistakenly identified?

There would be a 1% chance that the base was identified incorrectly.

2. What are the characteristics of a high quality base?

High quality bases are bases that can be identified with high confidence. The peak for a high quality base is well separated from other peaks. It does not overlap any other peaks. It shows the same spacing characteristics and height as peaks in the surrounding area.

Section 2.7 Focus Question

Why may sequencing reads from different sequencing primers (such as forward and reverse sequencing primers) from the same DNA preparation be homologous to different genes?

Each sequencing primer sequences a different part of the cloned gene. Sections of the cloned gene may be more homologous to one gene in the database, while different sections might be more homologous to a different gene.

Section 3.8 Focus questions

1. What is a read? How is this different from a sequence?

Reads are 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. If the chromatogram information is missing, then the sequence of bases is only a sequence.

2. What is a contig?

A contig is a sequence that has been constructed by comparing and merging the information from multiple reads.

3. What does it mean if a read has a "-" sign after the name?

The "-" sign indicates that the reverse complement of the sequence was used in the assembly.

4. Does an E value of zero mean that your sequence matched the subject sequence well or poorly? Explain your answer.

An E value of zero indicates that the match is unlikely to be random and therefore matched the subject sequence.

5. What would it mean if you found a subject sequence with an E value of 3?

This would mean that if you searched a database of random sequences with your query sequence, you would expect to find three sequences that matched your query sequence to the same extent as the subject sequence with the E value of 3.

6. Did you find the same gene sequences when you searched GenBank with your individual sequences as when you searched using the contig?

It is possible that the single sequences found different genes sequences than the contig due to searching using shorter sequences. This is similar to getting a high Max identity due to shorter sequence.

7. What might be some differences when searching databases using genomic DNA versus cDNA?

Genomic DNA (gDNA) contains introns and introns have much more variability from species to species than do coding regions of DNA. Complementary DNA (cDNA) contains only the coding region and, since there is more evolutionary pressure to keep the protein sequence the same, there would be less variability from species to species in the cDNA sequence.

Appendix C

Glossary

Annotating – The process of identifying the protein coding sequences and other biological features within genomic DNA sequences and adding such information to the sequence.

Assembly – Aligning and merging shorter sequences of a much longer DNA sequence in order to reconstruct the original sequence. To generate a significant portion of a genomic DNA sequence, assembly is usually used because current technology only allows for sequencing of 600–1000 base pair fragments of DNA with high fidelity.

Base call – Reading a DNA sequencing chromatograph and assigning a base to each peak.

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 (NIH working definition: <http://www.bisti.nih.gov/CompuBioDef.pdf>).

BLAST – Basic Local Alignment Search Tool – a suite of computer programs that are used to compare DNA and protein sequences to those in libraries of databases to search for similarities (<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>).

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.

Consensus sequence – A sequence that has been constructed from the comparison of multiple sequences. The result represents the best guess of what the base calls (or amino acids in the case of protein alignments) should be at each location.

Contig – A sequence that has been constructed by comparing and merging the information from sets of overlapping DNA segments.

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.

Discrepancies – Differences in base calls between two or more different sequences of the same clone or between different clones of the same gene.

DNA sequencing – Determining the exact order of nucleotides in a DNA molecule.

Exon – Eukaryotic gene segment that is transcribed to RNA, retained after RNA processing, and will be (with other exons) part of the mRNA that is 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. Exons are also known as the protein coding sequences of genes and introns as the noncoding regions.

FASTA format – A format used for submitting sequence data (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 without any spaces within the line. FASTA format uses single letter codes for the sequence without 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.

GenBank – The sequence database maintained by NIH. As of February 2008, GenBank contained 85,759,586,764 bases in 82,853,685 sequence records (<http://www.ncbi.nlm.nih.gov/Genbank/>).

Genome – The total genetic material of an organism.

Genomic DNA (gDNA) – All of the chromosomal DNA found in a cell or organism.

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.

Indel – A sequence discrepancy due to either an inserted or a deleted base.

Paralogous – Genes that share a high level of homology and are from the same genome.

Orthologous – Genes that share a high level of homology but are from different species.

Quality score (or value) – A numerical value indicating the confidence level for base calls. A higher quality value means higher confidence that the base is correct. A lower quality value suggests that the base call has a lower chance of being reliable and thus accepted.

Query – In terms of Geneious and relational databases, this is a program written in SQL that is used to extract information from a sequence database.

Query sequence – The input sequence (or other type of search term) with which all of the entries in a database are to be compared (<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html>).

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).

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) – Programming language used to extract relationships between different data sets in a relational database. For Geneious, a program written in SQL that does this is called a query.

Subject sequence – A sequence found by BLAST to have similarity to a sequence entered by the user (the query sequence).

Appendix D

References

- Allison LA (2007). *Fundamental Molecular Biology*. (Malden, MA: Blackwell Publishing).
- 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).
- Huang X and Madan A (1999). CAP3: a DNA sequence assembly program. *Genome Res* 9, 868– 877.
- Lodge J et al. (2007). *Gene Cloning, Principles and Applications*. (New York: Taylor & Francis).
- Lyons E and Freeling M (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* 53, 661-673.
- 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).
- Mychaleckyj JC (2007). Genome mapping statistics and bioinformatics. *Methods Mol Biol* 404, 461–488.
- Petersen J and Cerff R (2003). Origin, evolution, and metabolic role of a novel glyco-lytic GAPDH enzyme recruited by land plant plastids. *J Mol Evol* 57, 16–26.
- Primrose SB and Twyman RM (2006). *Principles of Gene Manipulation and Genomics*. (Malden, MA: Blackwell Publishing).
- 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.
- Selzer PM et al. (2004). *Applied Bioinformatics: An Introduction*. (New York: Springer).
- Xia X (2007). *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*. (New York: Springer).

Bioinformatics Internet Resources

www.ncbi.nlm.nih.gov/

National Center for Biotechnology Information (NCBI) homepage

www.ncbi.nlm.nih.gov/blast/Blast.cgi

NCBI's BLAST homepage for homology searching

www.blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

NCBI's tutorial webpage for BLAST

www.pbil.univ-lyon1.fr/cap3.php

CAP3 Program for sequence assembly and generation of contigs

www.mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=cap3

Alternative CAP3 Program for sequence assembly and generation of contigs

www.molbiol.ru/eng/scripts/01_13.html

Six-frame translation program from MolBio

www.ebi.ac.uk/Tools/emboss/transeq/index.html

Six-frame translation program from the EMBL-EBI

Legal Notices

Bio-Rad's thermal cyclers and real-time thermal cyclers are covered by one or more of the following U.S. patents or their foreign counterparts owned by Eppendorf AG: U.S. Patent Numbers: 6,767,512 and 7,074,367.

Trademarks

Apple, Mac and Mac OS are trademarks of Apple Inc.

BLAST is a trademark of the National Library of Medicine.

CentOS and RHEL are trademarks of Red Hat, Inc.

Excel, Word, and Windows are trademarks of the Microsoft Corporation.

Geneious is a trademark of Biomatters Limited.

GenBank is a trademark of the United States Department of Health and Human Services.

Firefox is a trademark of the Mozilla Foundation.

Intel is a trademark of Intel Corporation.

WinZip is a trademark of WinZip International LLC.

Java is a trademark of Oracle America, Inc.

Ubuntu is a trademark of Canonical Ltd.

© 2012 Bio-Rad Laboratories, Inc.



1665027

BIO-RAD

**Bio-Rad
Laboratories, Inc.**

*Life Science
Group*

Web site bio-rad.com **USA** 1 800 424 6723 **Australia** 61 2 9914 2800 **Austria** 43 1 877 89 01 177 **Belgium** 32 (0)3 710 53 00 **Brazil** 55 11 3065 7550
Canada 1 905 364 3435 **China** 86 21 6169 8500 **Czech Republic** 420 241 430 532 **Denmark** 45 44 52 10 00 **Finland** 358 09 804 22 00
France 33 01 47 95 69 65 **Germany** 49 89 31 884 0 **Hong Kong** 852 2789 3300 **Hungary** 36 1 459 6100 **India** 91 124 4029300
Israel 972 03 963 6050 **Italy** 39 02 216091 **Japan** 81 3 6361 7000 **Korea** 82 2 3473 4460 **Mexico** 52 555 488 7670 **The Netherlands** 31 (0)318 540 666
New Zealand 64 9 415 2280 **Norway** 47 23 38 41 30 **Poland** 48 22 331 99 99 **Portugal** 351 21 472 7700 **Russia** 7 495 721 14 04
Singapore 65 6415 3188 **South Africa** 27 (0) 861 246 723 **Spain** 34 91 590 5200 **Sweden** 46 08 555 12700 **Switzerland** 41 026 674 55 05
Taiwan 886 2 2578 7189 **Thailand** 66 662 651 8311 **United Arab Emirates** 971 4 8187300 **United Kingdom** 44 020 8328 2000

