



# Tutorial

## Bisulfite Sequencing

November 28, 2018

---

— Sample to Insight —

## Bisulfite Sequencing

This tutorial takes you through a complete bisulfite sequencing analysis, using a collection of tools that appear in the **Epigenomics Analysis** folder in the Bisulfite Sequencing folder.

Bisulfite sequencing is the use of bisulphite treatment of DNA to determine its pattern of methylation. DNA methylation was the first discovered epigenetic mark, and remains the most studied. Changes in cytosine methylation levels are implicated in regulation of gene expression, and have been shown to persist over generations, thus providing mechanistic basis for epigenetic inheritance.

Treatment of DNA with bisulphite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. Thus, bisulphite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single-nucleotide resolution information about the methylation status of a segment of DNA.

In this tutorial we will focus on how to run the analysis and we will not go through the technical details of how the Bisulfite Sequencing analysis is implemented. Please see the CLC Genomics Workbench manual for details on the algorithm.

The workflow goes through the following steps:

- Importing the sequencing data, and a relevant segment of a reference genome
- Mapping the reads to a reference genome using special bisulfite mode
- Calling methylation levels, with simultaneous detection of differential methylation in the two sample types
- Inspecting and interpreting results and reports, and adjusting the parameters accordingly
- Generation of a special track type commonly used in reduced representation bisulfite sequencing
- Final compilation of results in a Track List, and a brief discussion of results
- Creation of a workflow to automate the steps of the analysis

We will look at a subset of a dataset which describes changes in DNA methylation that accompany maturation of blood cells [Hodges et al., 2011]. We will focus on methylation changes in a 2Mb region of human Chromosome 16, around the CD19 gene. We will compare just two sample types, the human stem cell from placenta (HSPCs), and mature b-cells that express the CD19 marker. The comparable analysis in the original paper is shown on figure 1 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412369/figure/F1/> of the original paper.

### Importing prepared sequence and reference data

First, download and save the archive set from our website:

[http://resources.qiagenbioinformatics.com/testdata/bisulfite\\_sequencing\\_tutorial\\_data.zip](http://resources.qiagenbioinformatics.com/testdata/bisulfite_sequencing_tutorial_data.zip) Start the workbench, and create a new folder named, for example, "bisulfite tutorial", where we will keep the data and the analysis results.

To import the archive, use the Standard Import tool:

### File | Import (📁) | Standard Import (📁)

Select the zip file, keep the default option **Automatic import** checked, and press **Next** to choose a folder where the result will be saved.

This will produce a number of new objects in your navigation area, grouped in a folder called 'bisulfite\_sequencing\_tutorial\_data', as shown in figure 1.

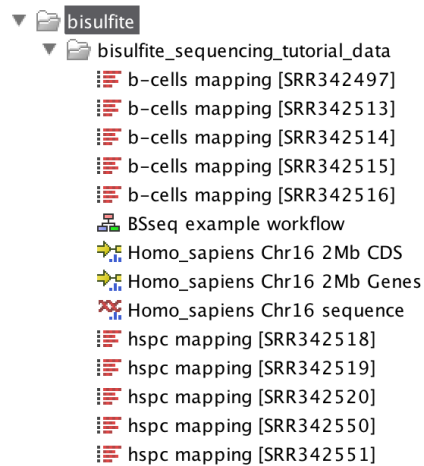


Figure 1: New objects in the Navigation Area after import.

You can now see:

- 10 sequence list objects with names that start with '**b-cells mapping...**' and '**hspc mapping...**'. They are paired end reads that have been downloaded and already imported into a Workbench for you from the SRA archive <http://www.ncbi.nlm.nih.gov/sra/>. Their corresponding accession numbers are in square brackets at the end of their names.
- Reference data (CDS, Gene and Sequence) extracted from human chromosome 16 around the CD19 gene.
- An example workflow that we will examine in the last chapter of this tutorial.

### Mapping the reads to a reference genome using special bisulfite mode

You will now map separately b-cells and hspc reads to a reference genome that was just imported.

1. Start the **Map Bisulfite Reads to Reference** tool, and select one set of reads as shown in figure 2.
2. Select the directionality to "directional" as it is the case of the protocol used in this tutorial.
3. On the next screen of the wizard, select the reference and leave the Reference masking set as "No masking" (figure 3).
4. In the next wizard window, you can leave the parameters for the Read mapping as defaults (shown on figure 4).
5. And finally, tick the box **Create a report** and **Save** your results in the folder "bisulfite tutorial" (figure 5).

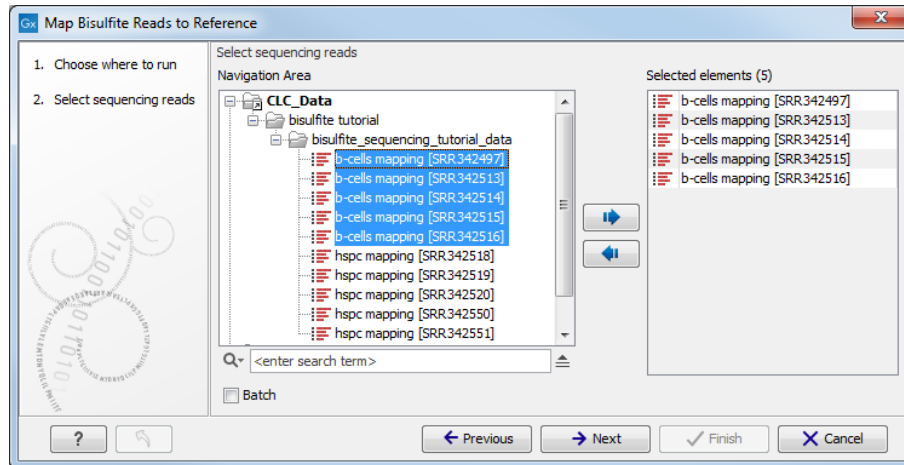


Figure 2: Selection of b-cells reads for bisulfite mapping.

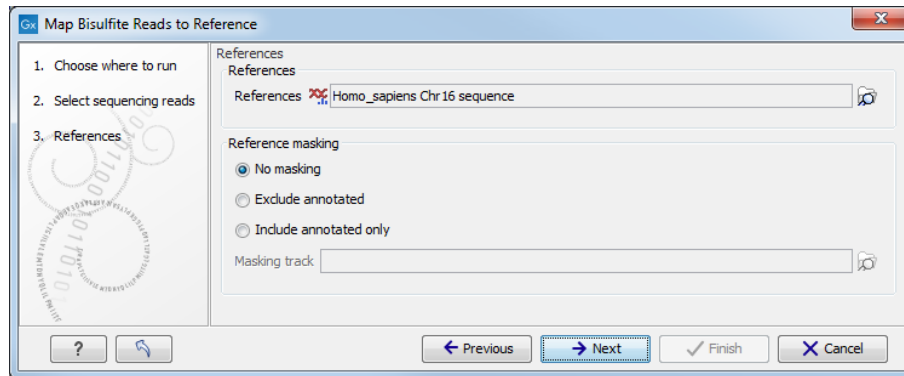


Figure 3: Selection of a reference from imported files for bisulfite mapping.

6. Let the tool run in the background while you restart the wizard and repeat the previous steps for the other subset of reads.

This tutorial dataset contains only a small number of reads (about 100K in each) that should map to only about 2Mb of the genome, so the mapping can finish in less than a couple of minutes. Your results are now in the "bisulfite tutorial" folder as shown in figure 6. Open and examine the mapping reports: close to 100% of reads should map in both cases as unbroken pairs.

### Calling methylation levels with simultaneous detection of differential methylation in the two sample types.

In this section, we will use the tool **Call Methylation Levels** to simultaneously determine methylation status of cytosines in the two mappings we have just produced, and to compare methylation levels between the two samples using Fisher exact test.

1. Start the **Call Methylation Levels** tool and select one of the two mappings (for example b-cells) as input before clicking on **Next** (figure 7):
2. Leave the parameters as default on the "Methylation call settings" window (figure 8):
3. In the wizard window called "Statistical tests and thresholds settings", choose **Fisher exact** as statistical mode, and select the other mapping (for example hspc..) as **Control Reads**

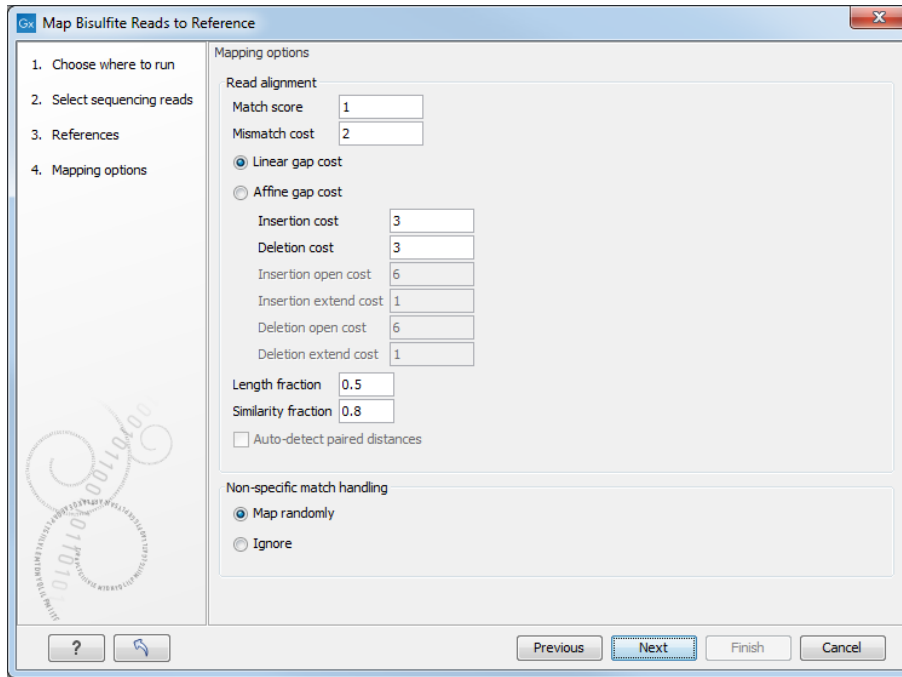


Figure 4: Setting the parameters for bisulfite mapping.

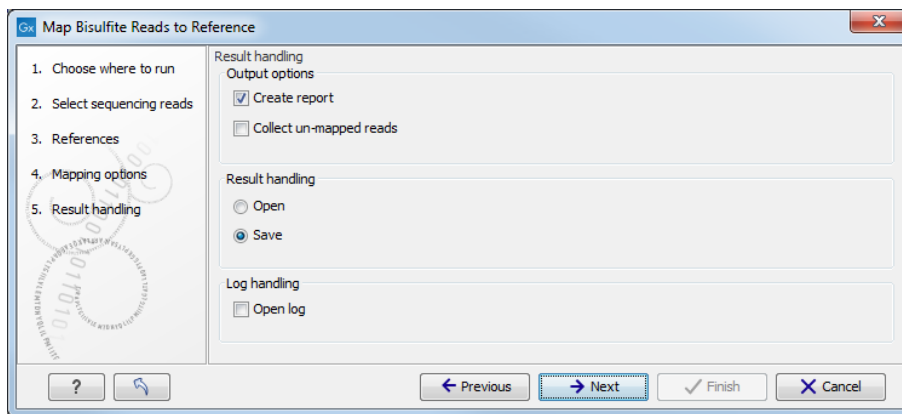


Figure 5: Selecting and saving the results of bisulfite mapping.

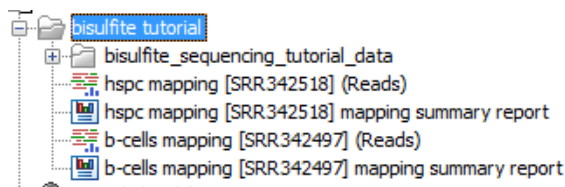


Figure 6: Results of bisulfite mapping.

**Track.** In addition reduce the window length to 200. Leave the other parameters as default (shown in figure 9).

4. Finally, check the boxes **Create track of methylated cytosines** and **Create methylation reports** (figure 10). Save your results in the "bisulfite tutorial" folder.

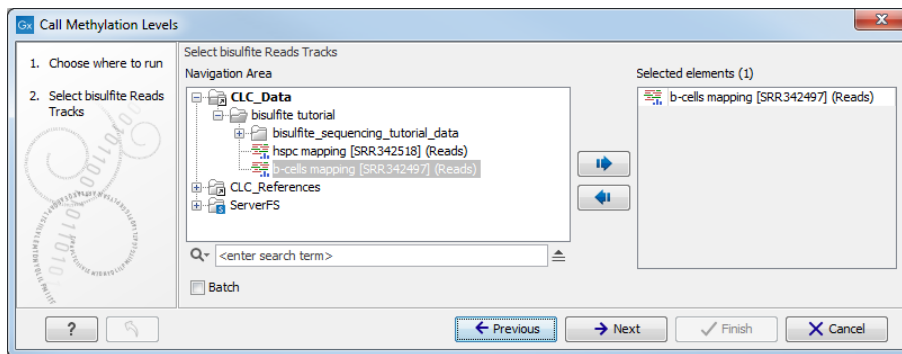


Figure 7: Selecting input to call methylation levels.

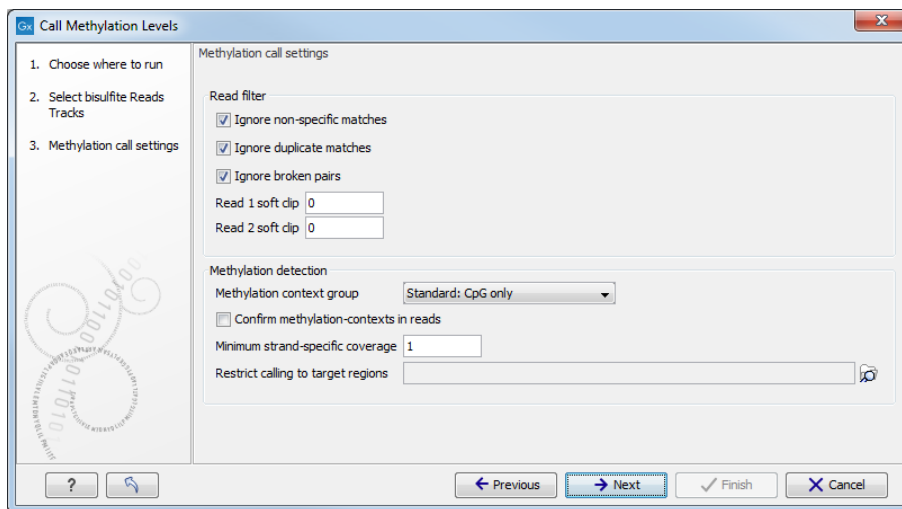


Figure 8: Setting parameters to call methylation levels.

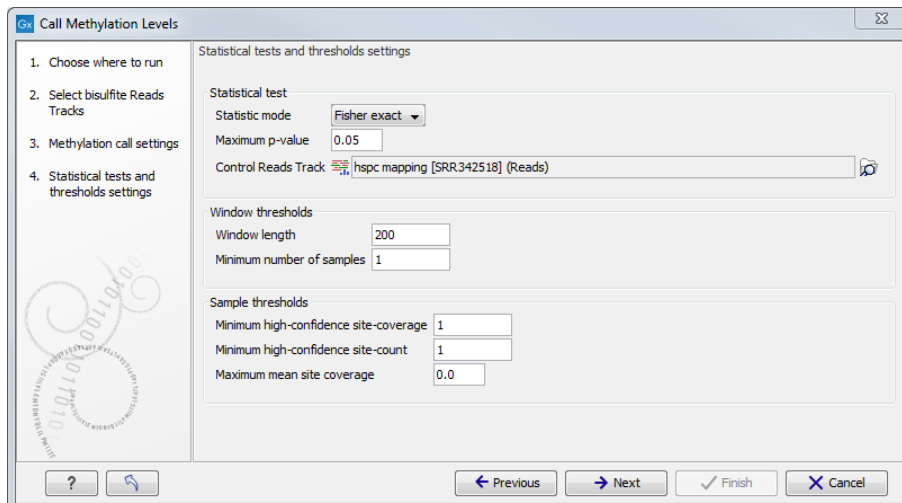


Figure 9: Configuring statistical tests.

**Inspecting and interpreting results and reports, and adjusting the parameters accordingly.**

After the **Call Methylation Levels** tool finishes, your Navigation area should look like figure 11.

Select Differential methylation (CG) (👉), open it to view, and switch to the table view by clicking on (📄) at the bottom of the view area. Verify that there are 246 differentially

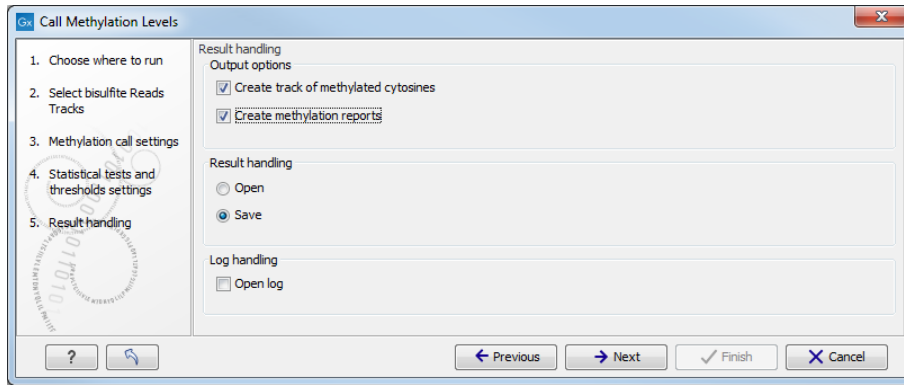


Figure 10: Configuring output options.

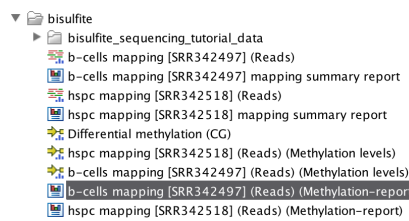


Figure 11: View of Navigation area after calling methylation levels.

methylated regions annotated given the chosen inputs and settings.

We are going to examine the methylation report to see if we can adjust some settings. Open one of the two Methylation-report objects, and inspect the statistics of the mapping in Chapter 4, **Methylation bias**. The figure 12 suggests that there is indeed methylation detection bias in the first few bases of the reads, especially in a second read of each pair, presumably due to end repair during library preparation, which erases methylation marks after bisulfite conversion. This suggests that the parameters **Read 1 soft clip** and **Read 2 soft clip** for the **Call Methylation Levels** tool need to be adjusted.

**4.1 Detects 5-methylated cytosines in CpG contexts.**

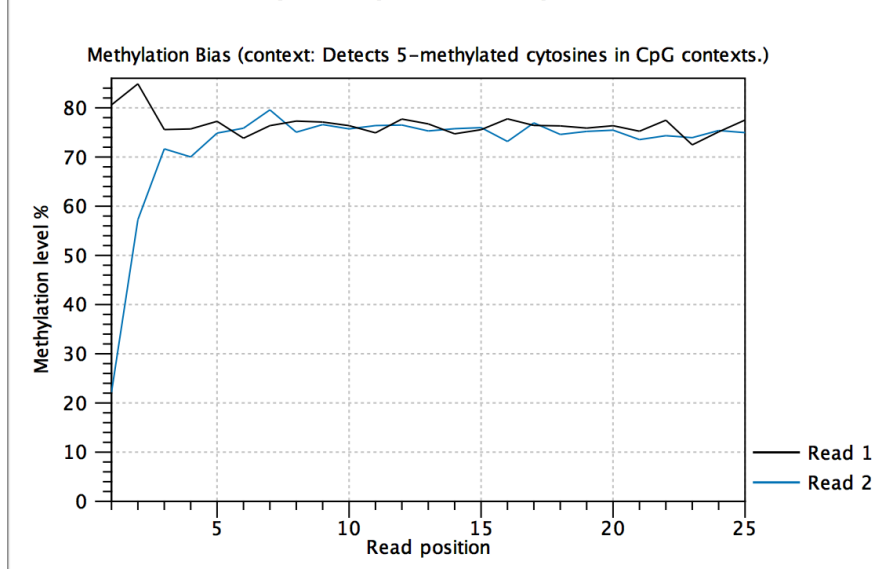



Figure 12: Detection of 5<sup>1</sup>-methylcytosines in CpG contexts.


Please repeat the steps in the Chapter **Calling methylation levels with simultaneous detection**

**of differential methylation in the two sample types**, but this time set the softclip parameter to "5" instead of "0" for both Read 1 and Read 2, and save the results in a subfolder called "soft clip 5".




### Generation of a special track type commonly used in reduced representation bisulfite sequencing.

In the previous exercise, differential methylation levels between the two samples were detected in 200bp windows consecutively along the entire genome. Sometimes it helps to focus the analysis to specific regions where CpG islands are more common. One experimental approach is to sequence and analyse bisulfite reads around certain restriction enzyme sites. The commonly used one is *MspI*, which cuts the sequence  $5' \dots \overset{\text{C}}{\text{C}}\text{GG} \dots 3'$  around CpG islands. Restriction digest fragments are typically size selected in a certain range before being analysed in bisulfite sequencing.

1. Start the tool **Create RRBS-fragment Track**.
2. Select the `Homo_sapiens Chr16 sequence` .
3. Leave parameters as default for the 'settings' wizard window, meaning that the restriction enzyme selected is *MspI*.
4. Save the result.

The resulting object `Homo_sapiens Chr16 sequence (MspI)`  can be used as input in **"Restrict calling to target regions"** option in figure 8. You can repeat the steps in the Chapter **Calling methylation levels with simultaneous detection of differential methylation in the two sample types** but in the rest of this tutorial we will use the results produced without this special track, and will only use the track in a Track List for illustrative purposes.

### Final compilation of results in a Track List, and a brief discussion of results

1. Start the tool Create Track List using the Launch button .
2. Select the input objects as shown in figure 13. Note that we are using the results of the **Map Bisulfite Reads to Reference** tool with soft clip parameters set at 5.
3. Click on the button labeled **Finished**. The track list opens in the View Area of your workbench. You can save it by dragging and dropping the tab in the "bisulfite tutorial" folder in the Navigation Area.
4. Double-click on the title of the `Homo_sapiens Chr16 2Mb Genes`  track in the list of tracks (it should be near the top, but should not be mistaken for the `Homo_sapiens Chr16 2Mb CDS`) to open the table view of the track in a split view. In the filter box type **"CD19"**, as shown in as shown in figure 14:
5. Clicking on a row with the CD19 gene name will automatically select the corresponding region in the browser view. Zoom in to that region by clicking on the  zoom tool, to produce the view shown in figure 15.



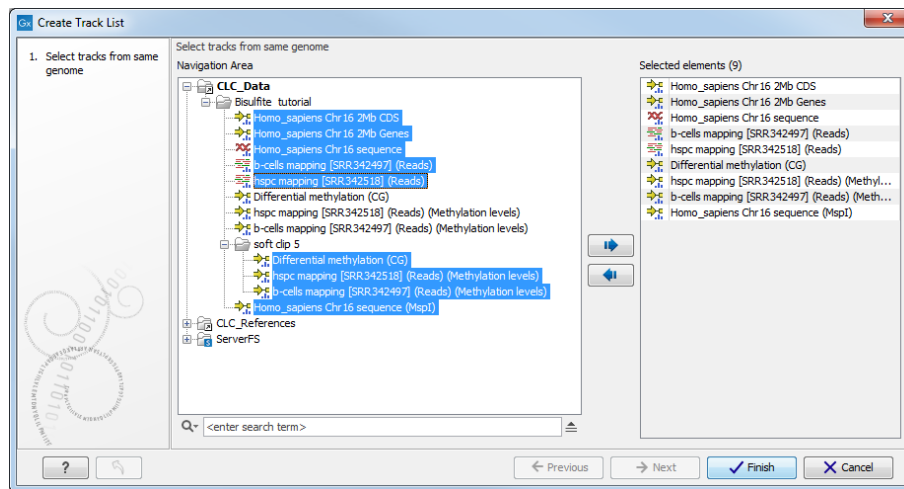


Figure 13: Selection of input objects to create a Track List.

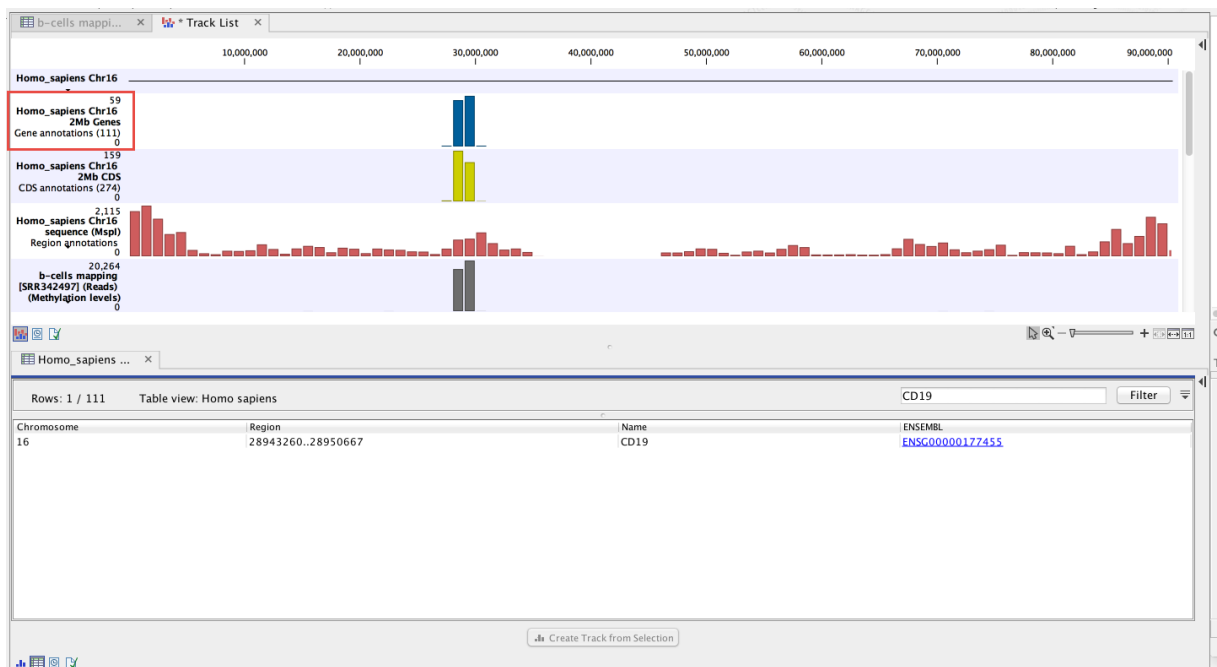


Figure 14: Opening table view of genes, and filtering to select the gene of interest.

6. Inside the gene, there is only one region in the **"Differential Methylation (CG)"** track identified; if you hover the mouse over that region, the detailed information will appear about it, as shown in the figure 15. Recall that to produce that differential methylation result, we used the b-cells sample as input, and the hspc sample as control, therefore the identified region is hypermethylated in the b-cells. Rename the **"Differential Methylation (CG)"** to **"b-cells Hyper Methylation (CG)"** to note this.
7. To identify regions that are hypermethylated in hspc sample compared to the b-cells sample, repeat the steps starting with figure 7, but reverse the input and the control mappings, using hspc as input and b-cells as control. You can also uncheck the **"Create track of methylated cytosines"** and **"Create methylation reports"** options when configuring the output, so that the only output produced will be the **"Differential Methylation (CG)"** track.
8. Rename this newly produced differential methylation track to **"hspc Hyper Methylation**

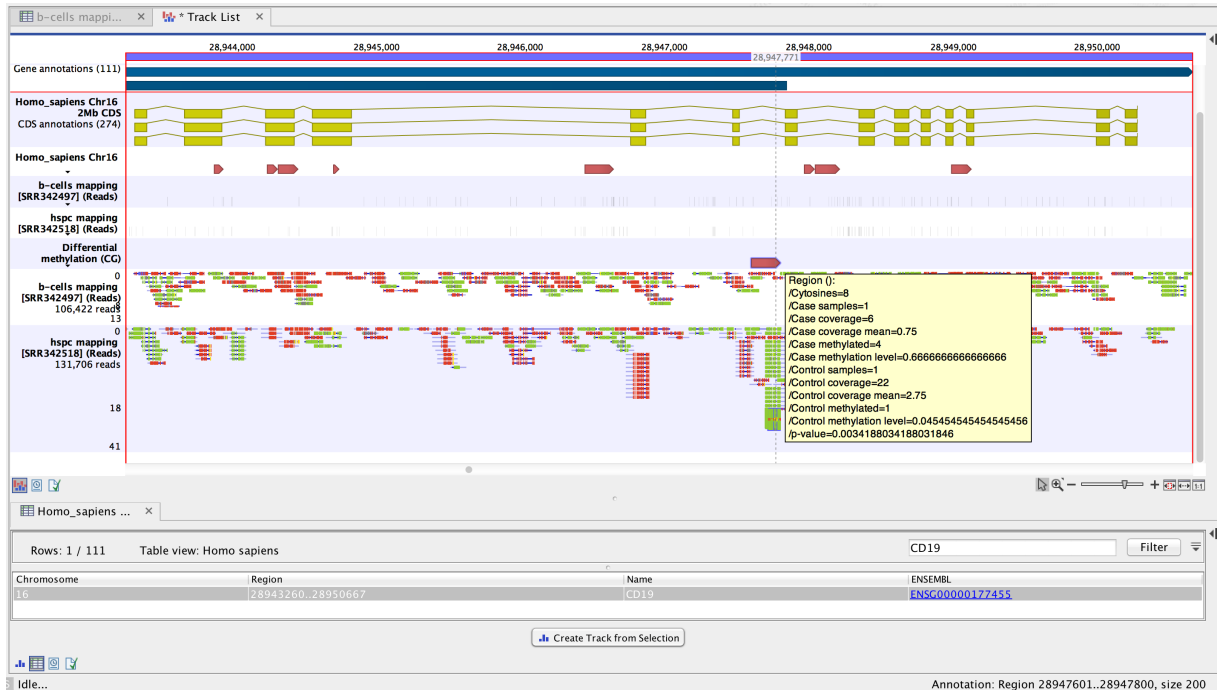


Figure 15: Highlighting the hypermethylated region in the CD19 gene of b-cells.

**(CG)".**

9. Add it to the Track List view by dragging and dropping it from navigation area to the view tab. This should produce the view similar to the one shown in figure 16. You can re-arrange the relative order and size of tracks by dragging them up and down, and by dragging the resizing bar that appears when a mouse is held over the vertical border of tracks.

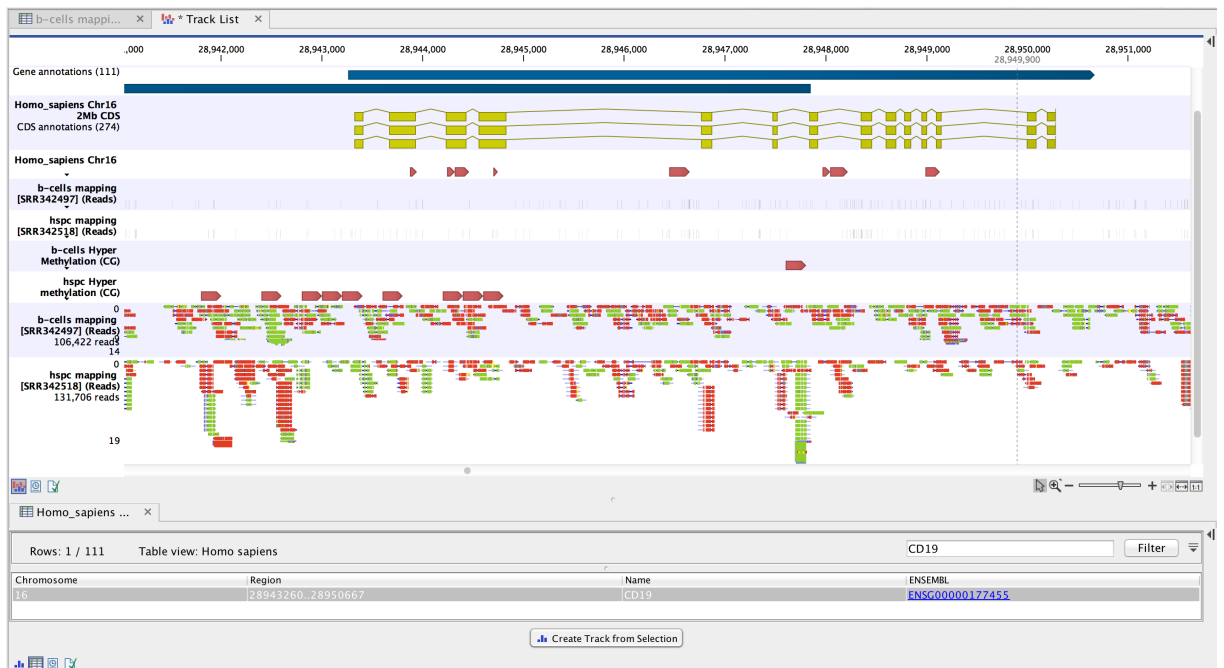


Figure 16: Combining the two directions of methylation level changed in one view.

You will notice that in the 5'-end region of the CD19 gene there are now multiple regions that are hypermethylated in the hspc sample; zoom into any one of them, and examine which individual methylated bases were responsible for the difference, along with the actual read mappings. One example is shown in figure 17:

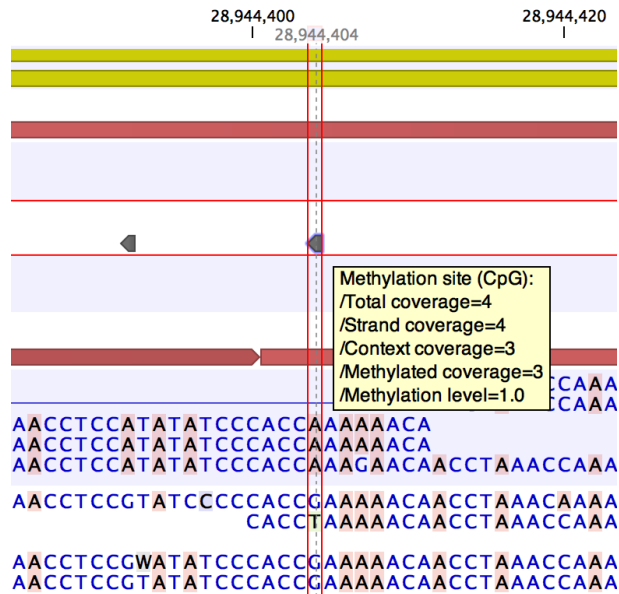


Figure 17: Base-level view of the differential methylation event.

Compare our view of the genome near the CD19 gene to the one in the published paper by Hodges et al, that also annotates the hypomethylation area in b-cell development in the 5'-region of the gene.

### Creation of a workflow to automate the steps of the analysis

The steps of analysis we have done so far can be automated, and executed in a single workflow. This tutorial dataset, when imported, also contains `BSseq example workflow` (📄) that is designed to do this. Open it by double-clicking, and examine its components. The graphical workflow editor view should look similar to figure 18:

Notice that the workflow contains twice the tool for calling differential methylation to conduct the reciprocal analysis of methylation, so that both hyper- and hypo-methylated regions can be identified in one workflow. These two instances are arbitrarily named "Call Methylation Levels" and "reverse methylation", and the order of inputs with regard to reads track and control reads track is reversed. You can edit the parameters of various tools in the workflow by double-clicking on them and going through the corresponding wizards in the view area, or do it when running the workflow wizard. If you are using the first option, parameters can be set and locked to certain values, allowing to run the exact same workflow configuration each time.

Click the "Run..." button to start the workflow wizard. It will ask you to provide the inputs, such as the reference sequence, control reads, and sample reads. It will also offer you to inspect and/or change (depending on lock settings) configurable parameters. Finally, it will offer you to save results, for which we will create a new folder. When the workflow execution completes, which should not take more than a couple of minutes, this new location in your navigation area should contain objects similar to ones shown in this figure 19:

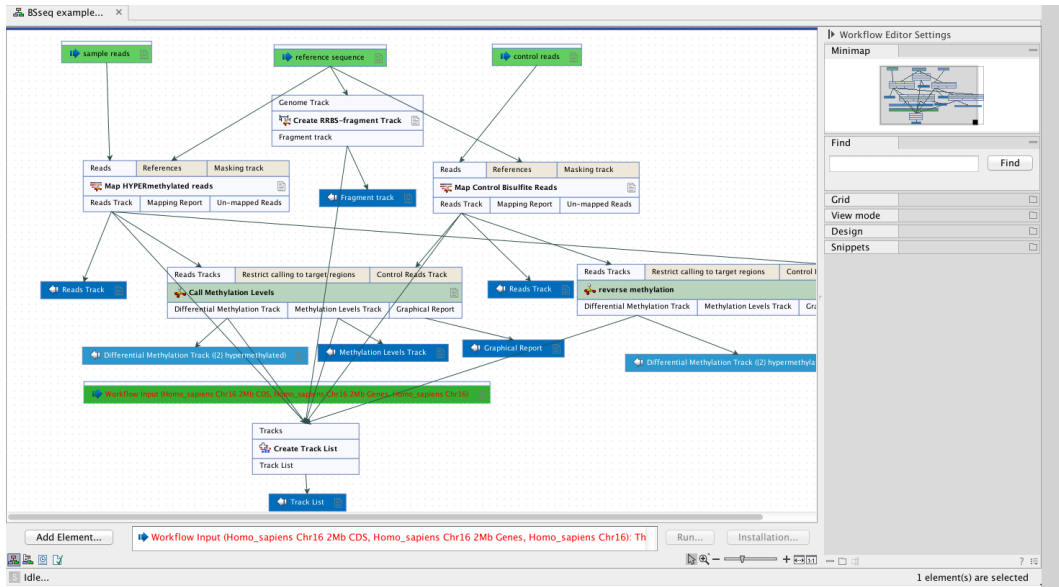


Figure 18: Complete workflow for bisulfite sequencing in the graphical editor view.

- ▼ workflow run results
  - Homo\_sapiens Chr16 sequence (MspI)
  - b-cells mapping [SRR342497] (Reads)
  - hspc mapping [SRR342518] (Reads)
  - hspc mapping [SRR342518], Homo\_sapiens Chr16 sequence, b-cells mapping [SRR342497] hypermethylated
  - b-cells mapping [SRR342497], Homo\_sapiens Chr16 sequence, hspc mapping [SRR342518] hypermethylated
  - b-cells mapping [SRR342497] (Reads) (Methylation levels)
  - hspc mapping [SRR342518] (Reads) (Methylation levels)
  - b-cells mapping [SRR342497] (Reads) (Methylation-report)
  - hspc mapping [SRR342518] (Reads) (Methylation-report)
  - Track List

Figure 19: Results generated in a workflow run.

The last one highlighted in the figure is the Track List view that aggregates all the results and reference in a way similar to the one we did manually for the figure 14.

## Bibliography

[Hodges et al., 2011] Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., et al. (2011). Directional dna methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell*, 44(1):17-28.