Tutorial

Creating a Methylation Database for Three Cell Types

June 22, 2022

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Creating a Methylation Database for Three Cell Types

This tutorial demonstrates how to build a Methylation Database that is useful for prediction of methylation profiles in mixture samples. The tutorial data set is constructed from cell line samples using the QIAseq T cell Infiltration panel (MHS-202Z).

In more detail, during this tutorial, we will:

- Test the methylation database provided as reference data on different pure cell types.
- Build databases based on pure sample methylation level tracks using the **Create Methylation Database** tool.
- Inspect reports and investigate how parameter selection affects the selected CpG sites distribution.
- Test the quality and performance on pure and mixture samples using the **Predict Methylation Profile** tool.
- Finally, we will highlight other use cases.

The aim of this tutorial is to guide you in constructing your own methylation database.

Prerequisites For this tutorial, you must have the Biomedical Genomics Analysis plugin version 22.1 or newer installed. How to install plugins is described here:

http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html.

General notes In this tutorial, we use the tool **Create Methylation Database** to build a methylation database. The tool is designed to select CpG sites that can be used to distinguish between two or three pure sample types. Please find a thorough description of the tool in the manual:

https://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsanalysis/current/index.php?manual=Create_Methylation_Database.html.

For testing and validation of the databases we use the tool **Predict Methylation Profile**. Please consult the manual for a description of the underlying algorithm, known limitations and biases:

https://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsanalysis/current/index.php?manual=Predict_Methylation_Profile.html.

Download tutorial data set Download the tutorial data and store the .zip file locally. Use the **Standard Import** 'Automatic import' option from the Import (📥) menu in the top left corner of the Workbench.

Upon import you will see a folder named Tutorial data populated with 15 Methylation Level Tracks. The names of these indicate their cell type content: Fib, Epi, and IC refer to Fibroblasts, Epithelial cells and Immune cells, respectively. Eleven samples are pure samples and numbered from 1 up to 5. They will be used for constructing databases. The remaining four samples are mixtures of different cell types that will be useful for testing performance of the databases. The numbers in the names sum to 100% and illustrate the approximate contribution of the different cell types.

Testing Methylation Database Available as Reference Data

- The QIAseq T Cell Infiltration Panel has been designed to analyze immune cell content in tissue. In the References ( ) manager we provide a database with cell types. Open the References manager by clicking the icon in the top right corner of the workbench. Now go to the **QIAGEN Sets** tab and look for QIAseq Methyl Panels hg19. This dataset includes a methyl reference database element. Select and download the element. Use the **Show in Navigation Area** button to locate the element.

- Navigate to the tutorial data folder in the navigation area. Start the **Predict Methylation Profile ( )** from the Launch ( ) menu in the top left corner. Simply start typing the name of the tool in the search bar or find it in the toolbox here:

  Toolbox | Epigenomics Analysis ( ) | Bisulfite Sequencing ( )| Predict Methylation Profile ( )

- Tick **Batch** in the lower left corner of the dialog and select Fib 1, Epi 1, IC 1 and the 4 elements with mixtures of the different cell types as shown in figure 1. Then click **Next** which will take you to the batch overview. If the seven samples are correctly selected move on to the settings step using **Next** again.

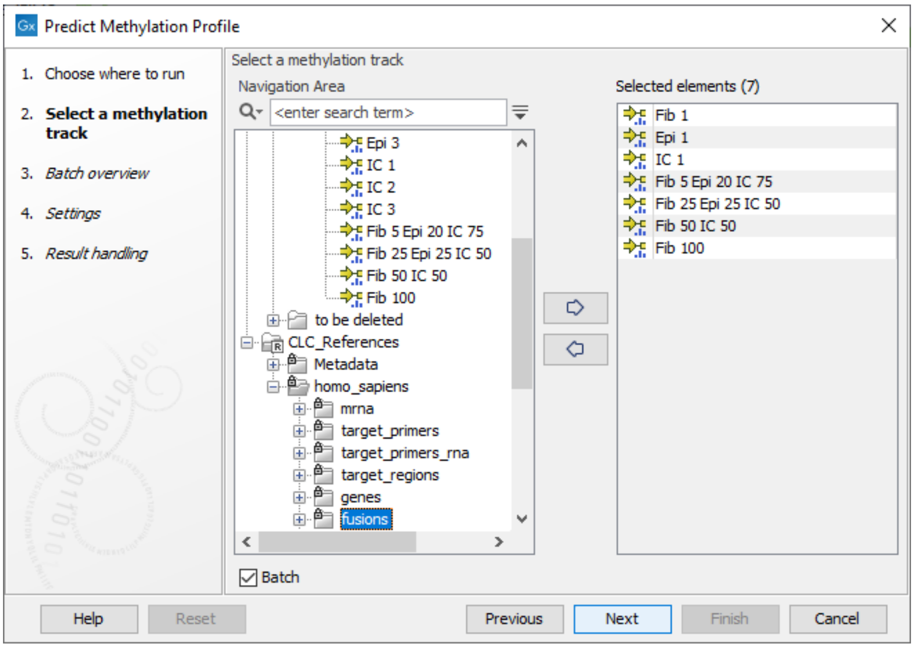
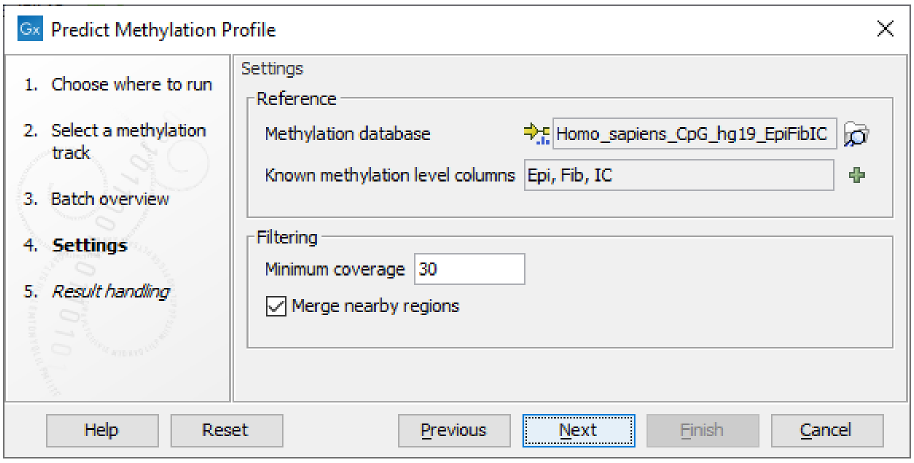


Figure 1: *Wizard step showing selection of the 7 elements that should be used in this analysis.*

- Select the methylation database you just downloaded as the reference element. Use the browse icon to locate the element among the CLC_References. Use the plus icon to set Known methylation level columns to the three cell types. Keep the default parameters otherwise, see figure 2.

- Save the outputs in a new folder called **reference data predictions** using the **Save in specified folder** option. If you tick **Create subfolder per batch unit** a folder will be created for each input sample, however it is not needed here. Finally select **Finish** to run the tool. It takes a few minutes to analyze the samples.



Settings

Reference

Methylation database: Homo_sapiens_CpG_hg19_EpiFibIC

Known methylation level columns: Epi, Fib, IC

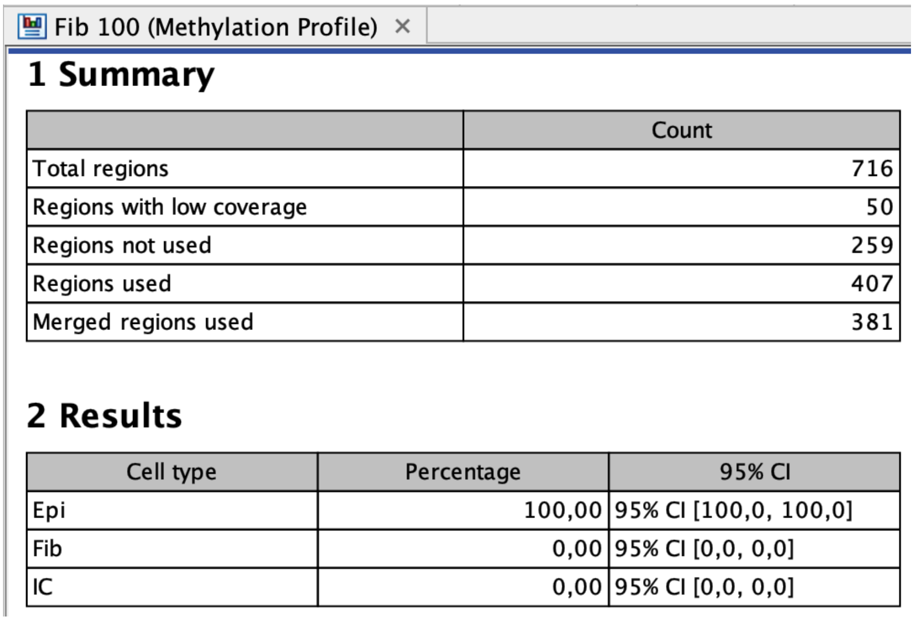
Filtering

Minimum coverage: 30

☑ Merge nearby regions

Figure 2: *Selection of the reference data element. The Known methylation level columns are filled out based on the content in the database. Leave the minimum coverage at 30 and use Merge of nearby regions as selected by default.*

- Open the generated reports and inspect the results. Note that the expected Fib contribution is not always captured correctly, see figure 3



1 Summary

	Count
Total regions	716
Regions with low coverage	50
Regions not used	259
Regions used	407
Merged regions used	381

2 Results

Cell type	Percentage	95% CI
Epi	100,00	95% CI [100,0, 100,0]
Fib	0,00	95% CI [0,0, 0,0]
IC	0,00	95% CI [0,0, 0,0]

Figure 3: *Wrong prediction of the 100% Fibroblast cell type as Epithelial cells.*

Looking at the other reports it is evident that the predictor is not handling the fibroblasts very well, although it helps a bit when all three cell types are in the mixtures. The most likely cause is that the fibroblasts in these samples have different methylation levels than those in the reference database. Hence it is necessary to build a new database that uses the fibroblasts in the samples.

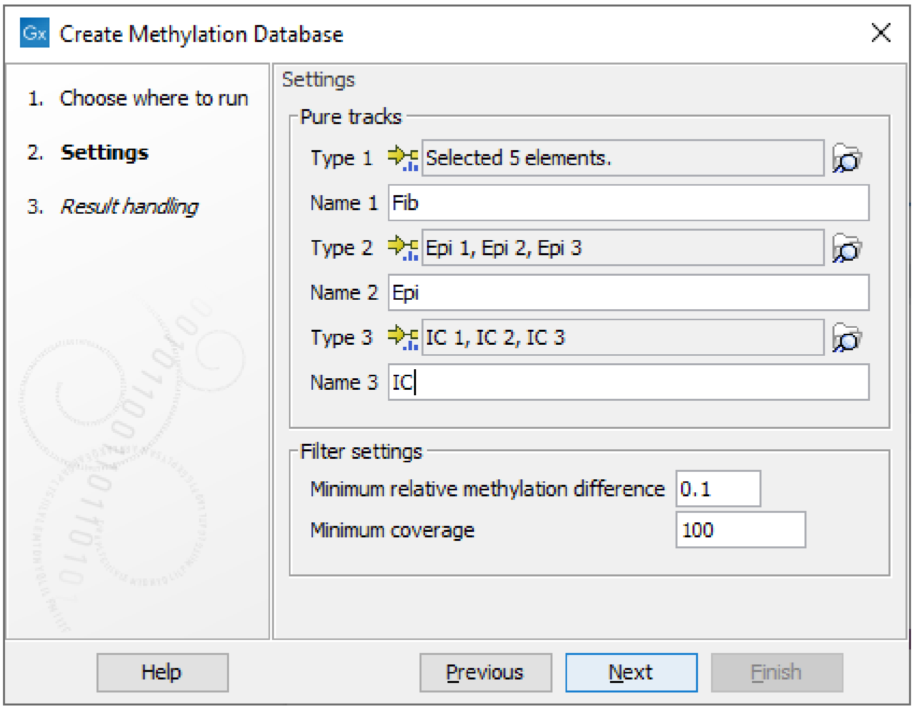
Creating a Methylation Database for Three Cell Types

Now let's build a database that will hopefully improve the predictions. We will use the eleven pure samples for this purpose.

- Launch the **Create Methylation Database** tool from the Launch menu in the top left corner. Simply start typing the name of the tool in the search bar or find it in the toolbox here:

  Toolbox | **Epigenomics Analysis** () | **Bisulfite Sequencing** ()| **Create Methylation Database** ()

- Select the methylation level tracks of the three different cell types. Use Fib, Epi and IC to name the cell types, see figure 4.



Settings

Pure tracks

Type 1: Selected 5 elements.

Name 1: Fib

Type 2: Epi 1, Epi 2, Epi 3

Name 2: Epi

Type 3: IC 1, IC 2, IC 3

Name 3: IC

Filter settings

Minimum relative methylation difference: 0.1

Minimum coverage: 100

Figure 4: *Selection of 5 Fib, 3 Epi and 3 IC samples. Filter settings are the same as the default settings.*

- Save the database in a folder called **creating databases** and press **Finish** to start the tool.
- Rename the created database to **DB 1**.
- Create two more databases by restarting **Create Methylation Database** (), one where the coverage is the same as before and one where the Minimum relative methylation level is the same as the first database created, but the other parameter varies:
  - rerun 1: Minimum relative methylation difference = 0.4 and Minimum coverage = 100 to keep the coverage the same.
  - rerun 2: Minimum relative methylation difference = 0.1 and Minimum coverage = 400 to keep the Minimum relative methylation level the same.

- When the databases are created, rename to **DB 2** and **DB 3**, respectively, and open the different database reports. As it can be seen, the report is divided into four sections:
  - **Summary** contains information on how many CpGs are filtered out and what the content is in the final database. Note here that you would need to retain a relatively high number of sites in order to have a good working database.
  - **Average Methylation Level** is a table containing information on the average methylation for each sample. Samples should have approximately the same level of methylation to produce a good database. When types have very different profiles it can be difficult to have a good balance of hyper and hypo methylated sites representing the given type.
  - **Methylation ranges for included sites** gives a good overview of the hyper and hypo contribution making up the cell type signature. Here a profile with many sites in both extremes is preferred.
  - **Relative Methylation levels in CpGs** is represented as a ternary plot when three samples are used and a histogram when two samples are used. Note how the filter parameters chosen in the different databases affect how the sites are selected, see figure 5.

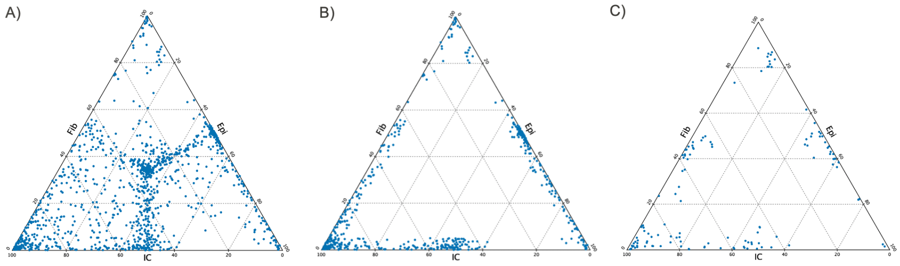


Figure 5: *Ternary plots of the three different databases created in the tutorial. A) plot showing many included sites and in many cases with little difference in methylation across types. These sites are non-informative and will not add value to the prediction. B) more diverse sites and with good coverage of the edges of the triangle. The coverage seems good and captures many CpGs that are potentially informative. C) plot showing the effect of increasing the average coverage, here fewer sites can contribute to the database profile prediction. This is not necessarily a bad thing as long as many regions are represented.*

Identifying the best database

Now let's test the databases and see which one performs best. For testing we will use the same samples as in the test of the reference data element we did previously (Fib 1, Epi 1, IC 1, Fib 100, Fib 50 IC 50, Fib 25 Epi 25 IC 50 and Fib 5 Epi 20 IC 75).

> Note that we strongly recommend not to validate using pure samples included in methylation databases. In this tutorial we make an exception strictly for illustrating purposes.

Start **Predict Methylation Profile** () and select the first database. Tick batch to run all samples at the same time. Repeat this for each database. Save the output in new folders called DB 1, DB 2 and DB 3. Remember to switch the input database between the runs.

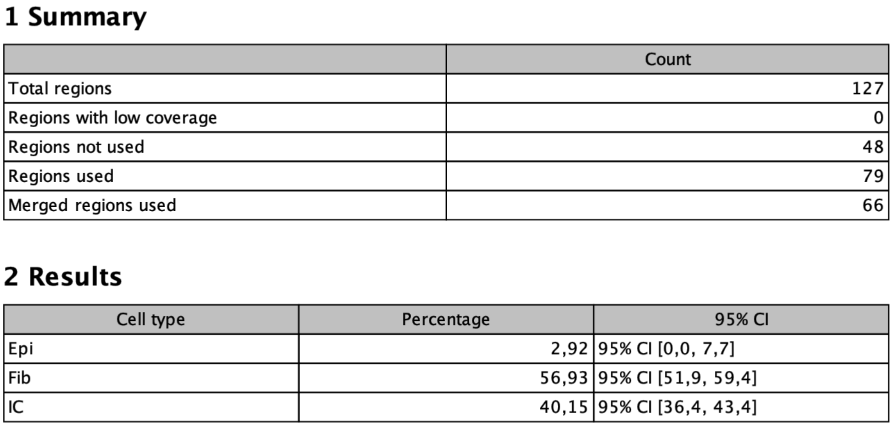
Inspect the results and compare to the expected.

A number of things can contribute to the decision when selecting a database.

1. The number of sites needed to make a proper estimation.
2. How close the estimated values are to the true values.
3. How tight the confidence interval is, and whether it captures the expected value.
4. How quickly the prediction is made.

Number of sites From the report generated when performing the prediction it is possible to see how many merged regions contribute to the prediction (information provided in Summary of the methylation profile report). As the tool does a dynamical assessment of which sites are informative for a respective sample it might differ between individual samples, however a trend can be seen (look at the methylation regions track differences between samples to inspect which sites are used for a given sample (Yes/No)). In the three cases represented here the average number of merged regions are 70, 500 and 1000 for DB 3, DB 1 and BD 2, respectively.

Only using 70 merged regions (DB 3) causes some problems with misprediction of types that are not in the mix, see figure 6. Comparing to selecting sites such that more regions are assessed seem to help the prediction, see figure 7. Note that the prediction is not exactly 50/50 Fib and IC, however the two databases agree. As the samples were mixed by adding an equal number of cells from each cell line, the NGS results might show some deviation from the expected values.



1 Summary

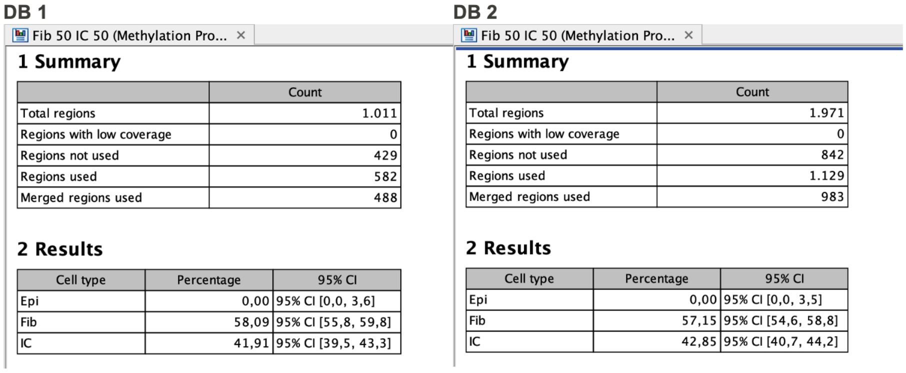
	Count
Total regions	127
Regions with low coverage	0
Regions not used	48
Regions used	79
Merged regions used	66

2 Results

Cell type	Percentage	95% CI
Epi	2,92	95% CI [0,0, 7,7]
Fib	56,93	95% CI [51,9, 59,4]
IC	40,15	95% CI [36,4, 43,4]

Figure 6: *Report of prediction using DB 3 and a mixture sample with 50% Fibroblasts (Fib) and 50% Immune cells (IC). The results indicate a small fraction of Epithelial cells (Epi), however they were not in the mix. The confidence interval also indicates that they could be absent, but it is difficult to judge.*

Estimated values Now let's look at the pure samples. In most cases the pure samples are predicted to be close to 100%, however one sample stands out, see figure 8. That the prediction on a sample that was used as input for the algo is this uncertain indicates that more sites are needed for a more precise estimate.



DB 1

Fib 50 IC 50 (Methylation Pro... ×

1 Summary

	Count
Total regions	1.011
Regions with low coverage	0
Regions not used	429
Regions used	582
Merged regions used	488

2 Results

Cell type	Percentage	95% CI
Epi	0,00	95% CI [0,0, 3,6]
Fib	58,09	95% CI [55,8, 59,8]
IC	41,91	95% CI [39,5, 43,3]

DB 2

Fib 50 IC 50 (Methylation Pro... ×

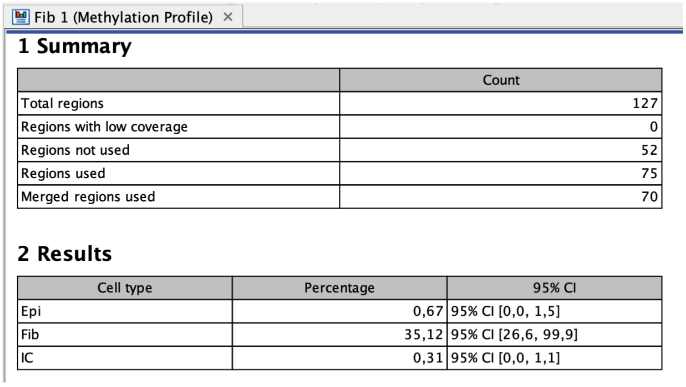
1 Summary

	Count
Total regions	1.971
Regions with low coverage	0
Regions not used	842
Regions used	1.129
Merged regions used	983

2 Results

Cell type	Percentage	95% CI
Epi	0,00	95% CI [0,0, 3,5]
Fib	57,15	95% CI [54,6, 58,8]
IC	42,85	95% CI [40,7, 44,2]

Figure 7: *Report of prediction using a mixture sample with 50% Fibroblasts (Fib) and 50% Immune cells (IC) and DB 1 and DB 2, respectively. Only minor differences can be seen. Epithelial cells (Epi) are not predicted to be in the samples which is correct and the IC contribution is higher for these two databases compared to DB 3 in figure 6.*



Fib 1 (Methylation Profile) ×

1 Summary

	Count
Total regions	127
Regions with low coverage	0
Regions not used	52
Regions used	75
Merged regions used	70

2 Results

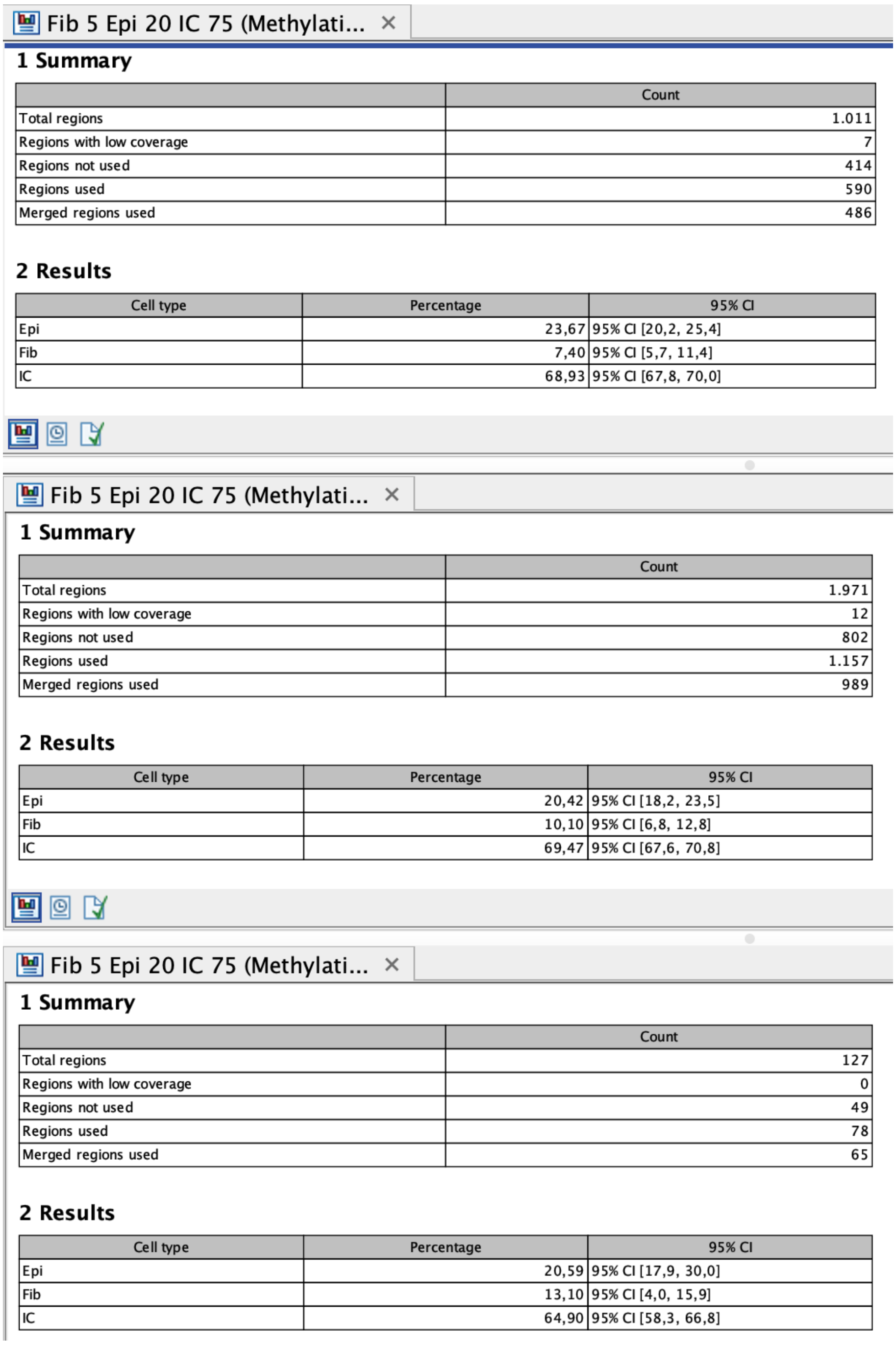
Cell type	Percentage	95% CI
Epi	0,67	95% CI [0,0, 1,5]
Fib	35,12	95% CI [26,6, 99,9]
IC	0,31	95% CI [0,0, 1,1]

Figure 8: *Report of the Fib 1 sample using the sample containing 100% Fibroblasts and DB 3. Even though the sample is predicted to mostly be Fib with little IC or Epi, it seems like the prediction is a bit uncertain and the combined predicted estimate do not sum up to 100%.*

Confidence interval Looking at the confidence intervals it is important to note that we do not guarantee the prediction algorithm will always capture the true value, however the confidence interval should be close to doing that. In perfectly designed databases the confidence interval will always capture the true value and will be relatively small.

Comparing the Fib 5 Epi 20 IC 75 mixture sample across databases reveals some benefits and disadvantages for the different databases. In figure 9, it can be seen that in all cases the IC cells are predicted to comprise less of the sample than expected. It can also be seen that the confidence intervals differ and that increasing the number of sites increases the values predicted slightly. (Note that the confidence intervals are calculated based on sampling error of sites. The fewer the number of sites, the larger the sampling error, and so the larger the intervals).

DB 1 is better at predicting the value of the Fib content, however only DB 3 captures the expected Fib value within the confidence interval, but the interval is large. The estimated value in DB 3 is twice as high as the expected value of Fib in the sample, see figure 9. The Epi value is more accurate predicted in DB 2 and 3, however the confidence interval in DB 1 is very close to capturing the true value and relatively narrow.


Fib 5 Epi 20 IC 75 (Methylati... ×

1 Summary

	Count
Total regions	1.011
Regions with low coverage	7
Regions not used	414
Regions used	590
Merged regions used	486

2 Results

Cell type	Percentage	95% CI
Epi	23,67	95% CI [20,2, 25,4]
Fib	7,40	95% CI [5,7, 11,4]
IC	68,93	95% CI [67,8, 70,0]

Fib 5 Epi 20 IC 75 (Methylati... ×

1 Summary

	Count
Total regions	1.971
Regions with low coverage	12
Regions not used	802
Regions used	1.157
Merged regions used	989

2 Results

Cell type	Percentage	95% CI
Epi	20,42	95% CI [18,2, 23,5]
Fib	10,10	95% CI [6,8, 12,8]
IC	69,47	95% CI [67,6, 70,8]

1 Summary

	Count
Total regions	127
Regions with low coverage	0
Regions not used	49
Regions used	78
Merged regions used	65

2 Results

Cell type	Percentage	95% CI
Epi	20,59	95% CI [17,9, 30,0]
Fib	13,10	95% CI [4,0, 15,9]
IC	64,90	95% CI [58,3, 66,8]

Figure 9: *The report of the Fib 5 Epi 20 IC 75 sample mix is shown for all three databases. From the top DB 1, in the middle DB 2 and at the bottom DB 3.*

In general the confidence intervals are narrower for DB 1 and DB 2 compared to DB 3, this indicates that there is a benefit to having some diversity, however at some point adding additional CpG sites does not increase the accuracy of prediction.

Speed The last thing to look at is speed. You probably noted that DB 2 took longer and DB 3 shorter to run. When building databases containing very many CpGs it is beneficial to select one that performs well and runs relatively fast as very little is gained by adding too much complexity.

Let's revisit the 4 points for selecting the database: number of sites, estimated values, confidence intervals and speed. When considering them jointly DB 1 performs best. Note that this is just a tutorial to illustrate the effect of the filter selections and that other selection might outperform what was selected here.

Creating methylation databases for other use cases

In addition to building a database from three cell types the **Create Methylation Database** tool also supports building databases from two conditions, this can be beneficial when analyzing case/control or cancer/healthy conditions. One very important thing here is that the methylation pattern needs to be well balanced so that both hyper and hypo methylation sites can be selected for predicting profiles in both conditions. The report section 3 shows the histograms of the CpG profiles of the individual conditions. Bars at both low and high methylation level should be visible. In addition, in section 2, make sure that the methylation contribution is not too different between the conditions, a little difference is however natural.

The final plot in section 4 of the report will be a histogram instead of the Ternary plot. It is important to make sure that this is also well balanced. Here is an example of an experiment that it not very well balanced, see figure 10.

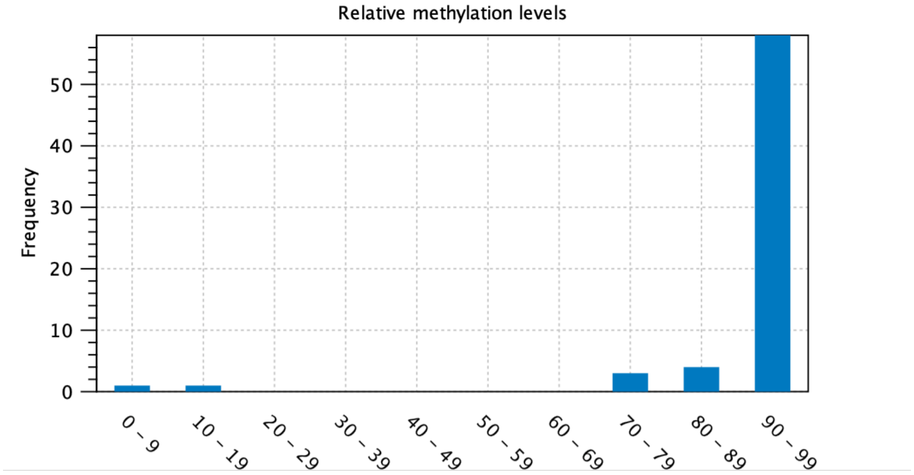


Figure 10: *Relative methylation levels plot of an experiment where an imbalance is seen.*