

CLC Academy

Training compendium CLC Genomics Workbench

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

CLC bio
a result of science

CLC Academy

Contents

1 Tutorial: De novo assembly and BLAST **3**

1.1 Importing the data 3

1.2 Assembly 3

1.3 BLAST the results 5

2 Resequencing tutorial: Map reads to reference followed by SNP and DIP detection **8**

2.1 Importing the data 8

2.2 Mapping the reads to the E. coli reference 9

2.3 Interpreting the read mapping results 11

2.4 Looking for SNPs 12

2.5 Looking for DIPs 13

CLC Academy

Chapter 1

Tutorial: De novo assembly and BLAST

This tutorial takes you through some of the tools for a typical *de novo* sequencing work flow with a 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]

1.1 Importing the data

If you don't already have this sample data set, 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 1.1

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 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 will be imported.

1.2 Assembly

The first step is to do a *de novo* assembly of the reads. Briefly explained, this means that we want the *CLC Genomics Workbench* to create longer contiguous sequences out of the relatively short reads that are on average 235 bp long:

Toolbox | **High-throughput Sequencing** () | **De Novo Assembly** ()

This shows the dialog in figure 1.2.

Select the `Ecoli.FLX` () sequence list and add it to the panel to the right. Click **Next**, **Next**

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

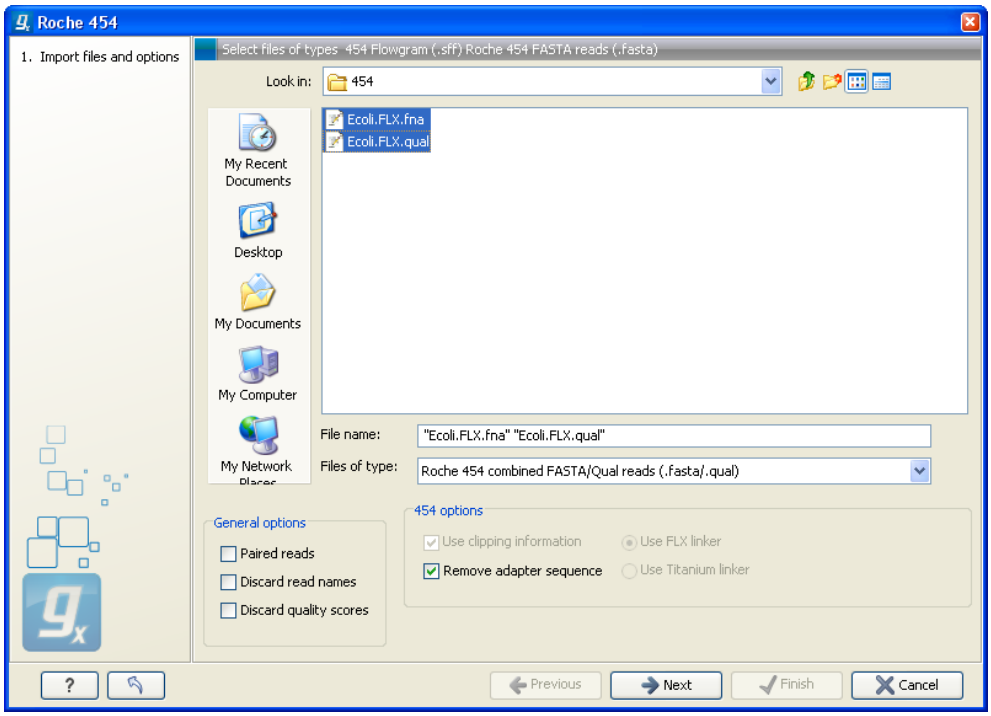


Figure 1.1: *Choosing the file you wish to import.*

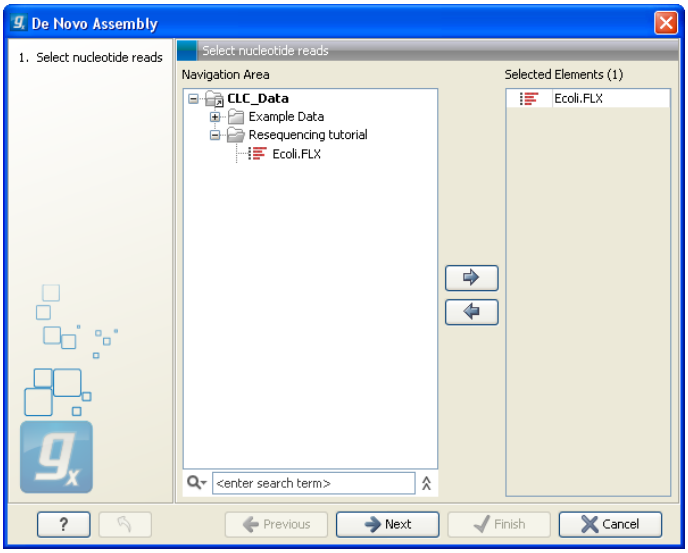


Figure 1.2: *Select sequence list containing the reads.*

and **Finish**.

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

The assembly will take several minutes. There are two steps in the assembly process. In the first step, contig sequences are created by assembling all the reads. Second, all the reads are mapped back using the contig sequences as references. This is done to create an alignment of all the reads (if you only wish to see the contiguous sequences without the reads, you should choose **Create simple contig sequence** before you clicked Finish).

The result of the assembly is a mapping table as shown in figure 1.3.

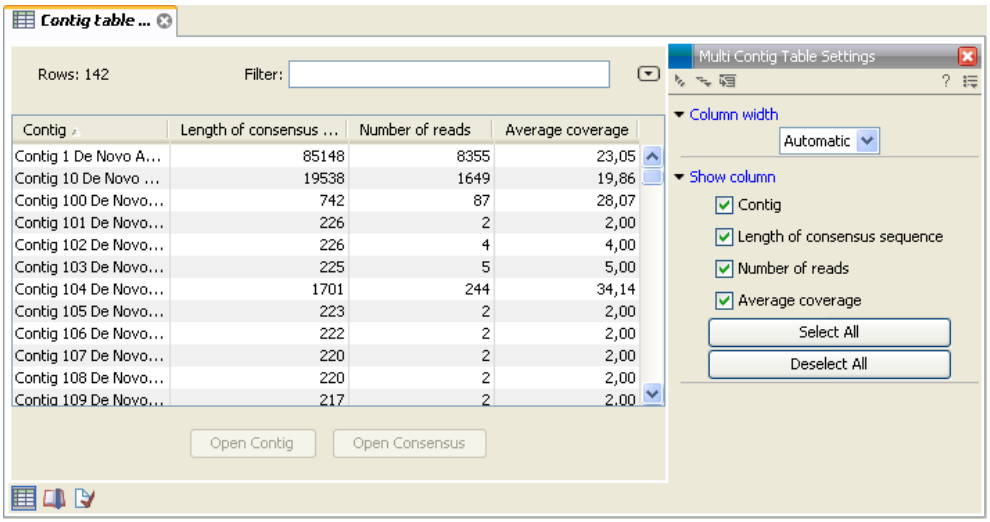


Figure 1.3: *A mapping table.*

Each row in the table represents a contig. There should be around 200 contigs (the number can vary a little from version to version). Double-clicking a row will open the contig with the mapped reads.

1.3 BLAST the results

One way to investigate the contigs would be to see if there are any similar sequences in public databases. This can be done using the BLAST program suite. In this tutorial we will run a BLAST search of five of our contigs against data held at the NCBI. You need to be connected to the internet to do this tutorial.

First, sort the table according to the contig length by clicking the **Length of consensus sequence** column header. Now find the contigs that are just above 10,000 nucleotides. Select the five first as shown in figure 1.4.

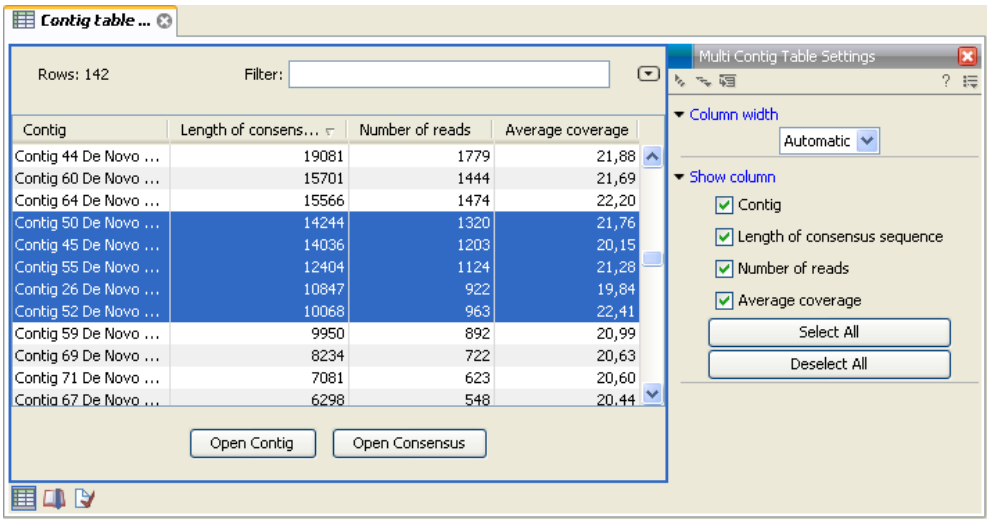


Figure 1.4: *Selecting five contigs that are just above 10,000 nucleotides long.*

Then click the **Open Contig** button at the bottom of the table. That will create a sequence list

with the five contig sequences that you have selected.

To BLAST these sequences, first make sure the sequence list with the five contigs has been **Saved** ( ). Then:

Toolbox | **BLAST Search ( )** | **NCBI BLAST ( )**

Select the sequence list that you have saved and click **Next**. To be completely open in terms of what kind of data there is in the sequences, choose the **Nucleotide collection** database (see figure 1.5).

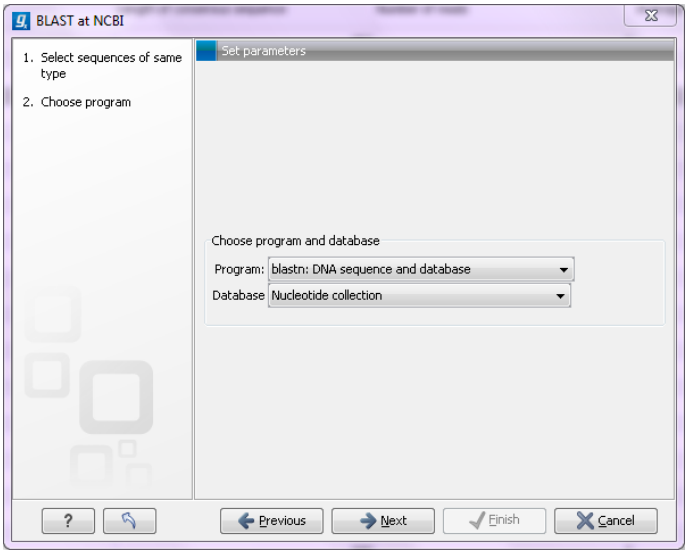


Figure 1.5: *Selecting the Nucleotide collection (nr) database.*

Click **Next** and leave the settings at their default (to make sure the settings are back to default, click the **Reset to CLC Standard Settings** ( ) button). Click **Next** again and **Finish**

The result is shown in figure 1.6.

For each contig, you can see the description of the top hit which is *E. coli* str. K12 in all cases. If you wish to inspect the individual BLAST results in more detail, simply double-click the row and the BLAST result will open. If you wish to use the result for reporting, e.g. in a spread-sheet, 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 (e.g. to only show BLAST results with an E-value lower than a certain threshold).

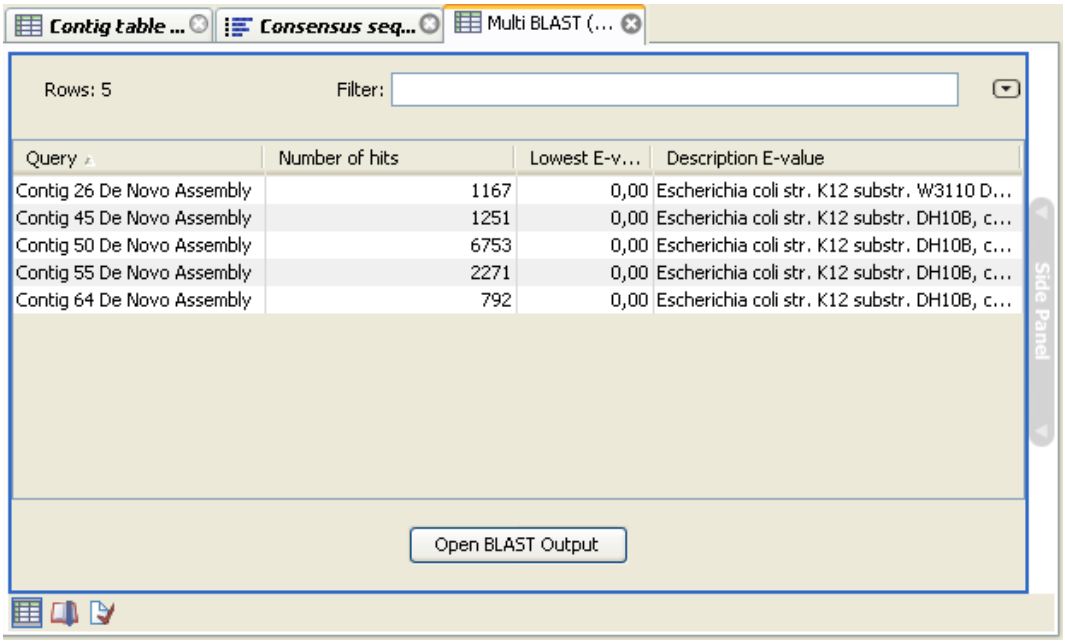


Figure 1.6: *The overview BLAST table.*

CLC Academy

Chapter 2

Resequencing tutorial: Map reads to reference 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]

2.1 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 2.1

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 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**

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).

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

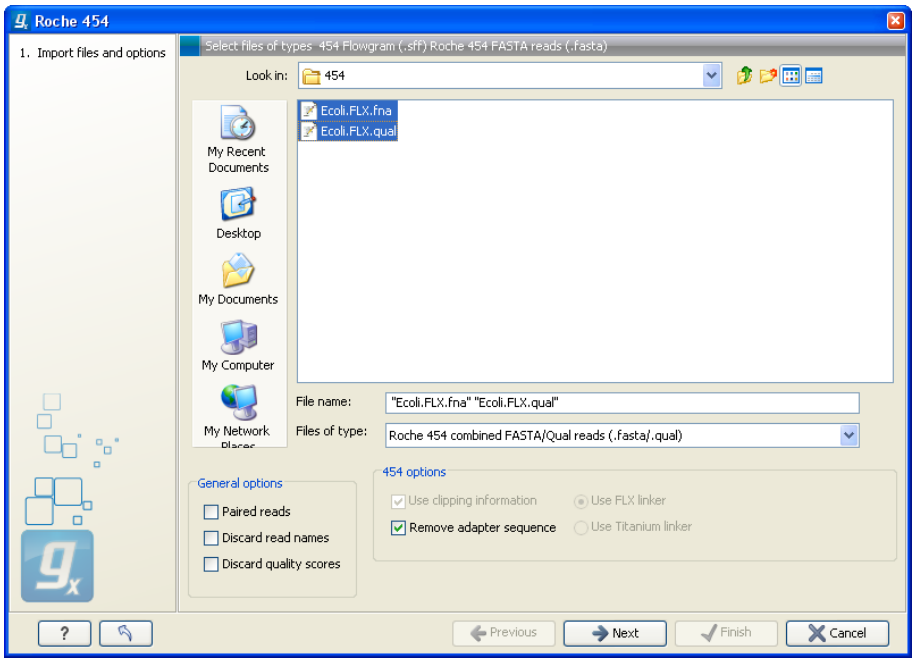


Figure 2.1: *Choosing what kind of data you wish to import.*

2.2 Mapping the reads to the E. coli 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 ( ) | Map Reads to Reference ( )**

This shows the dialog in figure 2.2).

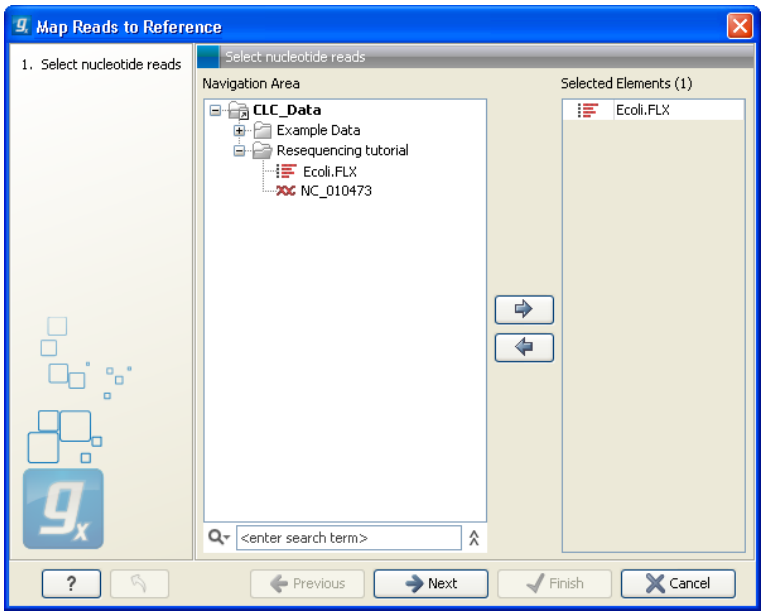


Figure 2.2: *Select sequence list containing the reads. The reference sequence will be selected in the next step.*

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 2.3.

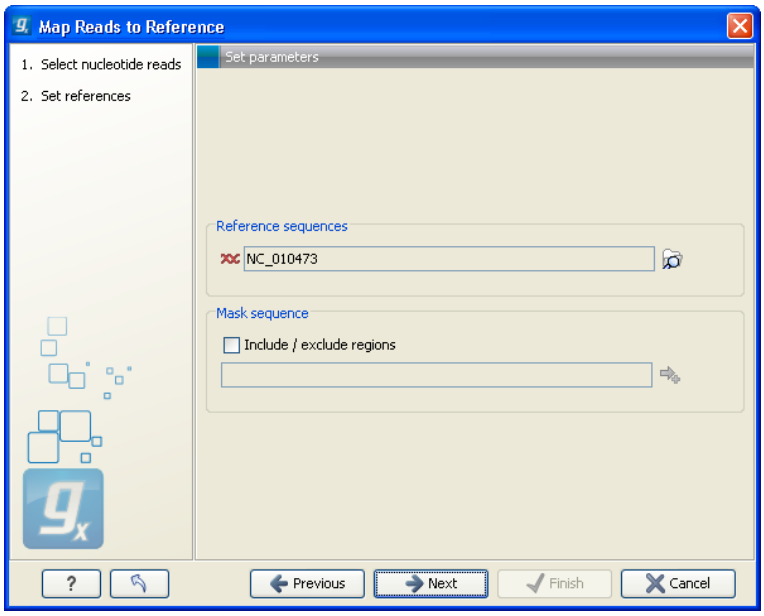


Figure 2.3: *Specifying the reference sequences and masking.*

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 2.4 (for more information about the other settings, please click the **Help** (?) button).

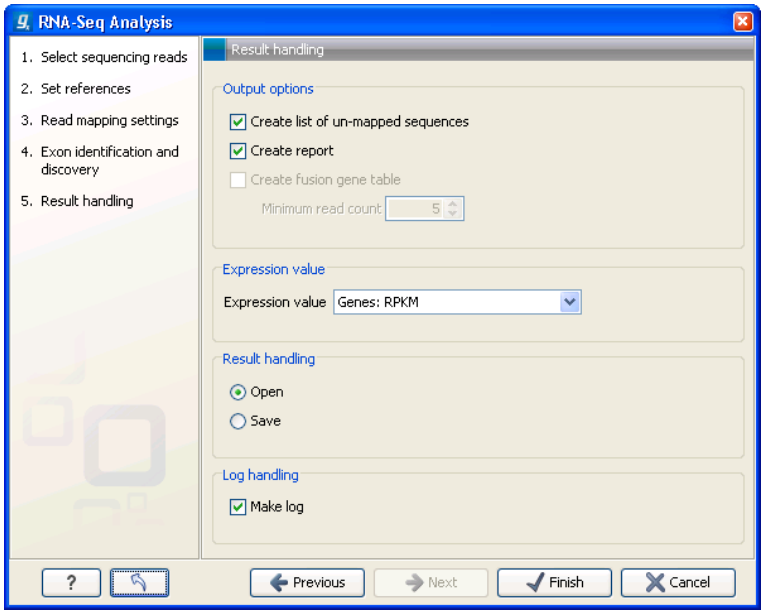


Figure 2.4: *Output options.*

Here you can select between different output options. Choose to **Create report** and **Create list of non-mapped reads**. Click **Finish**.

2.3 Interpreting the read mapping results

You can follow the mapping progress both in the status bar at the bottom left corner and the log (if you have chosen to show the log in figure 2.4). When the process is done, you will see the following results as tabs:

List of non-mapped 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 mapping. Most importantly, it shows the number of reads that matched the reference sequence.

Mapping ( ) The mapping itself shows the alignment of all the reads to the reference. We are going into details with the mapping now. Click the mapping tab ( )

The mapping (shown in figure 2.5) 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 2.5: *The reads mapped to the reference.*

You can use the zoom buttons for zooming in and out, and you can scroll along the reference (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 mapping, and there are a few that are particularly 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. 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 results.

2.4 Looking for SNPs

One way of inspecting the result is of course to go through the mapping 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 mapping result ( ) you saved before.

Clicking **Next** will display the dialog shown in figure 2.6

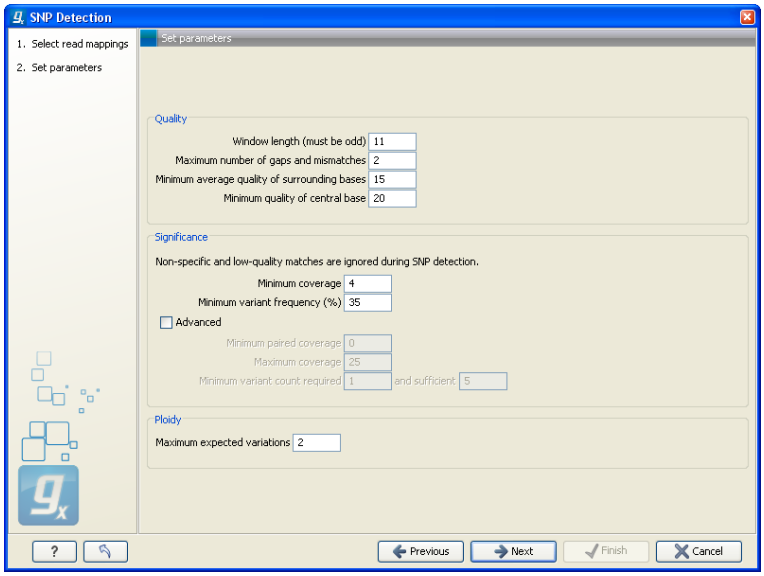


Figure 2.6: *SNP detection parameters.*

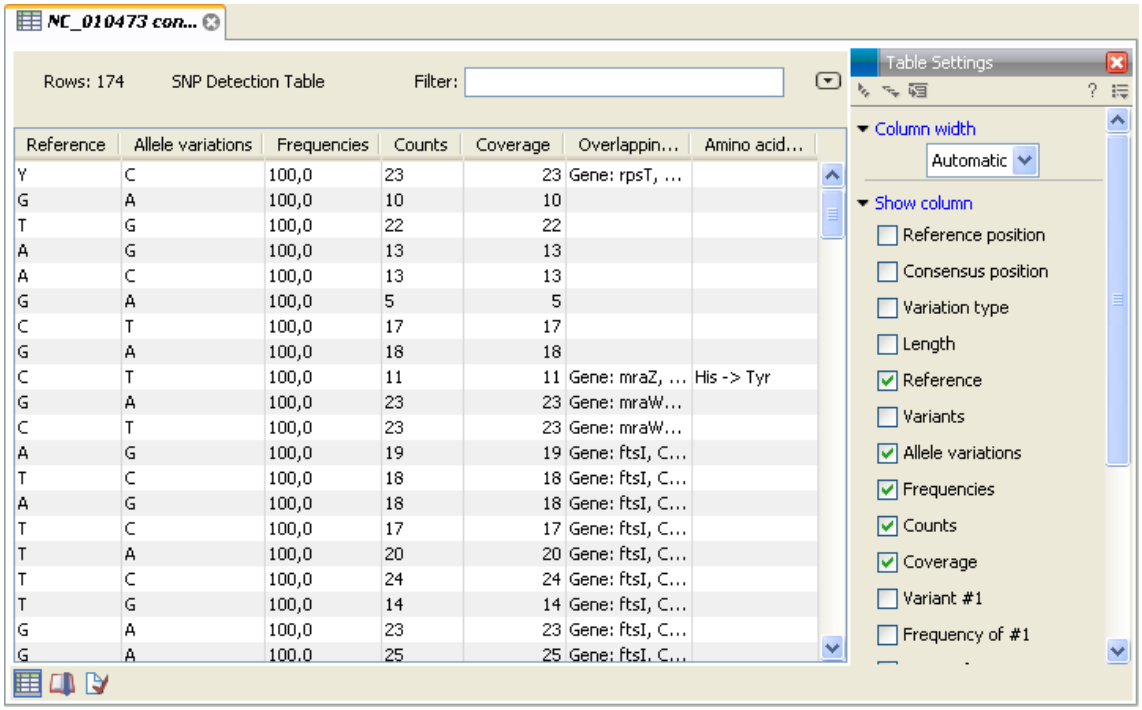
The default settings work well for this data set, but it is important to notice two of the settings that you should always consider when working with your own data: If the **Minimum coverage** is set to 4 but you have a mapping with an average coverage of 2, a lot of potential SNPs will not

be reported. The **Minimum variant frequency** should also be set, especially when working with mixed samples or 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 2.7 below the mapping.



Reference	Allele variations	Frequencies	Counts	Coverage	Overlappin...	Amino acid...
Y	C	100,0	23	23	Gene: rpsT, ...	
G	A	100,0	10	10		
T	G	100,0	22	22		
A	G	100,0	13	13		
A	C	100,0	13	13		
G	A	100,0	5	5		
C	T	100,0	17	17		
G	A	100,0	18	18		
C	T	100,0	11	11	Gene: mraZ, ...	His -> Tyr
G	A	100,0	23	23	Gene: mraW...	
C	T	100,0	23	23	Gene: mraW...	
A	G	100,0	19	19	Gene: ftsI, C...	
T	C	100,0	18	18	Gene: ftsI, C...	
A	G	100,0	18	18	Gene: ftsI, C...	
T	C	100,0	17	17	Gene: ftsI, C...	
T	A	100,0	20	20	Gene: ftsI, C...	
T	C	100,0	24	24	Gene: ftsI, C...	
T	G	100,0	14	14	Gene: ftsI, C...	
G	A	100,0	23	23	Gene: ftsI, C...	
G	A	100,0	25	25	Gene: ftsI, C...	

Figure 2.7: *A table of SNPs.*

By clicking in the table, you can browse through the SNPs. If you open the SNP result table and the read mapping view at the same time in a horizontally split view, the table and mapping view will be linked: double clicking on a row in the SNP table will take to that position in the read mapping view.

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 2.8).

Figure 2.8: *Filtering the SNP table to only display nonsynonymous SNPs.*

2.5 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 mapping, 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 is used outside the mapping), 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 mapping result ( ) you saved before.

Clicking **Next** will display the dialog shown in figure 2.9

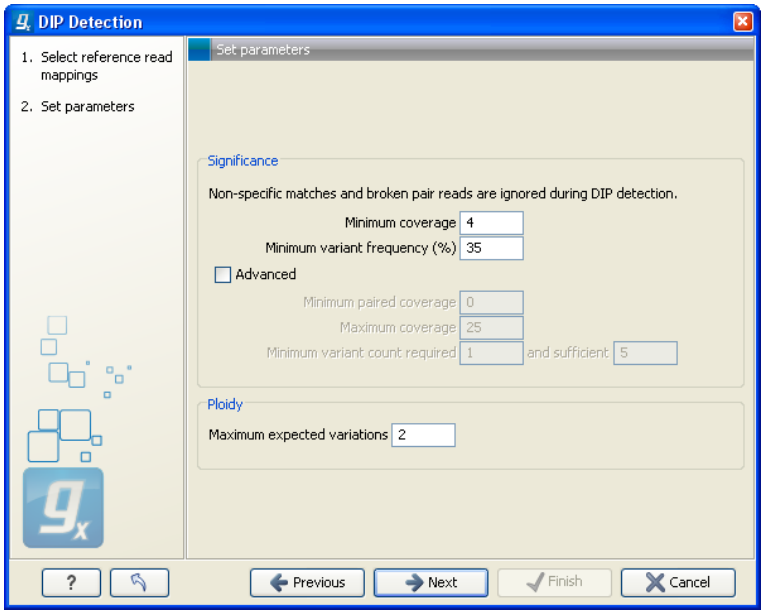


Figure 2.9: *DIP detection parameters.*

The default settings work well for this data set.

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

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

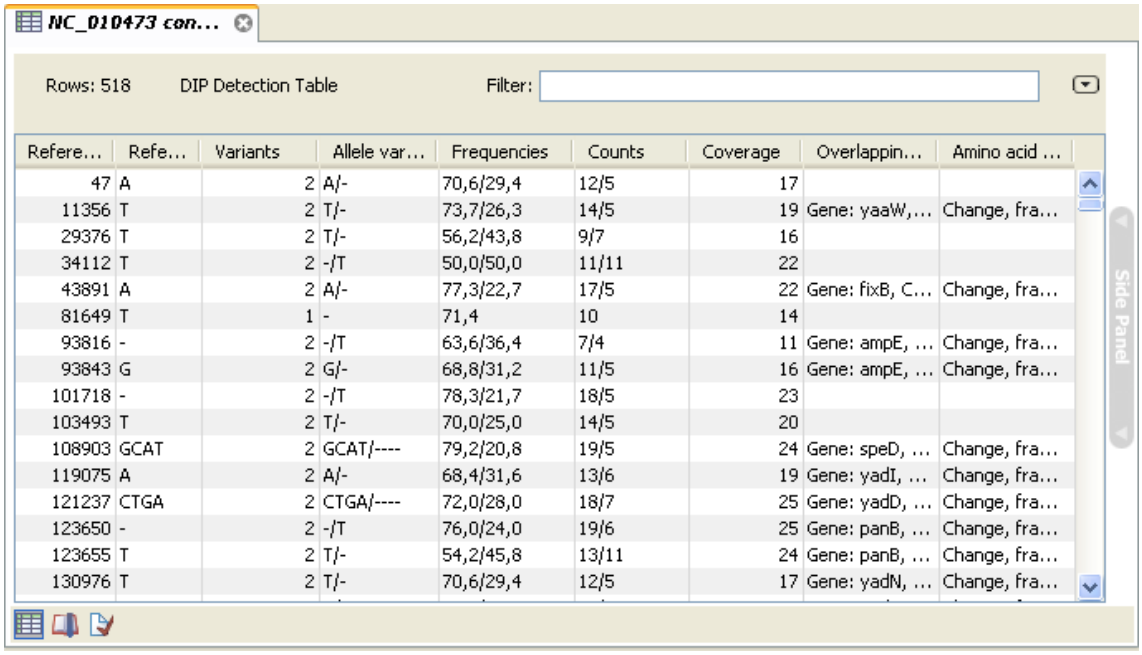
The result of the DIP detection will now open a table as displayed in figure 2.10 below the mapping.

By clicking in the table, you can browse through the DIPs. As with the SNP table, if you open the DIP result table and the read mapping view at the same time in a horizontally split view, the table and mapping view will be linked: double clicking on a row in the DIP table will take to that position in the read mapping view.

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



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 2.10: *A table of DIPs.*

rise to a lot of small insertions and deletions in a data set 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.