Tutorial

Resequencing tutorial: Reference assembly followed by SNP and DIP detection

December 15, 2009

CLC bio
Finlandsgade 10-12 · 8200 Aarhus N · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

Resequencing tutorial: Reference assembly followed by SNP and DIP detection

This tutorial takes you through some of the tools for analyzing a typical resequencing data set from a high-throughput sequencing machine. As an example we use an *E. coli* data set consisting of a little more than 400,000 reads from a 454 sequencer. [1]

Importing the data

First, download the data set from our web site: `http://download.clcbio.com/testdata/raw_data/454.zip`. Unzip the file somewhere on your computer (e.g. the Desktop).

Start the *CLC Genomics Workbench* and import the data:

File | **Import High-throughput Sequencing** | **Roche 454**

This will bring up the dialog shown in figure 1

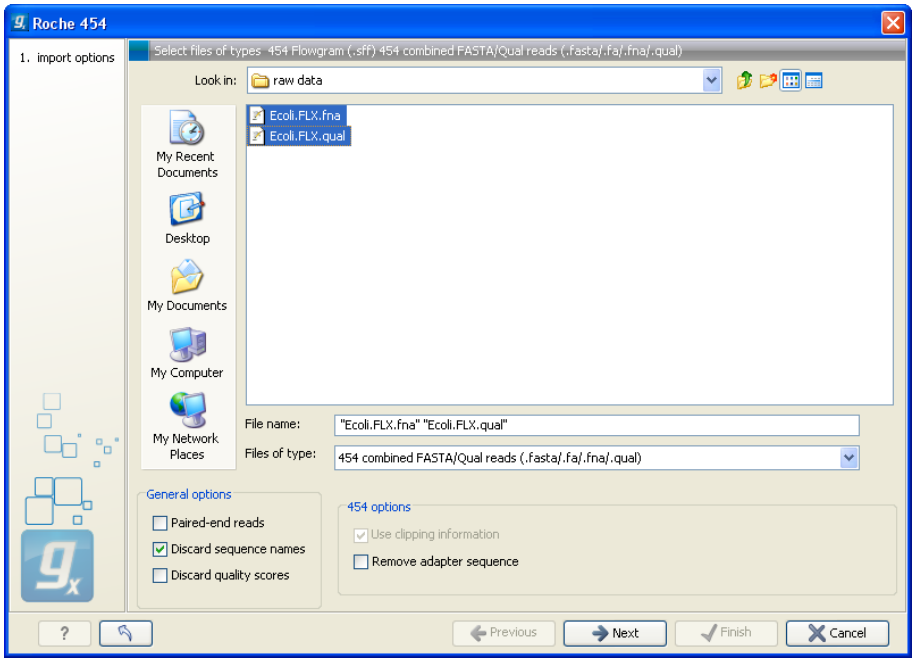


Figure 1: *Choosing what kind of data you wish to import.*

Select the `Ecoli.FLX.fna` and `Ecoli.FLX.qual` files that come from the downloaded zip file. Make sure the **Remove adapter sequence** checkbox is checked and that the **Paired-end reads** checkbox is NOT checked. The option to discard read names is not significant in this context because of the relatively small amount of reads. Click **Next**, choose to **Save** and click **Finish**.

After a short while, the reads have been imported. Next, import the reference genome sequence also included in the zip file:

File | **Import (****)** | **Locate "NC_010473.gbk"** | **Select**

[1] Note that there are special system requirements for *CLC Genomics Workbench*, see `http://www.clcbio.com/index.php?id=72`

Note that this is a genbank file imported using the "normal" import tool. Next-generation sequencing data needs to be imported using the special tool in the toolbox because they have a more complex structure (e.g. in this case two files - one with sequence and one with quality scores).

Mapping the reads to the reference

First step in the analysis is to map the reads to the reference genome:

> **Select the Ecoli.FLX sequence list** | **Toolbox** | **High-throughput Sequencing** ( ) | **Reference Assembly** ( )

This shows the dialog in figure 2).

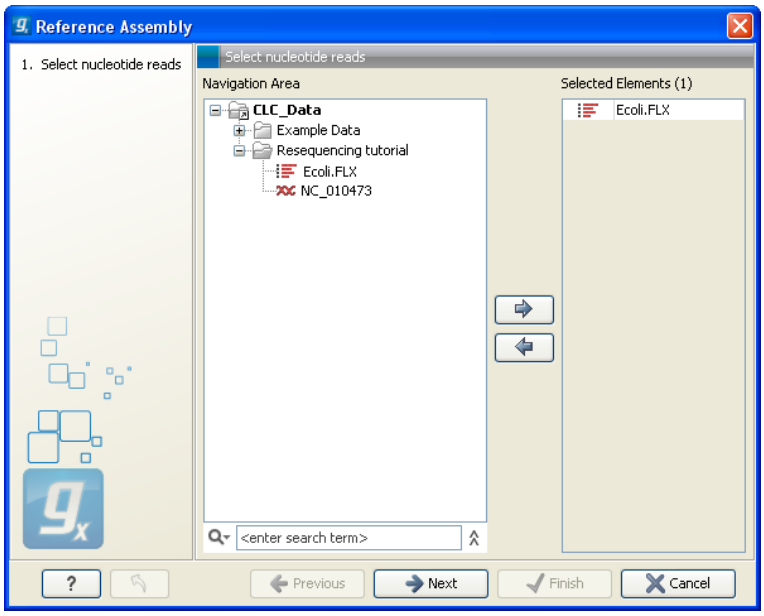


Figure 2: *Select sequence list containing the reads. The reference sequence will be selected in the next step.*

Select the `Ecoli.FLX` ( ) sequence list and add it to the panel to the right. Clicking **Next** will allow you to select a reference sequence as shown in figure 3.

At the top you select the `NC_010473` ( ) by clicking the **Browse and select element** ( ) button. You can select either single sequences or a list of sequences as reference sequences, but in this case just select this single genome sequence.

Clicking **Next** until you see the dialog shown in figure 4 (for more information about the other settings, please click the **Help** (?) button).

Here you can select between different output options. Choose to **Create report** and **Create list of unassemble sequences**. Click **Finish**.

Interpreting the reference assembly results

You can follow the progress of the assembly both in the status bar at the bottom left corner and the log (if you have chosen to show the log in figure 4). When the process is done, you will see

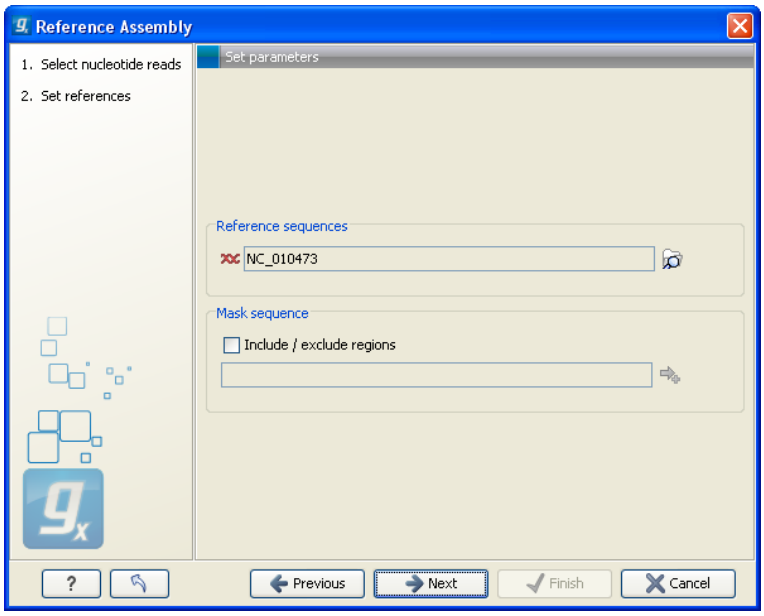


Figure 3: *Specifying the reference sequences and masking.*

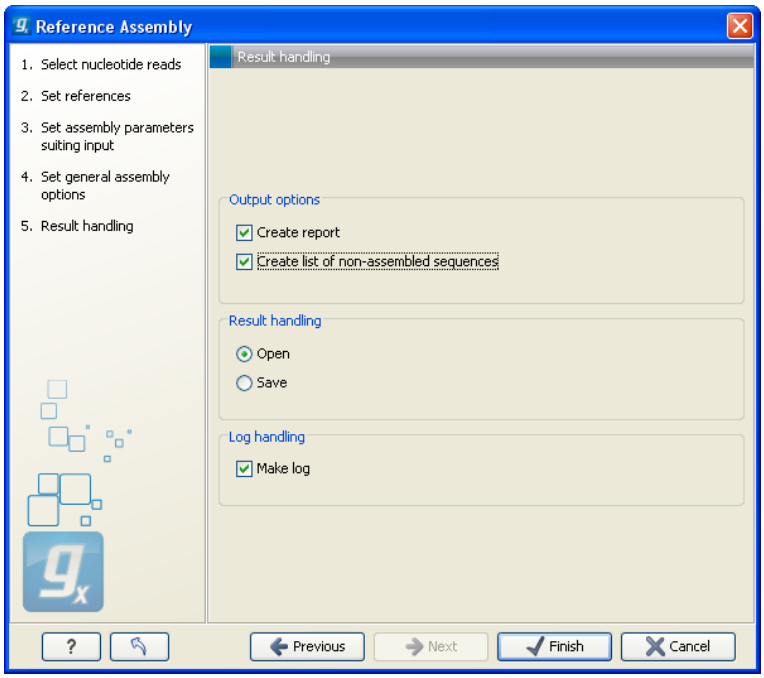


Figure 4: *Output options.*

the following results:

List of non-assembled reads These are the reads that did not match the reference sequence. You can use this list to investigate contamination in the sample or structural differences between the sequencing data and the reference sequence. Typically you will do a *de novo* assembly of these reads and then use BLAST to investigate the contigs (there is a separate tutorial showing how to do this).

Report The report shows information about the assembly. Most importantly, it shows the number of reads that matched the reference sequence.

Contig The contig itself shows the alignment of all the reads to the reference. We are going into details with the contig now.

The contig (shown in figure 5) displays the reads including quality scores. For annotated reference sequences like this, you can display the translation of the coding regions (the yellow CDS annotations) in the Side Panel in the **Nucleotide info** group under **Translation**. (If you scroll a little to the right, you can see the first translation).



Figure 5: *The contig with the reads mapped to the reference.*

You can use the zoom buttons for zooming in and out, and you can scroll along the contig (both horizontally and vertically to inspect the reads covering the region of the view). The **Side panel** includes a lot of settings to change the appearance of the contig, and for looking at these large contigs, there are a few that are good to know:

Sequence layout – Compactness This is used to determine the height of the reads (show/hide the sequence and quality information). This is very useful to get an overview when the reads themselves are not important.

Alignment info – Coverage Displays a graph of the coverage along the contig. There are more relevant graphs under the **Alignment info** group.

The read colors are green (forward) and red (reverse) by default.

Before we continue be sure to **Save** ( ) the contig.

Looking for SNPs

One way of inspecting the result is of course to go through the contig manually, guided by the annotations on the reference sequence. Given the huge amounts of data, manual inspection is only advisable if you have special loci that you are interested in. For a more systematic approach, the *CLC Genomics Workbench* provides two tools: SNP and DIP detection to help you get an overview of the differences between the reference sequence and the reads.

We first look at SNP detection:

Toolbox | **High-throughput Sequencing** ( ) | **SNP detection** ( )

This opens a dialog where you can select the contig you saved before.

Clicking **Next** will display the dialog shown in figure 6

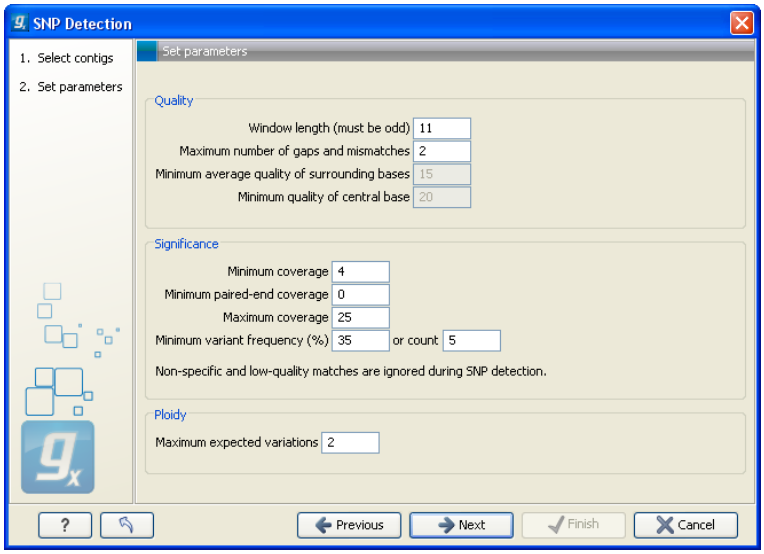


Figure 6: *SNP detection parameters.*

The default settings work well for this data set, but it is important to notice two of the settings that you should always consider when working with your own data: If the **Minimum coverage** is set to 4 but you have a contig with an average coverage of 2, a lot of potential SNPs will not be reported. The **Minimum variant frequency** should also be set, especially when working with non-haploid organisms (e.g. for diploids, it should be set below 50 % in order to report heterozygote SNPs). For this analysis we leave the setting at 35 %.

For more information about the settings, click the **Help** (?) button.

Click **Next** and **Finish**.

The result of the SNP detection will now open a table as displayed in figure 7 below the contig.

By clicking in the table, you can browse through the SNPs. You can **Copy** ( ) the contents of the table and paste into e.g. a spreadsheet for further processing. To select all the rows in the table, press Ctrl + A (⌘ +A on Mac). You can also **Export** ( ) the table in Excel format.

You can also use the filter at the top to sort the table, e.g. to show only nonsynonymous SNPs (filter the **Amino acid change** column to not being empty as shown in figure 8).

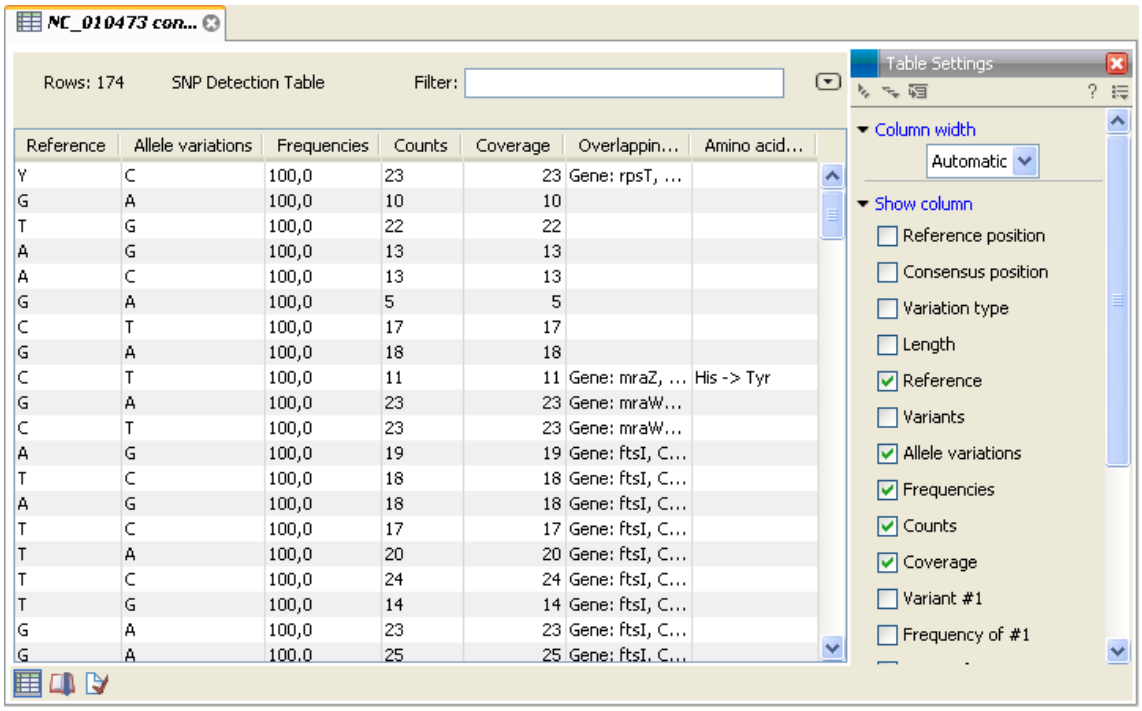


Figure 7: *A table of SNPs.*

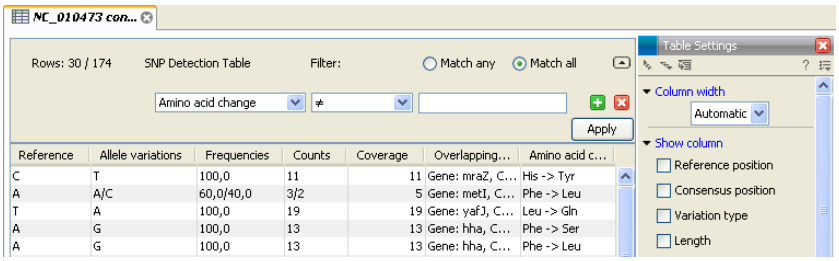


Figure 8: *Filtering the SNP table to only display nonsynonymous SNPs.*

Looking for DIPs

Besides looking for SNPs, you can also do a systematic search for DIPs (Deletion-Insertion Polymorphisms). If you have high coverage in your assembly, you will often find a lot of gaps in the consensus sequence. This is because just a single insertion in one of the reads will cause a gap in all other sequences at this position. The majority of all these gaps should simply be ignored, because they are introduce by single or very few reads because of sequencing errors. The automated DIP detection can be used to find the ones that are significant.

If you want to use the consensus sequence for other purposes, you can simply ignore all the gaps (they will disappear once the consensus sequence gets out of the contig), and the significant ones can then be annotated as DIPs:

Toolbox | **High-throughput Sequencing (**) | **DIP detection (**)

This opens a dialog where you can select the contig you saved before.

Clicking **Next** will display the dialog shown in figure 9

The default settings work well for this data set.

For more information about the settings, click the **Help** (?) button.

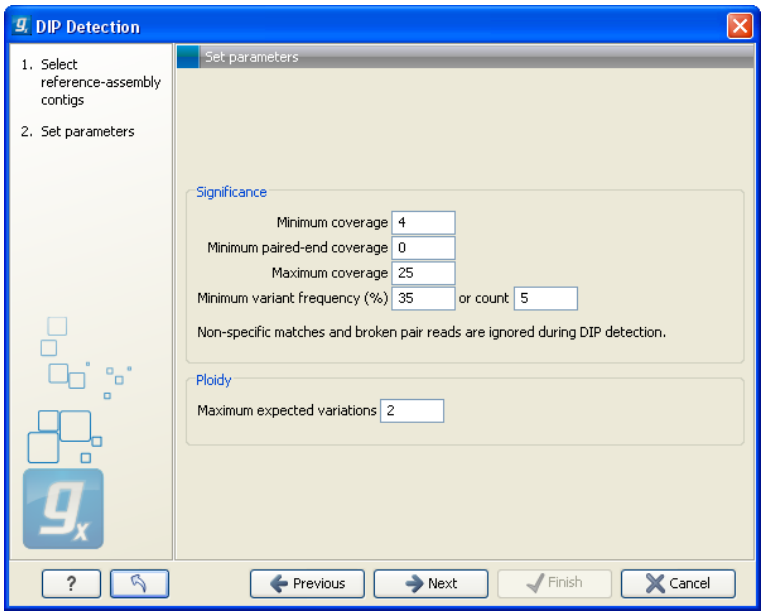
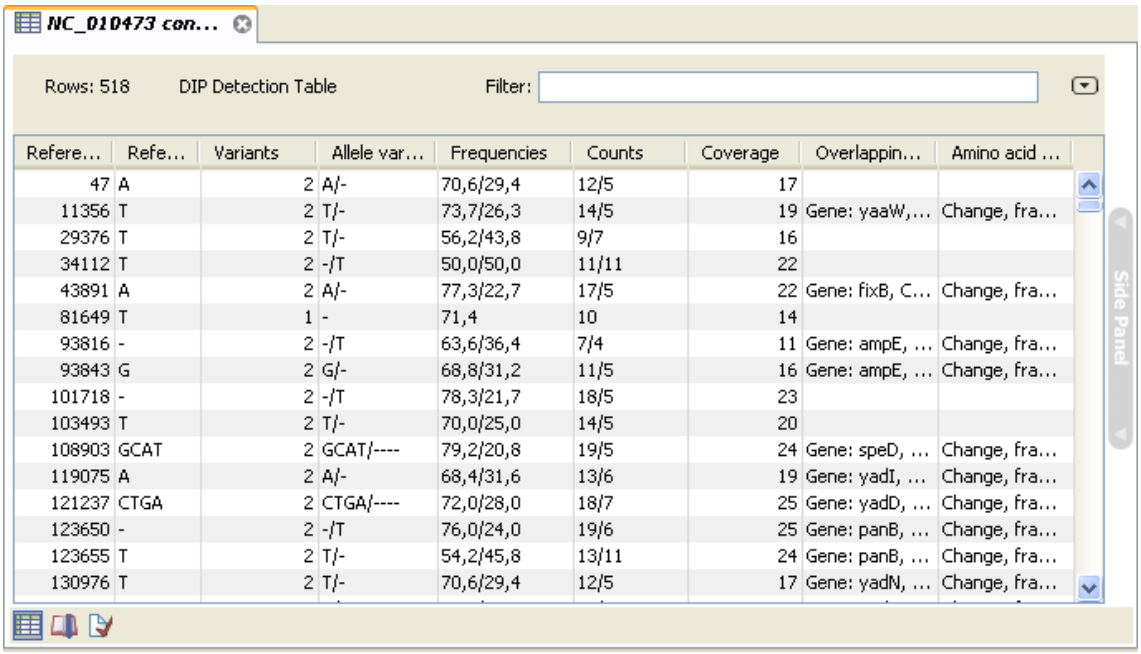


Figure 9: *DIP detection parameters.*

Click **Next** and **Finish**.

The result of the DIP detection will now open a table as displayed in figure 10 below the contig.



Refere...	Refe...	Variants	Allele var...	Frequencies	Counts	Coverage	Overlappin...	Amino acid ...
47	A	2	A/-	70,6/29,4	12/5	17		
11356	T	2	T/-	73,7/26,3	14/5	19	Gene: yaaW,...	Change, fra...
29376	T	2	T/-	56,2/43,8	9/7	16		
34112	T	2	-/T	50,0/50,0	11/11	22		
43891	A	2	A/-	77,3/22,7	17/5	22	Gene: fixB, C...	Change, fra...
81649	T	1	-	71,4	10	14		
93816	-	2	-/T	63,6/36,4	7/4	11	Gene: ampE, ...	Change, fra...
93843	G	2	G/-	68,8/31,2	11/5	16	Gene: ampE, ...	Change, fra...
101718	-	2	-/T	78,3/21,7	18/5	23		
103493	T	2	T/-	70,0/25,0	14/5	20		
108903	GCAT	2	GCAT/----	79,2/20,8	19/5	24	Gene: speD, ...	Change, fra...
119075	A	2	A/-	68,4/31,6	13/6	19	Gene: yadI, ...	Change, fra...
121237	CTGA	2	CTGA/----	72,0/28,0	18/7	25	Gene: yadD, ...	Change, fra...
123650	-	2	-/T	76,0/24,0	19/6	25	Gene: panB, ...	Change, fra...
123655	T	2	T/-	54,2/45,8	13/11	24	Gene: panB, ...	Change, fra...
130976	T	2	T/-	70,6/29,4	12/5	17	Gene: yadN, ...	Change, fra...

Figure 10: *A table of DIPs.*

By clicking in the table, you can browse through the DIPs. You can **Copy** ( ) the contents of the table and paste into e.g. a spreadsheet for further processing. To select all the rows in the table, press Ctrl + A (⌘ +A on Mac).

You can also use the filter at the top to sort the table.

Note: 454 data has an inherent flaw when it comes to homopolymer stretches. This means that the exact number of bases reported in a homopolymer stretch is not always right. This gives rise

to a lot of small insertions and deletions in an assembly like this. When you browse through the DIPs in the table, you will see that most of them are single-base insertions or deletions at the end of a homopolymer stretch. You can sort by length to see some of the longer and more interesting insertions/deletions.