BIO-RAD Explorer™

A Revised Workflow for Intron-Exon Boundary Mapping

Addendum

IMPORTANT: This revised workflow replaces chapter 9, sections 5.1–5.6 in the Cloning and Sequencing Explorer Series manual, Ver E, and sections 5.1–5.6 in the Sequencing and Bioinformatics Module manual, Ver C.

For technical service, call your local Bio-Rad office or, in the U.S., call **1-800-424-6723**

BIO-RAD

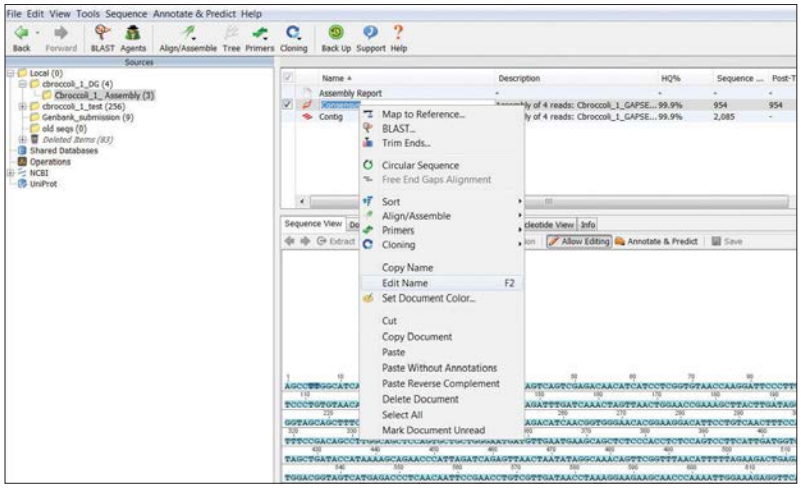
IMPORTANT: This revised workflow replaces chapter 9, sections 5.1–5.6 in the Cloning and Sequencing Explorer Series manual, Ver E, or sections 5.1–5.6 in the Sequencing and Bioinformatics Module manual, Ver C. The revised workflow prevents a problem that some users have encountered locating intron-exon boundaries after doing a BLAST search. You can determine the version (e.g. Ver E) of your manual by looking at the bottom of the last page of the pdf or the back of the printed manual.

Protocol

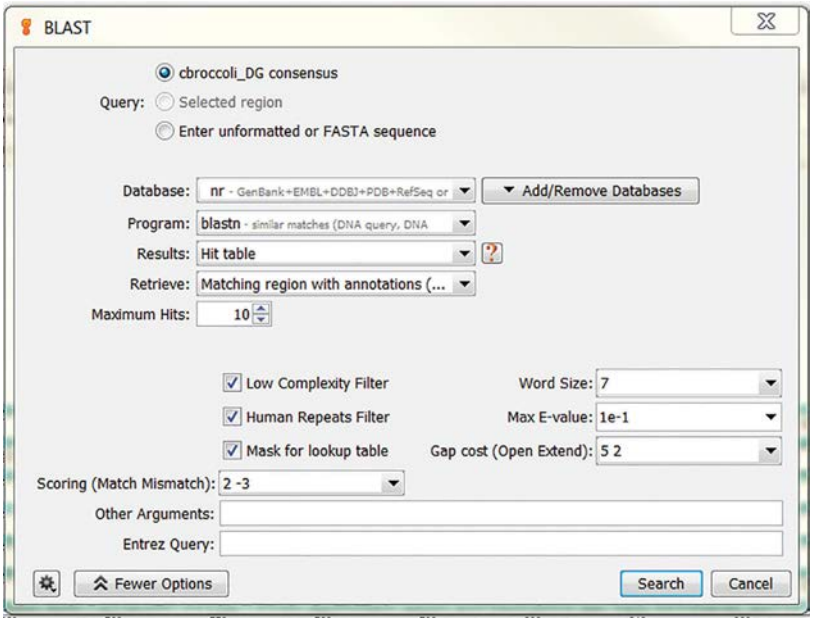
5.1 Using blastn to align the contig to sequences in the nr database.

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

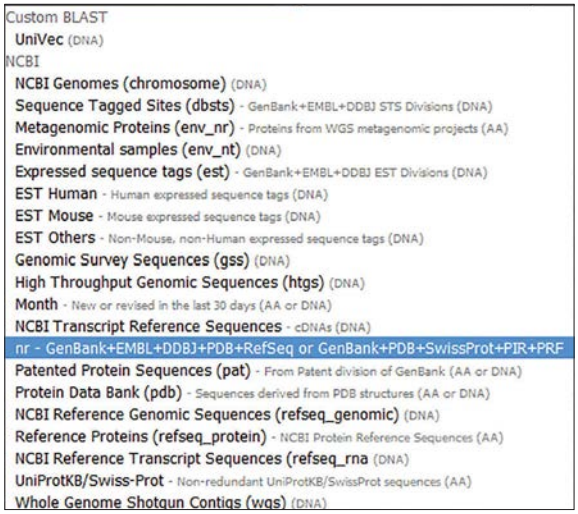
5.1.1 To avoid confusion with file names in the subsequent workflow steps, rename your consensus sequence. In your assembly folder, select your Consensus file and right click to open a menu. Select **Edit Name**. Rename your file to something easy to remember, such as the name of your sample, your initials, and then consensus. In the example below, the file will be renamed to **cbroccoli_DG_consensus**.



5.1.2 To begin your BLAST search, select your newly renamed consensus file. Select the BLAST icon from the menu bar. A new dialog box will appear:



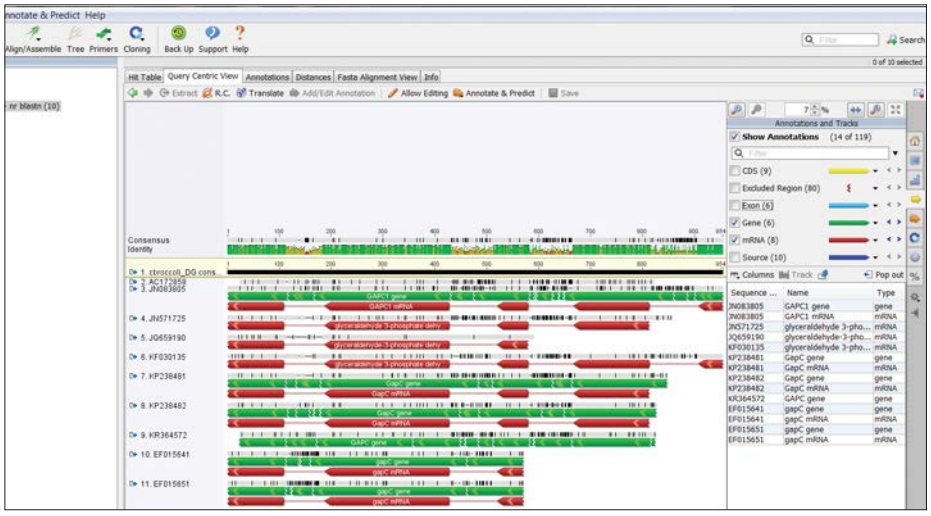
- Select **nr** as your database



- The default selection for Query should be your renamed consensus file
- Select **blastn** for Program
- Select **Hit table** for Results
- Select **Matching region with annotations (slow)** for Retrieve
- Set Maximum hits to **10**. This change will make the results easier to interpret because fewer sequences will be shown.
- Click **More Options**
- Change word size to **7**. This will increase the sensitivity of the blastn search and allow you to detect more distantly related sequences and short exons
- Click **Search**. A new folder will be created within your folder with the name of your consensus file — nr blastn (10)

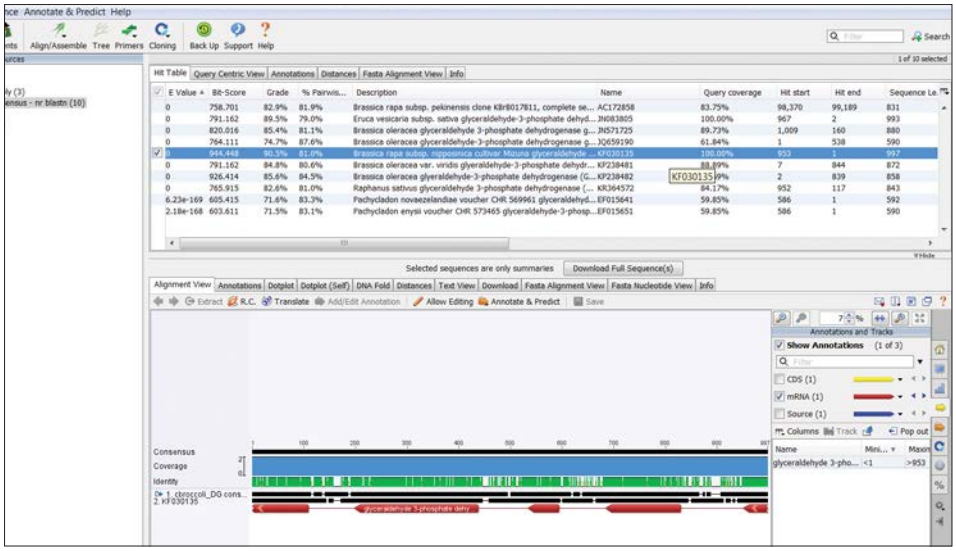
5.2 Interpreting the results and predicting the exon positions.

The Query Centric View will show the consensus sequence and nucleotide coordinates near the top of the page. Below, you will see the BLAST results alignment showing where portions of the sequences from the nr database align to your contig. If you do not see the annotations colored in red and green like the example below, navigate to the Annotations and Tracks tab  and check the boxes for **Show Annotations**, then **Gene** and **mRNA**. The mRNA annotations (in red) correspond to known exons.



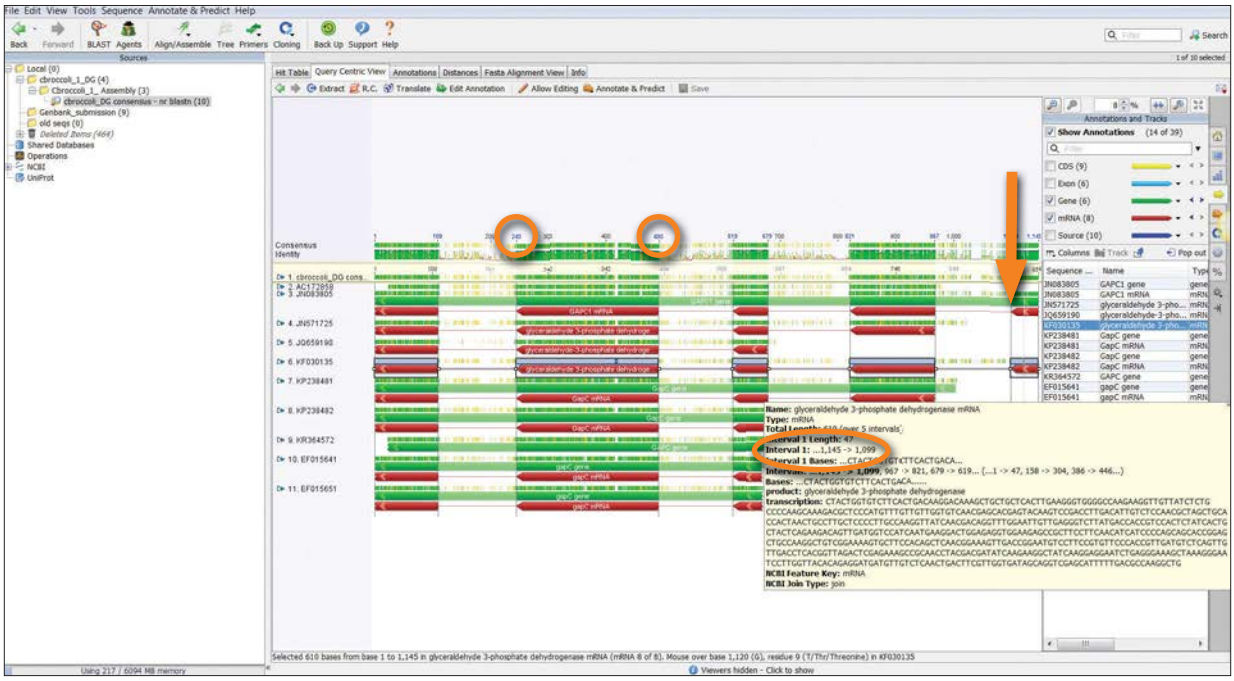
5.2.1 Identify the BLAST result with the longest and most extensive match to your contig. You can do this by looking at the statistics in the Hit Table and the corresponding alignments in Query Centric View. In the example below, the best match is the sequence named KF030135 *(Brassica rapa subsp. nipposinica cultivar Mizuna)* because it has the lowest E Value, Bit-Score, Grade, and % Pairwise values.

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.



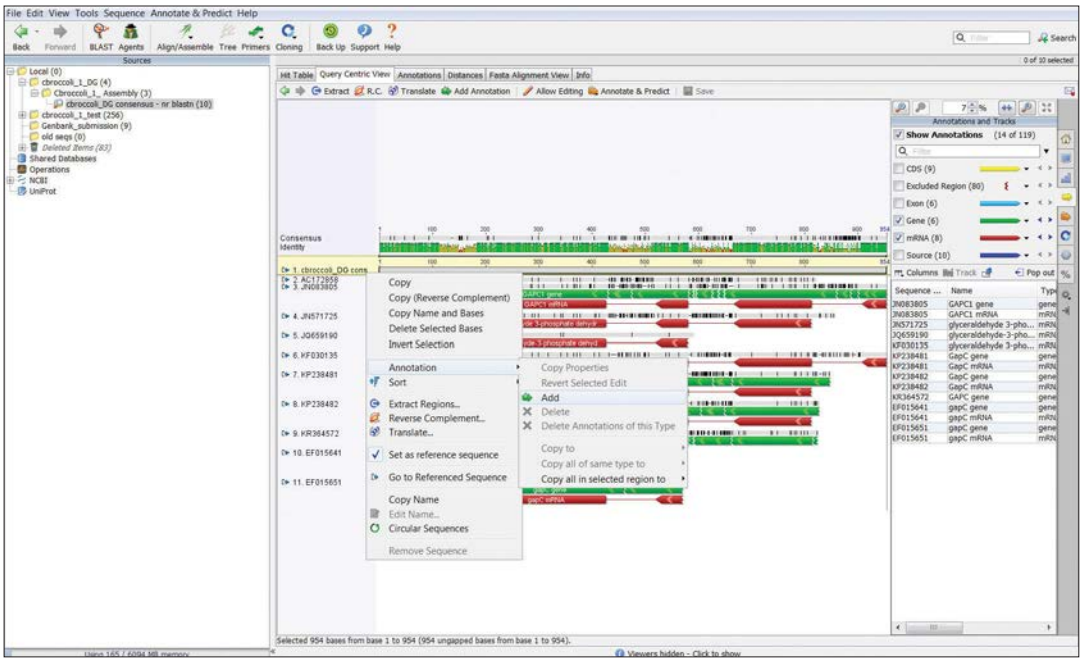
5.2.2 From the BLAST result with the best match to your contig, record the nucleotide coordinates of the mRNA intervals. These nucleotide coordinates will help you predict intron/exon boundaries in your consensus sequence.

- Go to Query Centric View.
- For your best match from the BLAST results, click on one of the mRNA annotations (red). You will see that all of the mRNA annotations will be selected.
- There are two places where you can look for the nucleotide coordinates:
  - o The coordinates for each interval appear in blue above the consensus sequence. See an example encircled in orange in the figure below. You now have all the nucleotide coordinates for all of your intervals.
  - o The coordinates for the interval you are hovering over can be found in the yellow pop up window. In the example below, interval 1 is highlighted, which spans bases 1145 to 1099. You will note that this range is decreasing; this is because this interval is in the reverse direction (see arrowhead at the end of the annotation). Repeat this for all the intervals.
- Record these coordinates and note whether these intervals are in the forward or reverse direction, based on which direction the interval arrowhead is pointing. In the example below, the orange arrow points to the arrowhead for interval 1, which is pointing in the reverse direction.
- Also note whether any of the intervals are truncated, or cut-off at the end. In the example below, interval 1 is truncated at the right side because it appears to continue beyond our sequence. Interval 5 is also truncated but on the left side, because it appears to continue to the left beyond our sequence.

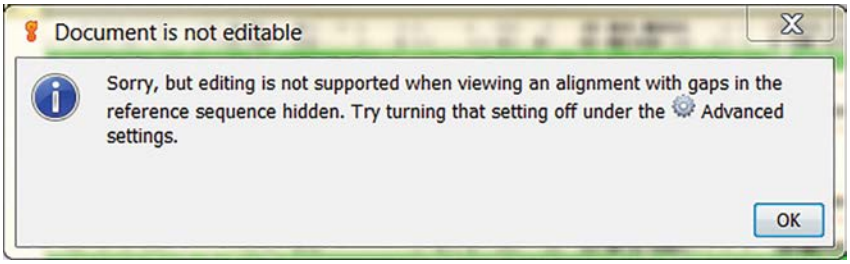


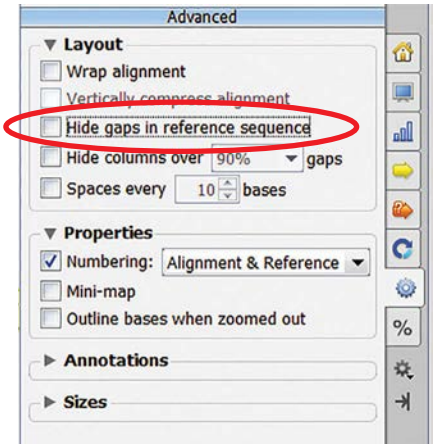
NOTE: There are two places to find nucleotide coordinates to map intron/exon boundaries: 1. Click on the mRNA annotation (red) for your best match and look for the blue numbers (see orange circles). 2. Hover over an interval and the yellow pop up window will list the nucleotide range (see orange oval). Also take note of the directionality of the interval by looking for the arrowhead (see orange arrow). An arrowhead on the left side indicates that the interval is in the reverse direction.

5.2.3 Find the intron/exon boundaries in your consensus sequence and mark them. In Query Centric View, right click on your consensus sequence. In the new menu, navigate to **Annotation**, the **Add**.



If you see a warning stating that your Document is not editable, follow the instructions to uncheck the box for **Hide gaps in reference sequence**.

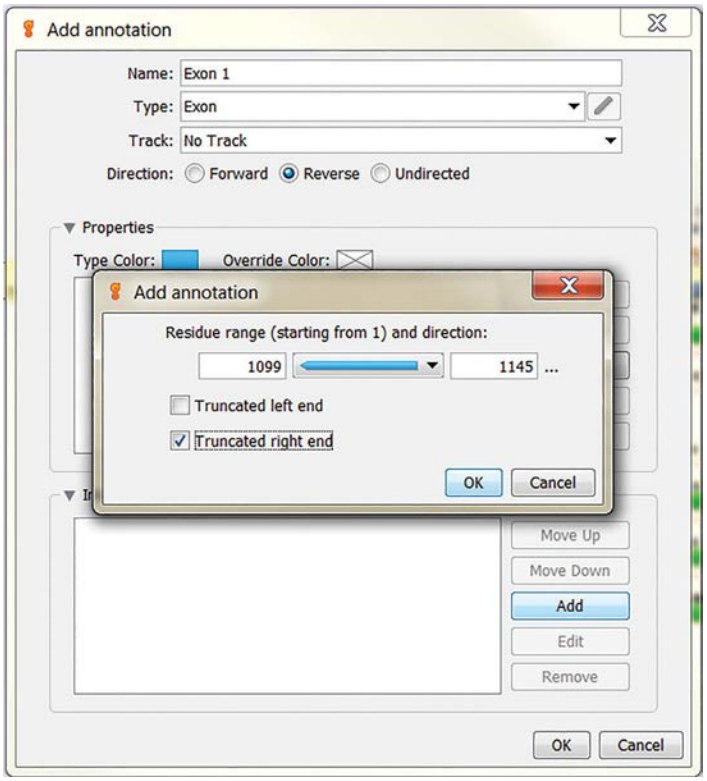




In the new window that appears, you will add all the nucleotide intervals that you noted down from section 5.2.2. Begin with the interval nearest the 5' end.

- For Name, the first interval will be named Exon 1
- For Type, select **Exon** from the dropdown menu
- For Track, keep the default selection
- Select the direction for that interval
- In the **Intervals** section, click to highlight the existing interval, and select **Remove** (this represents your entire consensus; you do not need to annotate this). Now click **Add**, and enter in the nucleotide coordinates that you previously noted down. Make sure the arrow is pointing in the correct direction. If your interval was cut off or truncated at either end, indicate as such by checking the correct box. Otherwise, leave them unchecked.

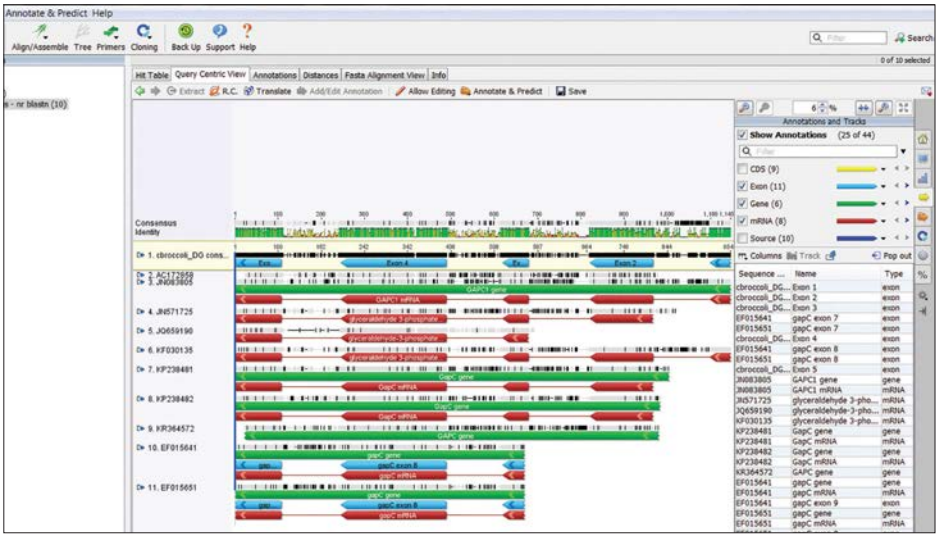
NOTE: Your arrows may not be blue, and could appear charcoal gray or any other color. If you wish, you can specify the color under the properties section under **Type Color**.



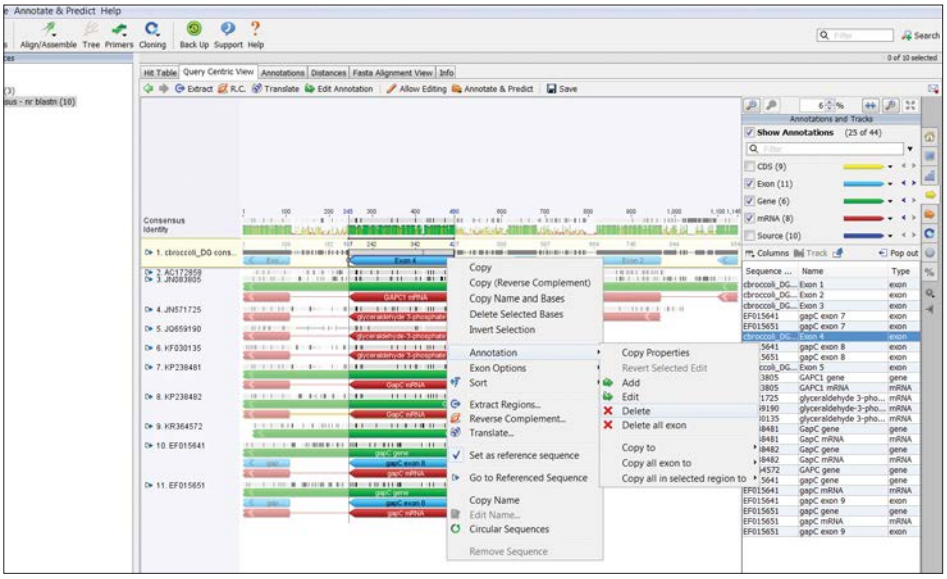
- Click **OK** to exit out of both windows. Your new interval called Exon 1 has been added as an annotation.
- For your next interval, repeat from the beginning of step **5.2.3** and label it exon 2.
- Go through all of your intervals and add them all as annotations on your consensus sequence.

NOTE: If you want to submit your sequence into GenBank, the exons must be numbered sequentially in this fashion. Otherwise, you can use the Add button to add all your intervals at the same time without having to open new windows each time, but all the exons will have the same name.

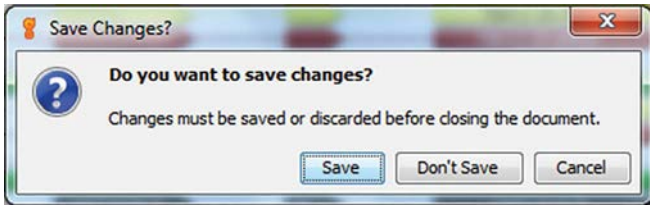
- When you are finished, your mapped exons will appear on your consensus document as blue annotations. (You may also see other blue annotations in the BLAST results, because adding your exons automatically checked the box to show all exon annotations)



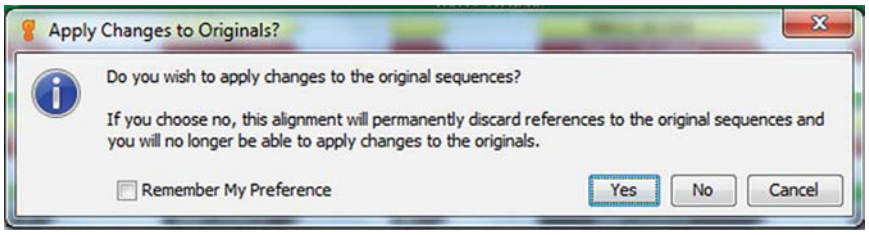
- If you made a mistake while annotating an exon(s) and want to delete it:
  o Click to highlight the exon. Right click and navigate to **Annotation**, then **Delete** (or **Edit**)



- In the Query Centric View toolbar, be sure to click **Save** to save your work. If you navigate away from the view, you will be prompted to save your changes.



- You will also be asked to apply changes to the originals. Click **Yes**.

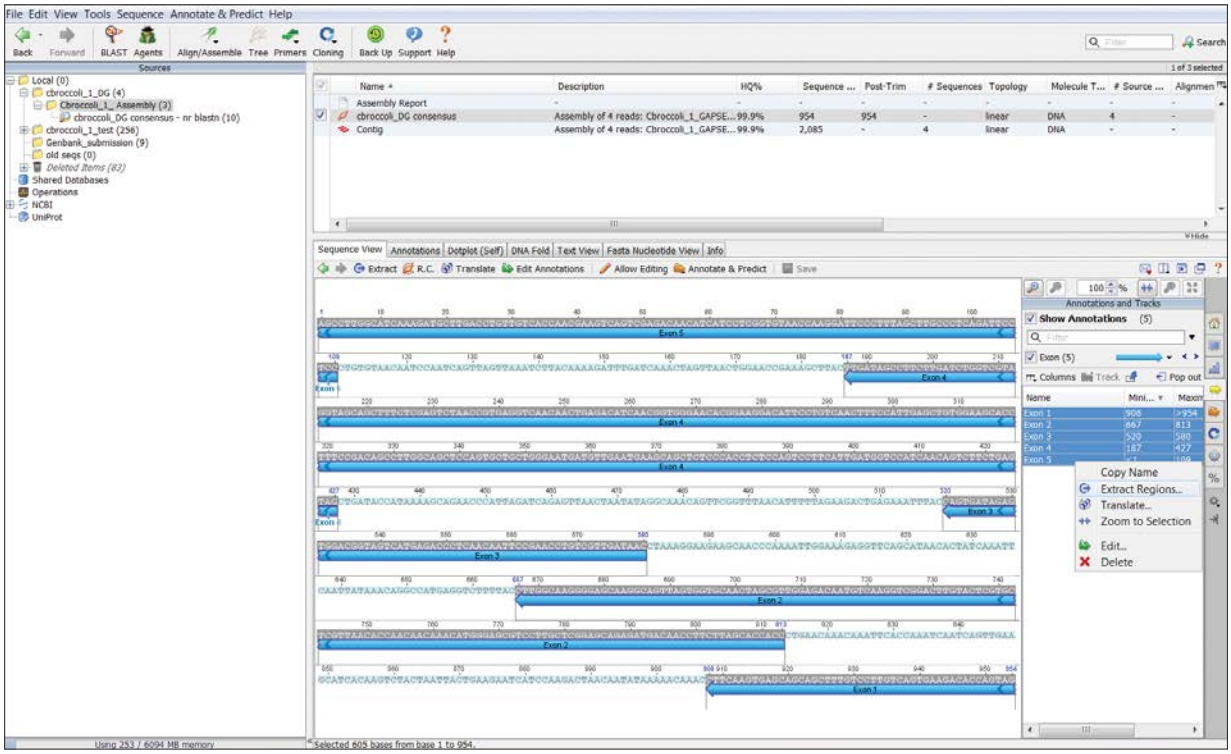


5.3 Check initial gene model (putative mRNA) with blastn and further refine model.

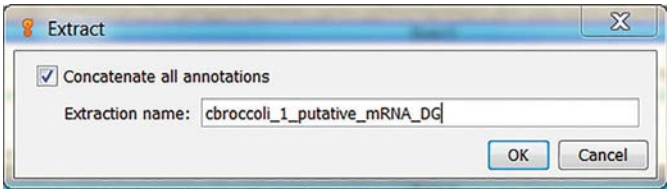
Now it is time to check your work by doing a blastn search with your proposed sequence for the mRNA. You will create a new document with only the exon sequences as your putative mRNA, run a new blastn search, and see how well the top results match your putative mRNA.

5.3.1 To create a document with only your exon sequences as your putative mRNA, go to your Assembly folder and select your annotated consensus sequence. Click on the Annotations and Tracks tab  . Select all the exons by holding the Shift button and highlighting the first and last exon on the list. In Sequence View, you will see that the sequences of these exons are highlighted.

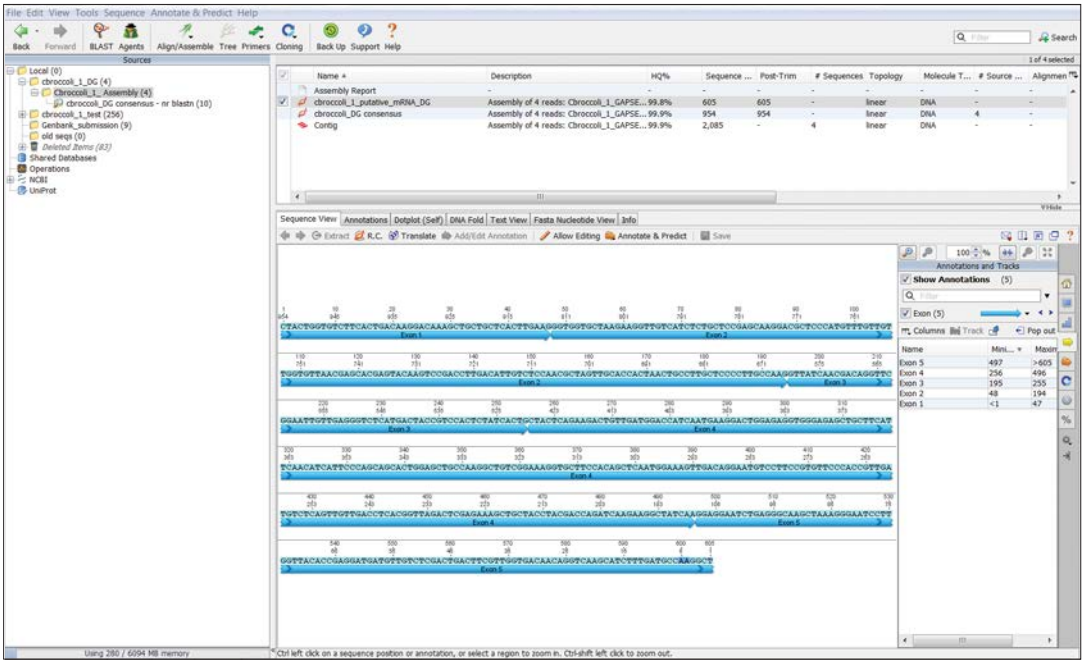
- Hover your mouse over the selected exons in the Annotations and Tracks tab. Right click and navigate to **Extract Regions**.



- In the new window that appears, check the box for **Concatenate all annotations**. For the extraction name, be sure to add “putative mRNA” and select a naming convention similar to your previous consensus document. In the example below, the new file will be named **cbroccoli_1_putative_mRNA_DG**.

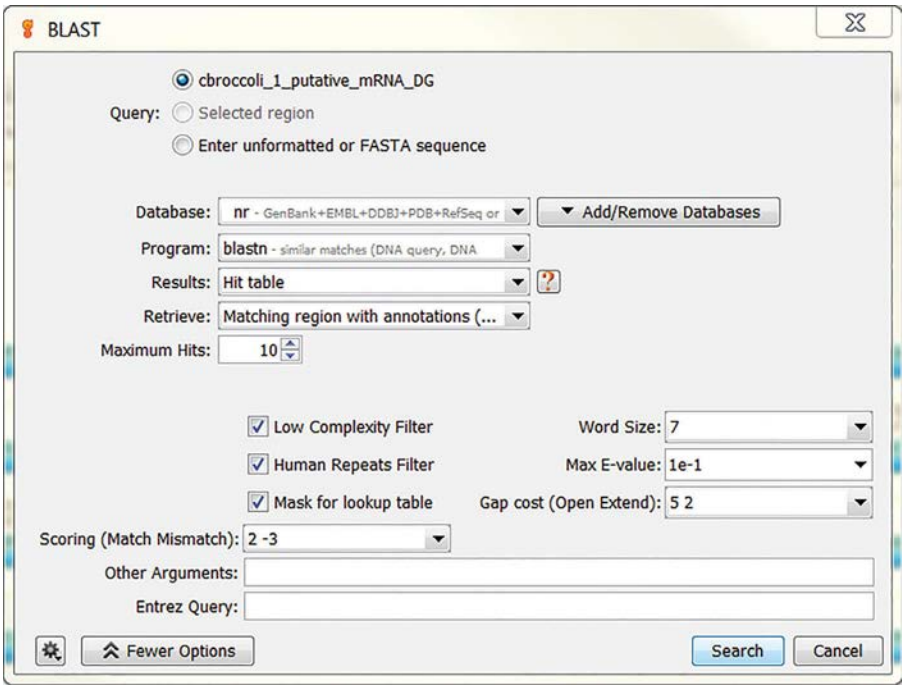


- Click **OK**. You will see your new putative mRNA document in the Assembly folder. In Sequence View, you’ll see that all the intronic sequences are removed and only the exonic regions are kept.



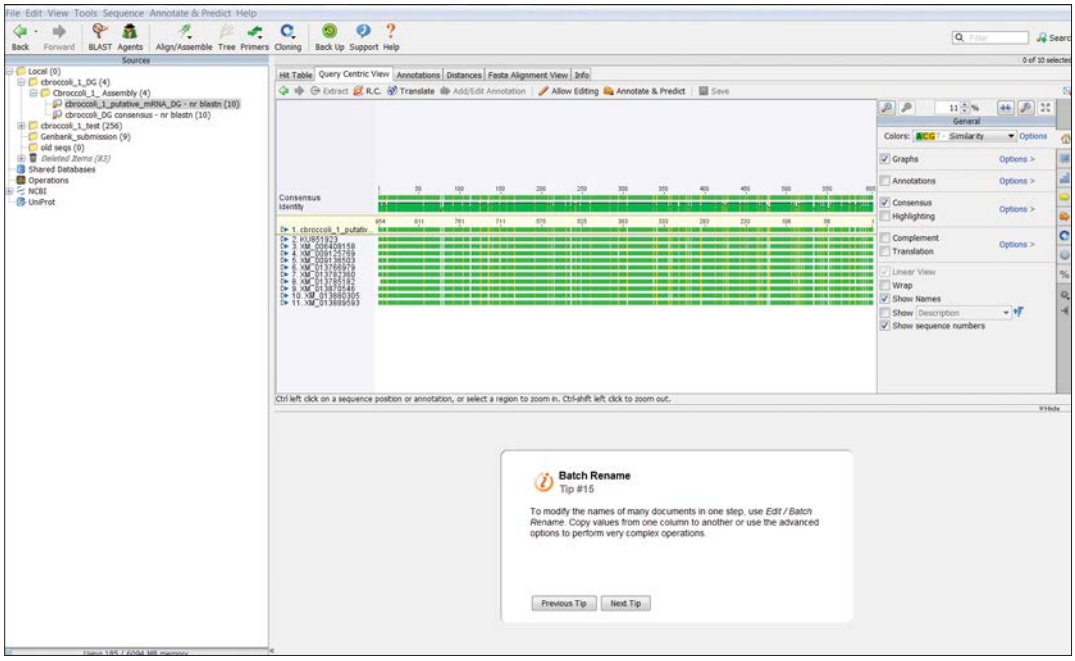
5.3.2 Run a blastn search on your putative mRNA sequence. Select your putative mRNA sequence file. Click the BLAST icon in the menu bar. Keep the defaults from the previous search, namely:

- Database should be **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.



5.4 Interpreting the results and finding and resolving errors.

In your putative mRNA blastn folder, you should now see your sequence and the ten best fits from the blastn search aligned in Query Centric View. These matches most likely will be different than the ones returned from the general nucleotide database when you searched for alignments for your single sequences and contig. This is because sequences in the reference database are scrutinized at a higher level than ones in the general database by NCBI scientists. In the Hit Table, there should be a high level of query coverage and a low E-value; the % pairwise identity will vary. There is much more variation in the intron regions of genes than in the exons (coding regions), so the level of homology should be high. However, there could still be a reasonable amount of variation between plants in the coding regions.



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 putative mRNA matches several subject sequences along the entire length with almost no 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 there may be regions where two joined exons have errors. There may also be errors that were missed on the first run-through.

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

Tip: You can also navigate to the Advanced Settings tab and toggle the box for **Hide gaps in reference sequence** (as in step 5.2.3) to see if the sequences look any different. If the sequences look the same, there are no gaps.

5.4.2 If gaps are observed, you will need to examine your results in further detail. Proceed to section 5.6.2 in Chapter 9 of the Cloning and Sequencing Explorer Series or the Sequencing and Bioinformatics Module manuals.

5.4.3 If there are no gaps, you are ready to proceed to section 5.7 in Chapter 9 of the Cloning and Sequencing Explorer Series or the Sequencing and Bioinformatics Module manuals.

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*

Bulletin 6858 Ver A US/EN *16-0943 0916*