Tutorial

Read Mapping in Detail

December 20, 2024

Sample to Insight

Read Mapping in Detail

This tutorial is about running read mappings and provides details about the mapping algorithm. You will learn what the parameters mean, which should help in making good decisions about the analysis of your own data.

We recommend going through the resequencing tutorial first, since it includes the basic introduction to this topic. The tutorial *Resequencing Analysis using Tracks* can be found at https://digitalinsights.qiagen.com/support/tutorials/.

In this tutorial we use a subset of an *E. coli* data set where we focus on a small part of the genome and use single and paired end Illumina GA reads.

Prerequisites   To run this tutorial, you must be working with the *CLC Genomics Workbench*, version 5.5 or higher.

Downloading and importing the data   First, we need to download and import the data.

1. Download the sample data from our website:

   https://resources.qiagenbioinformatics.com/testdata/Read_mapping_in_detail.zip.

2. Start the workbench.

3. To import the data, go to:

   File | **Import (🖳)** | **Standard Import (🖳)**

4. Choose the zip file you just downloaded. Leave the Import type set to **Automatic**.

   You will now have a folder structure like the one shown in figure 1.
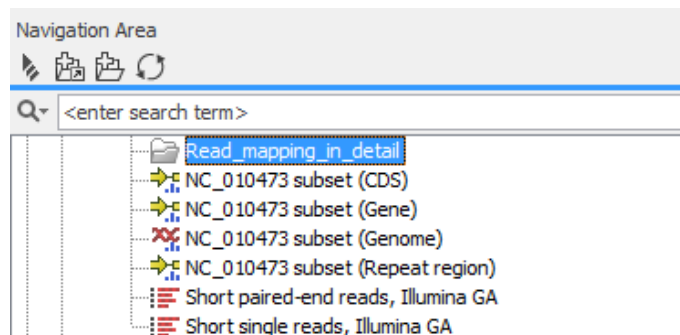

Figure 1: *The Navigation Area upon import of files.*

The data set consists of the **NC_010473 subset (Genome)** sequence track (subset of genome) as well as **CDS**, **Gene** and **Repeat Region** tracks for this genome. Also included are two sequencing data files (**Short single reads, Illumina GA** and **Short paired-end reads, Illumina GA**).

Mapping single reads

To investigate the different mapping parameters of the *CLC Genomics Workbench* we will perform several rounds of mapping of the Illumina single data file. We will have a look at and compare

the different mapping output to get a better understanding of how these parameters influence the result.

Default mapping   First, map the reads with standard settings:

1. To run the Map Reads to Reference tool, go to:

   Tools | Resequencing Analysis ( ) | Map Reads to Reference ( )

2. Select the **Short single reads, Illumina GA** as input. Click **Next**.

3. To select a reference, click the **Browse button ( )** and select **NC_010473 sub-set(Genome)**. Leave the reference masking options to No masking.

4. In the Mapping options dialog, click the **Reset** button to set the parameters back to default. Then adjust the similarity fraction to 0.9 (figure 2). Why we do this will be explained later in the tutorial.
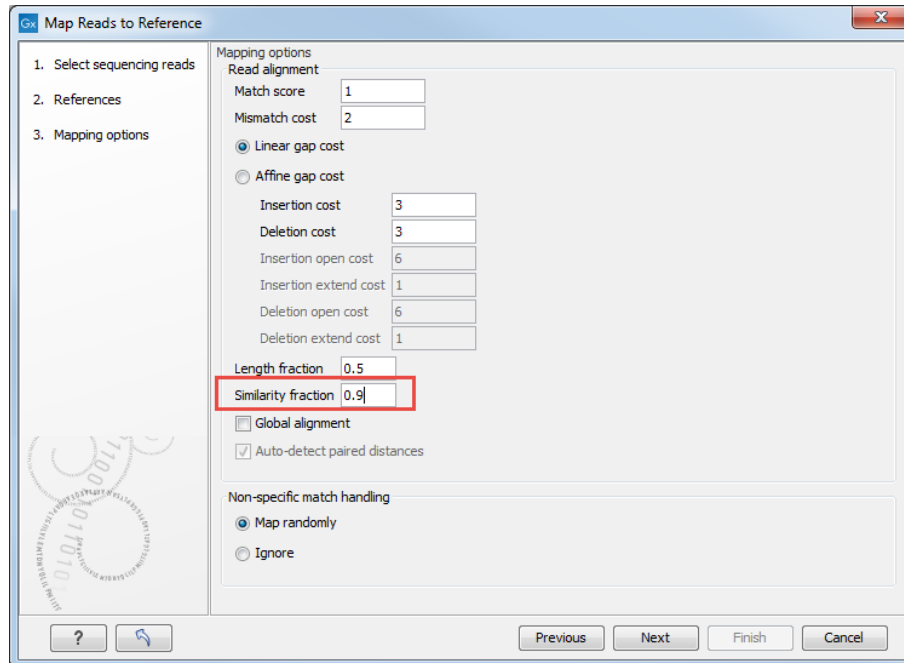

Figure 2: *Adjusting the similarity fraction.*

5. In the result handling dialog, choose to **Create reads track** and to **Save** the result in a new folder (that can be called Single Reads) before clicking **Finish**.

6. Rename the resulting track: click on the name of the track object in the **Navigation Area**, click again or press the **F2** key on your keyboard to rename the track **Default**.

For ease of overview and inspection we will create a track list. This will include both sequence tracks and the read mapping. Later we will add more read mappings to the same list.

1. To create a track list, go to:

   Track Tools ( ) | Create Track List ( )

2. Select the **NC_010473 subset (Genome)**, **(CDS)**, **(Gene)** and **(Repeat region)** tracks as well as the **Default** mapping track (figure 3).
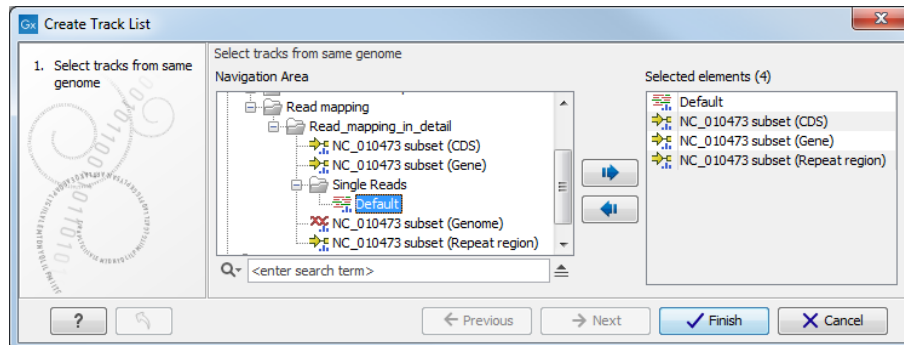

Figure 3: *Creating a track list.*

3. Click **Finish**.

4. Save the track list by dragging the view tab into the **Single Reads** folder in the **Navigation Area**.

Have a look at the Default reads track in the track list. In the right hand settings panel, set the Range to 2020 to 2110 to have a view like the one in figure 4.

Note that there is something special going on just at the border between the two annotated repeat regions. For many of the reads, the end of the read is faded because this part of the read does not match the reference sequence and therefore was not aligned during mapping.

The reason why you see this is because the *CLC Genomics Workbench* performs *local alignment* of the reads. This means that it only aligns the part of the read that matches well - the rest is left unaligned.

Mapping the reads in *CLC Genomics Workbench* is a two-step process:

1. For each read, the optimal local alignment between the read and the reference sequence is found.

2. All reads are filtered according to user-defined criteria for length and similarity of the local alignment.

To illustrate this we will perform two additional mappings with alternate parameter settings.

Mapping with length fraction 1 Next, run a read mapping with the length fraction set to 1.

1. Follow the read mapping steps 1-5 from above, but in the Mapping options, set the **Length fraction** to **1** (and again change the similarity fraction to 0.9 as shown in figure 5).

2. Rename the resulting track **Read length 1**.

3. Drag the saved track object from the **Navigation Area** into the open track list, below the **Default** track.

4. Activate the track list view by clicking the view tab and **Save** the track list.
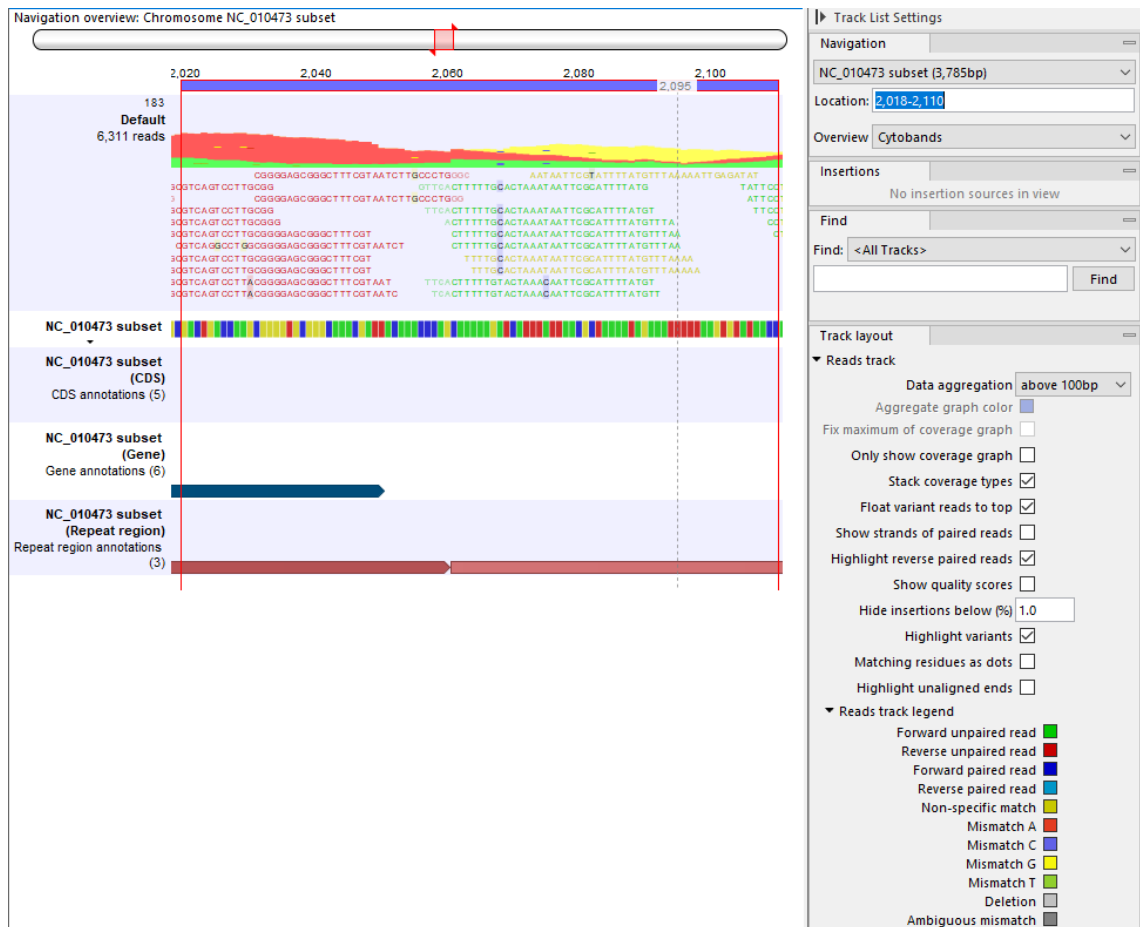
Figure 4: *Region 2020-2110 of the track list holding the Default mapping. Faded read ends signify non-aligned nucleotides.*

A length fraction setting of 1 means that the reads will be filtered so that only alignments matching in their entire length, with the specified similarity fraction, will be included in the mapping. A glance at the **Read length 1** track of the track list illustrates this. As expected, most reads with un-aligned ends are gone (see figure 6), and you are left with less reads. The number of mapped reads can be found below the track label. The reason why we adjusted the similarity fraction to 0.9 is that the read length is short in our example data set, and the default settings would not be stringent enough to reduce the occurrence of reads with unaligned ends.

Mapping with similarity 1   Looking at the **Default** track, you will see that besides the unaligned ends, there are also a number of mismatches in the reads. These mismatches are internal to the local alignment, whereas the unaligned ends are outside the local alignment.

Just as with the unaligned ends, you can control the level of similarity between reads and reference using the **Similarity** setting. To illustrate this, we will run a third mapping with the similarity fraction set to 1.

1. Follow the read mapping steps 1-5 from above, but in the Mapping options, set the **Similarity fraction** to **1** and set the **Length fraction** back to **0.5** (figure 7).

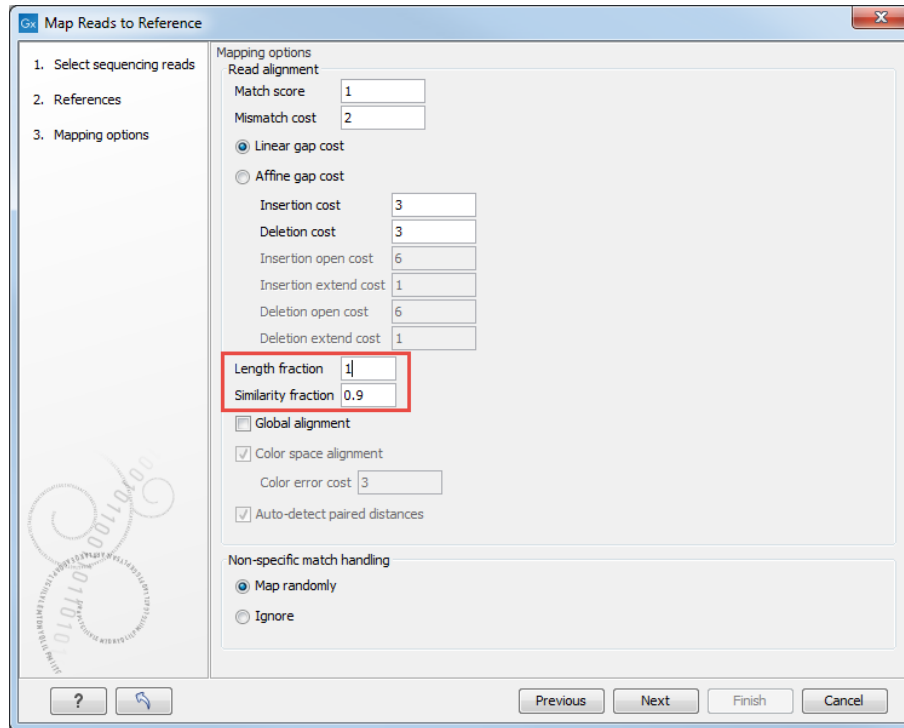2. Rename the resulting track **Similarity 1**.

Figure 5: *Alignment settings for the Read length 1 mapping.*

Figure 6: *Stricter alignment settings result in fewer reads mapped.*

3. Drag the saved track object from the **Navigation Area** into the open track list, below the **Read length 1** track.

4. Activate the track list view by clicking the view tab and **Save (⏎)** the track list.

Setting the Similarity fraction to 1 means that there must be a perfect match between the read and the reference. However, if you look a the **Similiarity 1** track in figure 8 you will find that several mapped reads contain mismatches.

The reason why we see mismatches despite a Similarity fraction of 1 is that the Similarity fraction does not apply to the entire read; it relates to the Length fraction parameter. With a Read fraction of 0.5, this means that at least 50% of the read must have 100% identity.
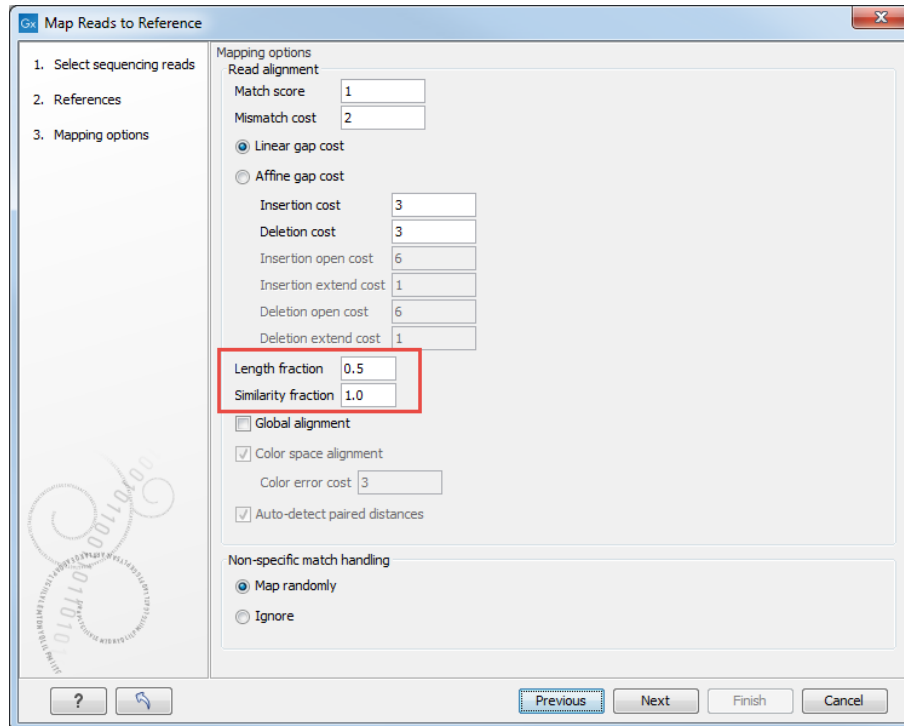
Figure 7: *Alignment settings for the Similarity 1 mapping*

Besides the **Length fraction** and **Similarity fraction**, you can also specify gap and mismatch costs (see figure 7). These determine how the initial local alignment should be performed. Setting a low mismatch cost and high insertion/deletion costs will favor mismatches over gaps in the local alignment. We are not going into details with these settings in this tutorial.

Making use of paired end information

In this section we will investigate the impact that the additional paired end information can have on a mapping.

Until now we have been working with the **Short single reads, Illumina GA** data set, but now we will have a look at the data set **Short paired-end reads, Illumina GA**. This is identical to the single reads except for the additional paired information.

We will perform a mapping with the paired data set and compare the result with the single read mapping.

1. Follow mapping steps 1-5 from the subsection headed *Default mapping*, except for selecting the **Short paired-end reads, Illumina GA** as input.

2. Click the (📁) button to create a new folder. Name the new folder **Paired Reads** and choose to save the results into this.

3. Click **Finish**.

4. Rename the resulting track **Paired**.

5. Create a new track list by going to:

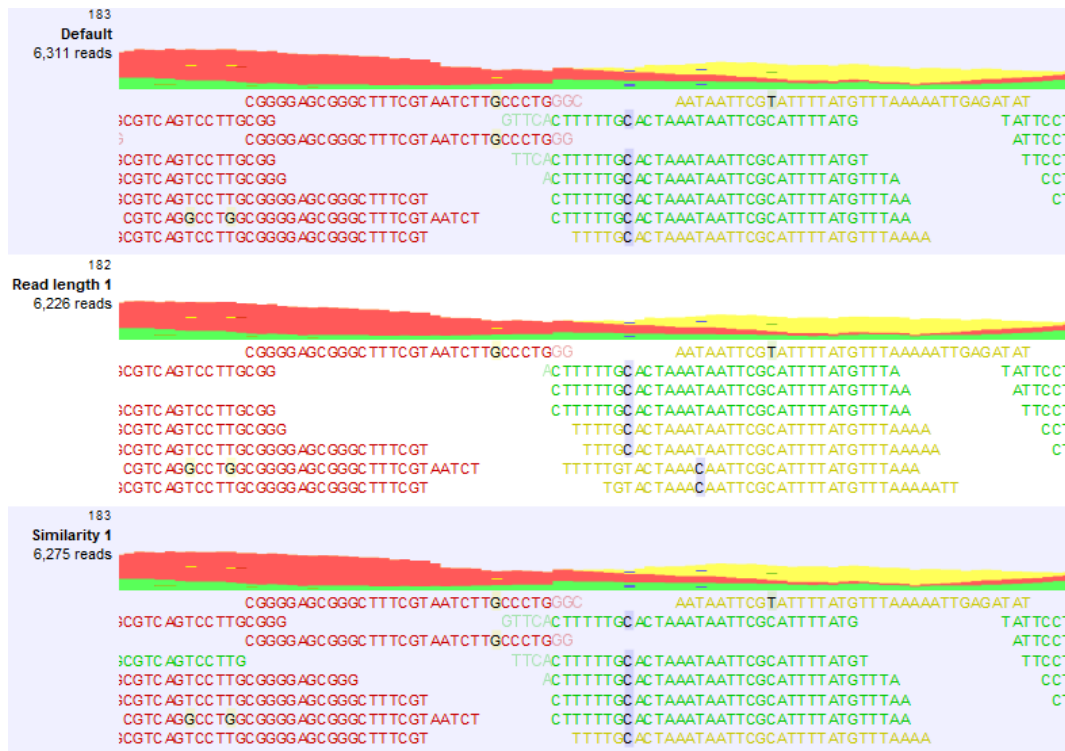   **Tools | Track Tools (📊) | Create Track List (📊)**

Figure 8: *The track list with the Similarity 1 mapping.*

6. Select the **NC_010473 subset (Genome)**, **(CDS)**, **(Gene)** and **(Repeat region)** tracks as well as the **Default** and **Paired** mapping tracks.

7. Click **Finish**.

8. Save the track list by dragging the tab into the **Paired Reads** folder in the **Navigation Area**.

In the right hand settings panel of the open track list set the Range to 1960 to 2210 to have a view like the one in figure 9.

For the **Default** track, you now see a number of yellow reads. The yellow color means that these reads could have been matched equally well to other positions on the reference sequence. They are what we call *Non-specific matches*. We see this because the reference sequence includes repetitive regions.

If you look at the **Paired** mapping track this displays no non-specific reads in this region.

If you zoom all the way out (fit width ⬚) you will see that, in fact, for the entire mapping region we find considerably less non-specific reads with the paired data set (figure 10). This is because the Workbench now has much more information that can be used to place the reads unambiguously: The reads in the single data set are very short – only 35 bp, whereas the paired reads have 70 bp total and span a region of approximately 215 bp of the reference sequence. Although the total number of base pairs are the same in the two data sets, the extra information in the paired read structure produces a mapping of much higher quality (figure 10).
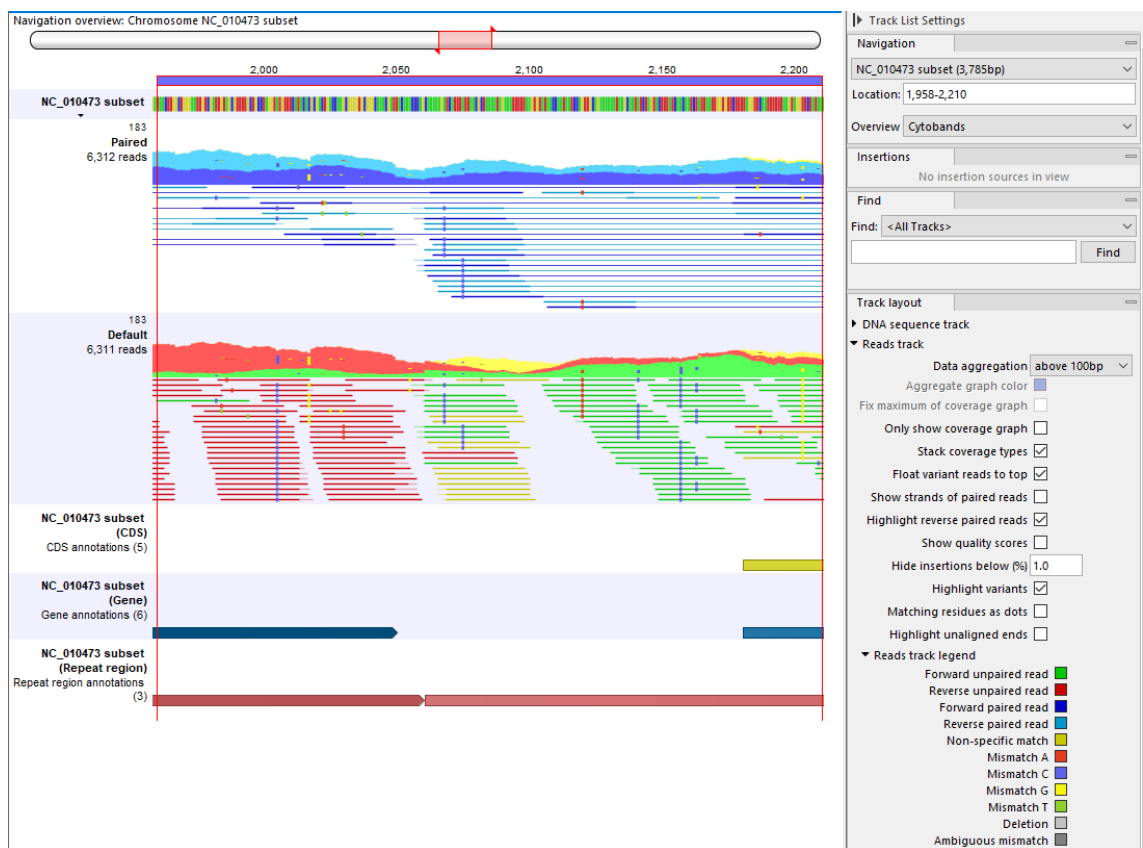
Figure 9: *Comparing the Default and Paired mappings.*

Tutorial

Figure 10: *The number of non-specific reads decreases dramatically when paired information is included.*