QC of aligned reads

These tools work on BAM files that contain read-related information (e.g. read DNA sequence, sequencing quality, mapping quality etc.). They are typically generated by read alignment programs such as bowtie2.

The following tools will allow you to inspect your BAM files more closely.

bamCorrelate

This tool is useful to assess the overall similarity of different BAM files. A typical application is to check the correlation between replicates or published data sets.

What it does

The tool splits the genomes into bins of a given length. For each bin, the number of reads found in each BAM file is counted and a correlation of the read coverages is computed for all pairs of BAM files.

Important parameters

bamCorrelate can be run in 2 modes: *bins* and *bed*.

In the *bins* mode, the correlation is computed based on **randomly sampled bins of equal length**. The user has to specify the *number* of bins. This is useful to assess the overall similarity of BAM files,
but outliers, such as heavily biased regions have the potential to skew the correlation values.

In the *BED-file option*, the user supplies a list of genomic regions in BED format in addition to (a) BAM file(s). bamCorrelate subsequently uses this list to compare the read coverages for these regions only. This can be used, for example, to compare the ChIP-seq coverages of two different samples for a set of peak regions.

In addition to specifying the regions for which the read numbers should be compared (random regions in the bins mode, selected regions in the BED-file mode), you can also specify what kind of correlation measure you would like to compute: Pearson or Spearman. In short, Pearson is an appropriate measure for data that follows a normal distribution while Spearman does not make this assumption and is generally less driven by outliers. As genome-wide sequencing data very rarely follows a normal distribution and we often encounter few regions that capture extremely high read counts (= outlier), we tend to prefer the Spearman correlation coefficient.

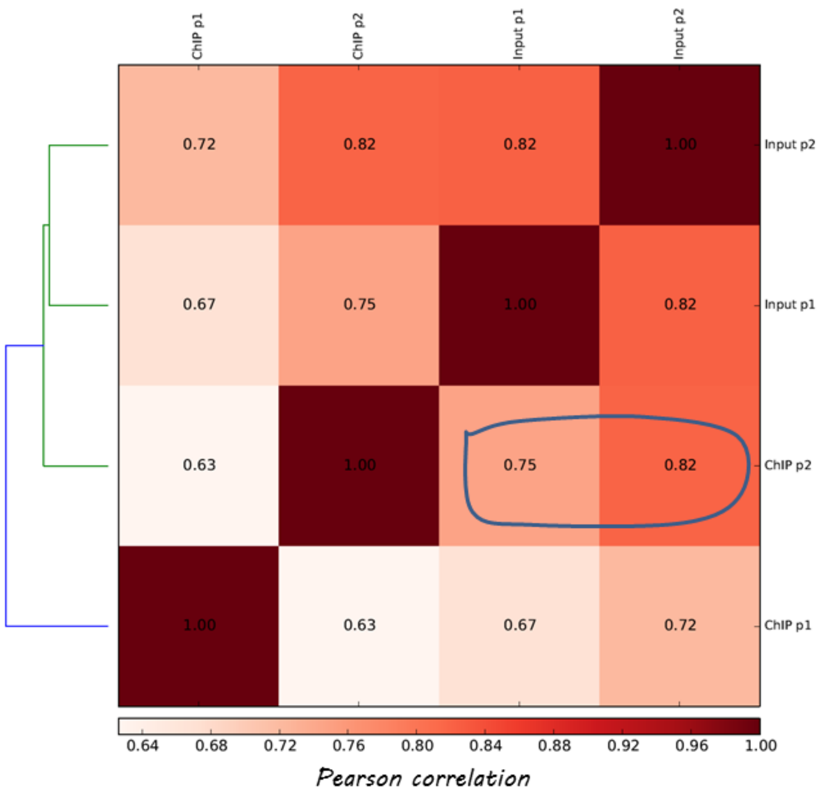
Output files:

- **diagnostic plot** the plot produced by bamCorrelate is a clustered heatmap displaying the values for each pair-wise correlation, see below for an example
- **data matrix** (optional) in case you want to plot the correlation values using a different program, e.g. R, this matrix can be used

Example Figures

Here is the result of running bamCorrelate. We supplied 4 BAM files that were generated from 2 patients - for each patient, there is an control (called "input") and a ChIP-seq sample (from the GEO sample GSE32222).

2 patient samples, same ChIP antibody



- heterogeneous tissues
- sample preparation (tissue sampling, sample handling, library prep...)
- differences between the patients' genomes
- ....

the ChIP sample from patient 2 is more similar to the input of patient 1 than to the other ChIP

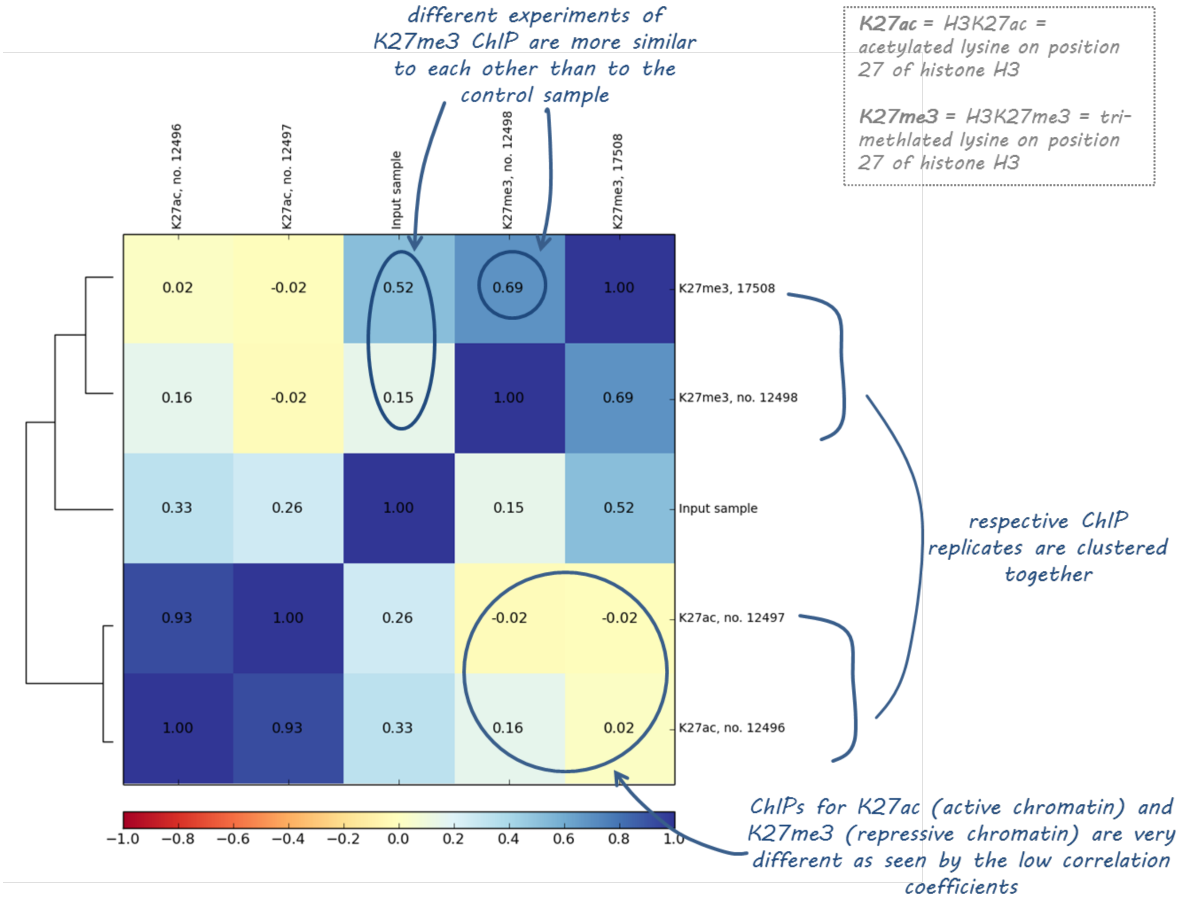
which ChIP is the odd one? or is the input from patient 1 the problem?

You can supply any number of BAM files that you would like to compare. In Galaxy, you just click "Add BAM file", in the command line you simply list all files one after the other (you can give meaningful names via the --label option).

Here's the command that was used with the command line version:

```
$ deepTools-1.5.2/bin/bamCorrelate bins --fragmentLength 200 \
--bamfiles GSM798383_SLX-1201.250.s_4.bwa.homo_sapiens_f.bam \
GSM798384_SLX-1881.334.s_1.bwa.homo_sapiens_f.bam \
GSM798406_SLX-1202.250.s_1.bwa.homo_sapiens_f.bam \
GSM798407_SLX-1880.337.s_8.bwa.homo_sapiens_f.bam \
--labels "ChIP p1" "ChIP p2" "Input p1" "Input p2" \
--plotFile bamCorrelate_result.pdf --corMethod pearson
```

Here is another example of ChIP samples for two different histone marks (the histone marks are abbreviated H3K27me3 and H3K27ac and have been shown to mark inactive and active chromatin, respectively. For our example, H3K27ac was ChIPed by the same experimentator for different cell populations while H3K27me3 was performed with the same antibody, but at different times. You can see that the correlation between the H3K27ac replicates is much higher than for the H3K27me3 samples, however, for both histone marks, the ChIP-seq experiments are more similar to each other than to the other ChIP or to the input. In fact, the signals of H3K27ac and H3K27me3 are almost not correlated at all which supports the notion that their biological function is also quite opposing.



different experiments of K27me3 ChIP are more similar to each other than to the control sample

K27ac = H3K27ac = acetylated lysine on position 27 of histone H3

K27me3 = H3K27me3 = tri-methlated lysine on position 27 of histone H3

respective ChIP replicates are clustered together

ChIPs for K27ac (active chromatin) and K27me3 (repressive chromatin) are very different as seen by the low correlation coefficients

computeGCbias

This tool computes the GC bias using the method proposed by Benjamini and Speed.

What it does

The basic assumption of the GC bias diagnosis is that an ideal sample should show a uniform distribution of sequenced reads across the genome, i.e. all regions of the genome should have similar numbers of reads, regardless of their base-pair composition. In reality, the DNA polymerases used for PCR-based amplifications during the library preparation of the sequencing protocols prefer GC-rich regions. This will influence the outcome of the sequencing as there will be more reads for GC-rich regions just because of the DNA polymerase's preference.

computeGCbias will **first calculate the *expected* GC profile** by counting the number of DNA fragments of a fixed size per GC fraction (GC fraction is defined as the number of G's or C's in a genome region of a given length)(a). This profile is then **compared to the *observed* GC profile** by counting the number of sequenced reads per GC fraction.

a) The expected GC profile depends on the reference genome as different organisms have very different GC contents. For example, one would expect more fragments with GC fractions between 30% to 60% in mouse samples (GC content of the mouse genome: 45 %) than for genome fragments from Plasmodium falciparum (genome GC content P. falciparum: 20%).

Output files

- **Diagnostic plot**
  - box plot of *absolute* read numbers per genomic GC fraction
  - x-y plot of *observed/expected* read ratios per genomic GC fraction (ideally, ratio should always be 1 (log2(1) = 0))
- **Data matrix**
  - tabular matrix file
  - to be used for GC correction with *correctGCbias*

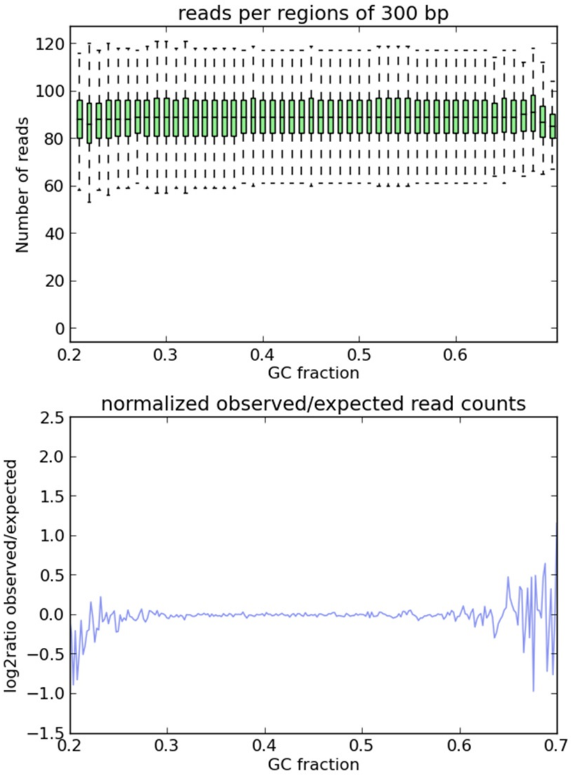
What the plots tell you

In an ideal sample without GC bias, the ratio of observed/expected values should be close to 1 for all GC content bins.

However, due to PCR (over)amplifications, the majority of ChIP samples usually shows a significant bias towards reads with high GC content (>50%) and a depletion of reads from GC-poor regions.

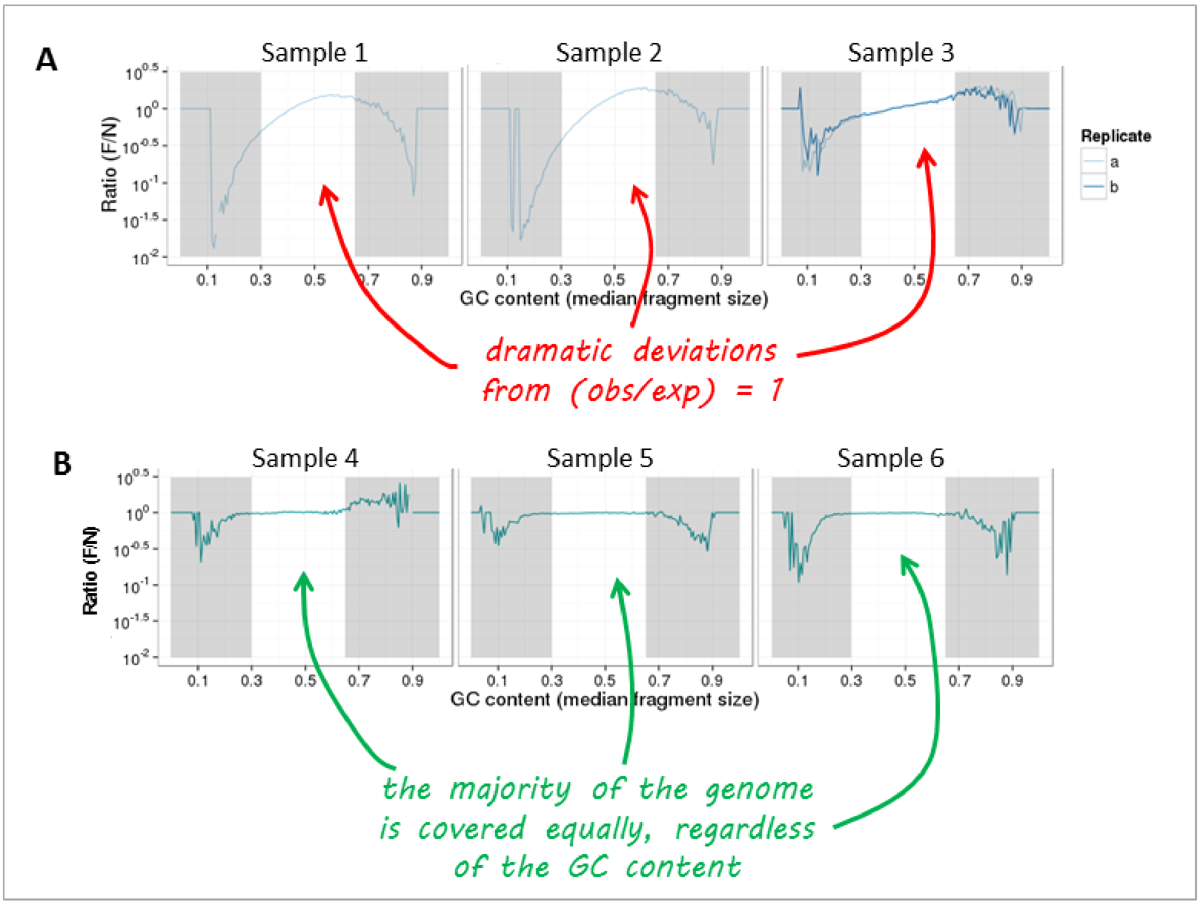
Example figures

Let's start with an ideal case. The following plots were generated with computeGCbias using simulated reads from the *Drosophila* genome.



As you can see, both plots based on **simulated reads** do not show enrichments or depletions for specific GC content bins, there is an almost flat line at log2ratio of 0 (= ratio of 1). The fluctuations on the ends of the x axis are due to the fact that only very, very few regions in the genome have such extreme GC fractions so that the number of fragments that are picked up in the random sampling can vary.

Now, let's have a look at **real-life data** from genomic DNA sequencing. Panels A and B can be clearly distinguished and the major change that took place between the experiments underlying the plots was that the samples in panel A were prepared with too many PCR cycles and a standard polymerase whereas the samples of panel B were subjected to very few rounds of amplification using a high fidelity DNA polymerase.



bamFingerprint

This quality control will most likely be of interest for you if you are dealing with ChIP-seq samples as a pressing question in ChIP-seq experiments is "Did my ChIP work?", i.e. did the antibody-treatment enrich sufficiently so that the ChIP signal can be separated from the background signal? (After all, around 90 % of all DNA fragments in a ChIP experiment will represent the genomic background). We use bamFingerprint routinely to monitor the outcome of ChIP-seq experiments.

What it does

This tool is based on a method developed by Diaz et al. and it determines how well the signal in the ChIP-seq sample can be differentiated from the background distribution of reads in the control sample. For factors that will enrich well-defined, rather narrow regions (e.g. transcription factors such as p300), the resulting plot can be used to assess the strength of a ChIP, but the broader the enrichments are to be expected, the less clear the plot will be. Vice versa, if you do not know what kind of signal to expect, the bamFingerprint plot will give you a straight-forward indication of how careful you will have to be during your downstream analyses to separate biological noise from meaningful signal.

The tool first samples indexed BAM files and counts all reads overlapping a window (bin) of specified length. These counts are then sorted according to their rank and the cumulative sum of read counts is plotted.

Output files:

- **Diagnostic plot**
- **Data matrix** of raw counts (optional)

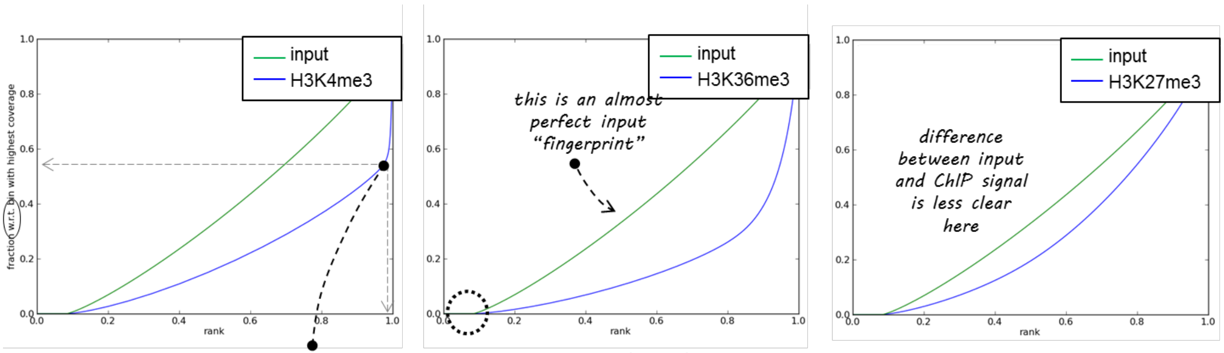
What the plots tell you

An ideal input with perfect uniform distribution of reads along the genome (i.e. without enrichments in open chromatin etc.) should generate a straight diagonal line. A very specific and strong ChIP enrichment will be indicated by a prominent and steep rise of the cumulative sum towards the highest rank. This means that a big chunk of reads from the ChIP sample is located in few bins which corresponds to high, narrow enrichments seen for transcription factors.

Example figures

Here you see 3 different fingerprint plots.

We chose these examples to show you how the nature of the ChIP signal (narrow and high vs. wide and not extremely high) is reflected in the "fingerprint" plots. Please note that these plots go by the name of "fingerprints" in our facility because we feel that they help us tremendously in judging individual files, but the idea underlying these plots came from Diaz et al.



when counting the reads contained in 97% of all genomic bins, only 55% of the maximum number of reads are reached, i.e. 4% of the genome contain a very large fraction of reads!

this indicates very localized, very strong enrichments!
(as every biologist hopes for in a ChIP for H3K4me3)

pay attention to where the curves start to rise - this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads)

H3K27me3 is a mark that yields broad domains instead of narrow peaks

it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed