htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R

Evarist Planet 1, Camille Stephan-Otto Attolini 1, Oscar Reina 1, Oscar Flores 2 and David Rossell 1

1Biostatistics and Bioinformatics Unit, Institute for Research in Biomedicine of Barcelona, Spain
2IRB-BSC Joint Research Program on Computational Biology, IRB Barcelona, Spain

Associate Editor: Prof. John Quackenbush

ABSTRACT

Summary: We provide a Bioconductor package with quality assessment, processing and visualization tools for high-throughput sequencing data, with emphasis in ChIP-seq and RNA-seq studies. It includes detection of outliers and biases, inefficient immuno-precipitation and overamplification artifacts, de novo identification of read-rich genomic regions and visualization of the location and coverage of genomic region lists.

Availability: www.bioconductor.org

Contact: david.rossell@irbbarcelona.org

Supplementary information: Supplementary data available at Bioinformatics online.

While analysis strategies for high-throughput sequencing data are proliferating, there remains a need for quality assessment, data processing and visualization methods. We provide tools to detect the presence of outliers and biases, inefficient immuno-precipitation (IP), over-amplification and strand-specific biases. We implement strategies to adjust for these biases. Also, we provide routines for quick data formatting, analysis and visualization. htSeqTools is integrated in Bioconductor (Gentleman et al, 2004), an environment offering a wide variety of analysis strategies. We take advantage of parallel computation and operations implemented in other packages to deliver computationally efficient solutions. htSeqTools can be a valuable complement for pipelines and advanced analysis strategies.

Below we show the main software features and use a Saccharomyces cerevisiae (GSE16926) and a human (GSE25836) ChIP-seq experiment as examples (www.ncbi.nlm.nih.gov/geo).

The supplementary material (SM) contains a more detailed description, comparisons with existing approaches, typical workflows, as well as 2 ChIP-seq and 1 RNA-seq additional examples.

1 QUALITY CONTROL

- Visualize sample correlations. Principal Component Analysis (PCA) is useful to assess quality and identify problematic samples. Unfortunately, it is not directly applicable to sequencing data.

Instead, we measure the distance in read coverage between samples i and j as $d_{ij} = 0.5(1 - \rho_{ij})$ where $\rho_{ij}$ is the Pearson, Spearman or Kendall correlation between their log(coverage+1). The log-scale reduces the influence of extremely high-coverage regions. We display $d_{ij}$ in 2-3 dimensions via Multi-Dimensional Scaling (MDS), so that Euclidean distances between points approximate $d_{ij}$. Figure 1 shows an MDS plot for the ChIP-seq experiment GSE25836. The distance between FOXA2 IP samples and their inputs is larger than the distance between replicates, indicating a satisfactory quality.

- Remove over-amplification artifacts. PCR over-amplification causes some reads to repeat an abnormally large number of times, which can induce biases in downstream analyses. Simultaneously, naturally occurring read repeats are expected. For instance, short genomes or IP samples typically show more read repeats than
longer genomes or control samples, as they focus on smaller genomic regions. We model the number of reads as a mixture of truncated negative binomial distributions (number of components set to minimize the BIC, Schwarz, 1978), and use an empirical Bayes approach akin to Efron et al., 2001 to estimate the FDR. We fit the model after truncating 0.001 of the reads (by default) with highest number of repeats, as these are more likely to be artifacts. Figure 2 shows more read repeats in the S. cerevisiae than in the human data (solid lines). Reads with more than 6 repeats were flagged as over-amplification artifacts in the human data at a 0.01 FDR, while for S. cerevisiae the cutoff was 138 repeats. The procedure adapts the cutoff to the nature of the data.

### Assess enrichment efficiency

In sequencing experiments such as ChIP-seq, MeDIP or DNase-seq, certain samples accumulate more reads in specific regions than their controls. A lack of such enrichment with the standard deviation (SD) and Gini’s coefficient G makes it possible to detect over-amplification artifacts. Figure 2 shows more read repeats in the S. cerevisiae than in the human data (solid lines). Reads with more than 6 repeats were flagged as over-amplification artifacts in the human data at a 0.01 FDR, while for S. cerevisiae the cutoff was 138 repeats. The procedure adapts the cutoff to the nature of the data.

#### Correct strand bias

ChIP-seq fragment sizes cause reads on the +/- strands to be shifted with respect to each other. With single-end reads this poses a challenge, as the fragment size distribution is unknown. Akin to Zhang et al., 2008, we scan for reads in high coverage regions and estimate the shift $\delta$ as the mean distance between reads in the + and − strands. We add/subtract 0.5$\delta$ to the location of reads on the +/- strand, respectively.

### Table 1. Coverage $SD_n$ and $G_n$ for dataset GSE25836

<table>
<thead>
<tr>
<th>ID</th>
<th>$SD_n$</th>
<th>$G_n$</th>
<th>Antibody</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM634613</td>
<td>0.115</td>
<td>0.0060</td>
<td>FOXA</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634615</td>
<td>0.114</td>
<td>0.0059</td>
<td>FOXA</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634617</td>
<td>0.110</td>
<td>0.0017</td>
<td>Input</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634619</td>
<td>0.106</td>
<td>0.0017</td>
<td>Input</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634614</td>
<td>0.124</td>
<td>0.0142</td>
<td>FOXA</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634616</td>
<td>0.135</td>
<td>0.0211</td>
<td>FOXA</td>
<td>GAI</td>
</tr>
<tr>
<td>GSM634618</td>
<td>0.120</td>
<td>0.0022</td>
<td>Input</td>
<td>GAI</td>
</tr>
</tbody>
</table>

Although many analysis strategies are available, the computational burden of applying them to the whole genome is often excessive. It is therefore convenient to pre-screen and focus the analysis. We implement a screening tool to detect all genomic regions with coverage above a user-specified threshold, count the number of reads in each region, and optionally compare the number of reads across samples via likelihood ratio or permutation chi-square tests. We also allow for refined peak calling within the selected regions.

#### Visualize hits

We facilitate the visualization of a list of genomic regions by plotting the distribution of their distances to the closest gene/feature (in bp or relative to the feature length) and average coverage profiles. Often it is useful to scan the genome for regions accumulating a large number of hits, e.g. peaks in ChIP-seq or differential expression in RNA-seq may reveal common regulatory mechanisms. We provide functions to detect and plot such areas.

### ACKNOWLEDGMENTS

Instituto de Salud Carlos III (INB-Genome España and COMBIOMED RETICS) funds OF.

### REFERENCES


