DeepTools: user-friendly tools for the normalization and visualization of deep-sequencing data

Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn Grüning

September 26, 2013

Contents

1 Tools for quality control

1.1 bamFingerprint

This tool is based on a method developed by Diaz et al. (2012). Stat Appl Genet Mol Biol 11(3). The resulting plot can be used to assess the strength of a ChIP (for factors that bind to narrow regions). 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 are plotted. 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.

1.2 computeGCBias

This tool computes the GC bias using the method proposed by Benjamini and Speed (2012). Nucleic Acids Res. (see below for more explanations) The output is used to plot the bias and can also be used later on to correct the bias with the tool correctGCbias. There are two plots produced by the tool: a boxplot showing the absolute read numbers per genomic-GC bin and an x-y plot depicting the ratio of observed/expected reads per genomic GC content bin.

Summary of the method used

In order to estimate how many reads with what kind of GC content one should have sequenced, we first need to determine how many regions the specific reference genome contains for each amount of GC content, i.e. how many regions in the genome have 50% GC (or 10% GC or 90% GC or...). We then sample a large number of equally sized genome bins and count how many times we see a bin with 50% GC (or 10% GC or 90% or...). These EXPECTED values are independent of any sequencing as it only depends on the respective reference genome (i.e. it will most likely vary between mouse and fruit fly due to their genome's different GC contents). The OBSERVED values are based on the reads from the sequenced sample. Instead of noting how many genomic regions there are per GC content, we now count the reads per GC content. In an ideal sample without GC bias, the ratio of OBSERVED/EXPECTED values should be close to 1 regardless of the GC content. Due to PCR (over)amplifications, the majority of ChIP samples usually shows a significant bias towards reads with high GC content (¿50%)

2 Tools for normalization

3 correctGCBias

This tool requires the output from computeGCBias to correct the given BAM files according to the method proposed by Benjamini and Speed (2012). Nucleic Acids Res. The resulting BAM files can be used in any downstream analyses, but be aware that you should not filter out duplicates from here on.

3.1 bigwigCompare

This tool compares two bigwig files based on the number of mapped reads. To compare the bigwig files the genome is partitioned into bins of equal size, then the number of reads found in each BAM file are counted for such bins and finally a summarizing value is reported. This value can be the ratio of the number of reads per bin, the log2 of the ratio, the sum or the difference.

3.2 bamCoverage

Given a BAM file, this tool generates a bigWig or bedGraph file of fragment or read coverages. The way the method works is by first calculating all the number of reads (either extended to match the fragment length or not) that overlap each bin in the genome. Bins with zero counts are skipped, i.e. not added to the output file. The resulting read counts can be normalized using either a given scaling factor, the RPKM formula or to get a 1x depth of coverage (RPGC).

3.3 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.

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

3.4 bamCompare

This tool compares two BAM files based on the number of mapped reads. To compare the BAM files, the genome is partitioned into bins of equal size, the reads are counted for each bin and each BAM file and finally, a summarizing value is reported. This value can be the ratio of the number of reads per bin, the log2 of the ratio or the difference. This tool can normalize the number of reads on each BAM file using the SES method proposed by Diaz et al. (2012). Stat Appl Genet Mol Biol 11(3). Normalization based on read counts is also available. The output is either a bedGraph or a bigWig file containing the bin location and the resulting comparison values. If paired-end reads are present, the fragment length reported in the BAM file is used by default.

4 Tools for visualization

4.1 computeMatrix

This tool summarizes and prepares an intermediary file containing scores associated with genomic regions that can be used afterwards to plot a heatmap or a profile. Typically, these genomic regions are genes, but any other regions defined in a BED or INTERVAL format can be used. This tool can also be used to filter and sort regions according to their score.

4.2 Profiler

This tool creates a profile plot for a score associated to genomic regions. Typically, these regions are genes, but any other regions defined in a BED or INTERVAL format will work. A preprocessed matrix generated by the tool computeMatrix is required.

4.3 Heatmapper

The heatmapper visualizes scores associated with genomic regions, for example ChIP enrichment values around the TSS of genes. Those values can be visualized individually along each of the regions provided by the user in INTERVAL or BED format. In addition to the heatmap, an average profile plot is plotted on top of the heatmap (can be turned off by the user; it can also be generated separately by the tool profiler). We implemented vast optional parameters and we encourage you to play around with the min/max values displayed in the heatmap as well as with the different coloring options. If you would like to plot heatmaps for different groups of genomic regions individually, e.g. one plot per chromosome, simply supply each group as an individual BED file.

5 Credits

If you would like to give us feedback or you run into any trouble, please send an email to deeptools@googlegroups.com

This tool is developed by the `BioinformaticsandDeep-SequencingUnit`http://www3.ie-freiburg.mpg.de/facilities/research-facilities/bioinformatics-and-deep-sequencing-unit/ at the `MaxPlanckInstituteforImmunobiologyandEpigenetics`http://www3.ie-freiburg.mpg.de.