Steps taken to analyze Blast-Parser results within Excel
Results from BlastN

Initial preparations

- Read output of Blast-Parser into Excel (make sure all files are visible, including text files; choose delimited file type (click on **Next**); choose comma as a delimiter (click on **Next**); click on **Finish**)
- Insert row before line 1 (put cursor at line 1; click on **Insert**; click on **Rows**)
- Add to top row titles for each column. If you don't know what the columns should be, visit the program that generated the output, particularly the subroutine that printed the output.
- Set the top row apart from the rest by underlining it. To do this, click on the row to select it and then click on the **Borders icon** in the tool bar. Select **single underline**. (Warning: you'll get less than satisfactory results if you click on **U**. That will just underline the individual words of the title line).
- Continue setting the top row apart by changing its color. To do this, click on arrow next to the **Fill color** icon (a tipping bucket) and click on your favorite color.
- Scan the data to see if the values are reasonable. If not, then back to the program!

  Interlude: What are we trying to do here? We need to pick out those lines in which a match of one end is near the end of another and the two are oriented properly with respect to one another. Deal with the first criterion first.

Calculate Distances

- Sort the output so that the lines are arranged in the order that they appear in the chromosome. To do this select the entire table (click on the upper left gray corner, above **row 1**; click on **Data**; click on **Sort**; make sure the **My list has Header row** is clicked; sort by the beginning of the hit in the chromosome/target sequence – whatever you called that column; click on **OK**).
- Make a new column header in the first available space (probably **L1**) saying something like *distance*
- Enter a formula in **L3** that will subtract the end of the previous chromosomal location (in **I2**?) from the beginning of the chromosomal location on the current line (in **H3**?):

  =H3-I2

  Problem: Are these locations correct? Are these locations correct for both the **d**irect and the **c**omplementary orientations?
- Whoops! In my output at least, the locations are correct for the direct orientation but reversed for the complementary orientations. This isn't a mistake in the Blast Parser, just a feature. One could reasonably list the beginning and end of a hit backwards or forwards for a backward looking hit. Unfortunately, listing it backwards (higher coordinate first) is inconvenient for my calculation of distance.

- Make all beginnings low number first, in the following way: **Insert** two columns before the column labeled distance. Do this by clicking the letter at the top of the column, then click on **Insert** and **Columns**. Do this again for the second column.
- Label the columns something appropriate (e.g. TargetLowEnd, TargetHighEnd).
- Enter a formula in the second row of the column labeled TargetLowEnd (or whatever):

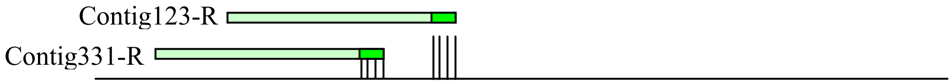
  =MIN(H2,I2)
- Enter a formula in the second row of the column labeled TargetHighEnd (or whatever):

  =MAX(H2,I2)
- Block out the two cells you just entered and extend them to the bottom of the table (grab the little black box at the lower right of the blocked out cells).
- Check to make sure that the formula worked for both direct and complementary hits.
- OK, back to calculating distance. In the column labeled *Distance*, third row, enter the formula:

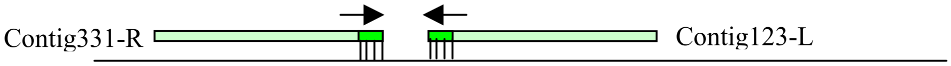
  =L3-M2
- Extend the formula to the bottom of the table
- Check to make sure that the subtraction worked properly

Check orientation

Interlude: What geometry are we looking for? This will take some thought. Certainly we don't want hits of the following variety:



The ends of these two contigs are close, but they are not nearby contigs at the opposing termini of a gap. What we DO want is:



If you found ends in the orientation shown, it would be worthwhile making primers and trying to bridge the gap by PCR. So it is important which ***types*** *of ends (L or R) of contigs are near each other. Consider also that what is labeled Contig331-R might well be Contig-447-L but flipped over. In other words, a complementary hit of a left-hand end will be oriented the same way as a direct hit of a right-hand end. So we are left with the following possibilities:*

	Second hit			
First hit	**Left, direct**	**Right, direct**	**Left, complement**	**Right, complement**
Left, direct				
Right, direct	*good*	*ignore*		
Left, complement	*good*			
Right,complement				

> *You can fill in the rest of the table. We already have a column that says the orientation, D or C, but we need to extract information regarding which end of the contig was matched. That information is in the first column, i.e., the name of the contig.*

- Label a new column (maybe "EndType"). In row 2, enter the following formula:

  =RIGHT(A2,1)

  This formula extracts the right-most character of cell A2, which happens to contain the letter L or R.

- Extend this formula to the bottom of the table.
- Label a new column (maybe "OrientationCombo") and in row 3 insert the formula:

  =IF(AND(O2="R",G2="d",O3="L",G3="d"),"good","bad")

  This means that the cell is set to "good" if the previous hit is from a righthand end in the direct orientation and the current hit is from a lefthand end in direct orientation; otherwise the cell is set to "bad".

- If the formula seems to work properly (extend the formula to many cells to check), modify it to consider all the other combinations in the table above that you want to consider good. Do this using the OR function:

  =IF(OR(AND(..),AND(..),AND(..),AND(..)),"good","bad")

  where AND(..) contains one of the conditions you determined to contain an appropriate set of orientations. If ***any*** of the conditions are met, then the cell will be set to "good", otherwise it will be set to "bad".

> *Interlude: Programming in Excel is clumsy. It's difficult to see what the formulas mean as you enter them and far more so later. For that reason, it's important to go slowly, checking each step to make sure it does what you want.*

Combine distance and orientation criteria and identify good PCR candidates

- You want to consider only those instances where the orientations are right and the distance between the ends are within range of PCR (let's say within 8000 bp). Label a new column (maybe "PCRcandidate"), and in row 3 insert the formula below and extend it to the bottom of the table.

  =IF(AND(P3="good",N3<8000),"yes","no")

- The lines marked “yes” are the ***second*** element of a two-end pair. To include the first element as well, label a new column (maybe “PCRpair”), and in row 2 insert the formula below and extend it to the bottom of the table.

  ```
  =IF(Q3="yes","left",IF(Q2="yes","right",""))
  ```

  This identifies the left element of a pair as one in which the ***next*** row is marked as a PCR candidate and identifies the right element of a pair as one in which the ***current*** row is so marked. The syntax reads: IF the next row is marked, then put “left” in the cell; otherwise IF the current row is marked, then put “right” in the cell; otherwise leave the cell blank.

- To examine only those hits that are PCR candidates, use Excel’s filtering capability. Click on the column labeled PCRpair (or whatever you labeled it), then click on **Data** in the tool bar, then **Filter**, then **Autofilter**. Click on the **down-arrow** that should have appeared at the side of the PCRpair column and click on **(NonBlanks)**.