Sequence analysis

htSeqTools: high-throughput sequencing quality control, processing and visualization in R

Evarist Planet1, Camille Stephan-Otto Attolini1, Oscar Reina1, Oscar Flores2 and David Rossell1,∗

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

Received on July 13, 2011; revised on December 15, 2011; accepted 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, inefficient immuno-precipitation artifacts, de novo identification of read-rich genomic regions and visualization of the location and coverage of genomic region lists.

Supplementary information: Supplementary data available at Bioinformatics online.

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

$$d_{ij} = 0.5(1 - ρ_{ij})$$

where $ρ_{ij}$ 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 overamplification artifacts: PCR overamplification 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 repeats 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 False Discovery Rate (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 six repeats were flagged as overamplification 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 coverage variability can indicate sample preparation problems (e.g. inefficient IP) or a lack of pronounced peaks. We measure coverage inequality with the standard deviation (SD) and Gini’s coefficient G (Gini, 1912), a classical econometrics measure of wealth inequality. The expected value of the coverage SD is proportional to $\sqrt{n}$ (see Supplementary Material), where n is the number of reads. The expected E(G(n)) also depends on n, but no closed-form expression is available. We estimate E(G(n)) by generating n reads uniformly distributed along the genome. In order to make samples with different n comparable, we report $SD_n = SD/\sqrt{n}$ and $G_n = G – E(G(n)). Table 1 shows higher SDn and Gn in the IP samples than in their respective controls, suggesting no sample preparation problems. Samples sequenced with GAI present clearer peaks than GAI samples, thus indicating an improvement in the technology.
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 \( \hat{s} \) as the mean distance between reads in the + and − strands. We add/subtract 0.5\( \hat{s} \) to the location of reads on the ± strand, respectively.

2 ANALYSIS

Find read-rich regions: in many sequencing experiments, the goal is to identify de novo genomic regions of interest, e.g. binding sites, previously unannotated short RNAs or copy number variations. Although many analysis strategies are available, the computational burden of applying them to the whole genome is often excessive. It is therefore convenient to prescreen 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 base pair 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.


Conflict of interest: none declared.

REFERENCES


