RNA-seq data normalziation

Mikhail Dozmorov

Spring 2018

Proper normalization is important

- Transcripts with different lengths within a sample are NOT comparable
- Transcripts with the same length between samples are NOT comparable
- A combination of the two is even worse

Data generative process for one sample

The read counts from RNA-seq follow a sampling process. For gene i, $i = 1, ..., G$, let

- μ - the true expression (number of cDNA fragments)
- L_i - gene length
- The probability of a read starting from gene i is: $p_i = \mu_i * L_i / \sum_{i=1}^{G} \mu_i * L_i$
- If the total number of reads is N, the count for gene i, denoted by Y_i, can be modeled as a Poisson random variable. Let $\lambda_i = N * p_i$, $Y_i|\lambda_i \sim Poisson(\lambda_i)$
- Downstream DE test between sample 1 and 2 is: $H_0 : \mu_{1i} = \mu_{2i}$ which is NOT equivalent to $H_0 : \lambda_{1i} = \lambda_{2i}$ without proper normalization

Concerns in RNA-seq data normalization

- When comparing two samples, if the distributions of p_i are approximately the same, normalizing by N will be sufficient – this is what RPKM does.
- However if that's not true we will be in trouble.
  - A toy example: if there are only two genes in the genome, their read counts are 10 and 20 in one sample, and 10 and 100 in another one. We don't know how to compare!
  - A real example: RNA-seq is often used to compare one tissue type to another, e.g., brain vs. liver. Many genes may be liver-specific and not transcribed in brain, causing difference in **library composition**.
- The normalization procedure is to choose a proper "baseline" for different samples, then normalize data to the baseline so that the counts are comparable.

Single factor normalization methods – One normalization factor per sample

- Total or median counts, aka scaling to the library size
- Intuition - it is expected that sequencing a sample to half the depth will give, on average, half the number of reads mapping to each gene
- Problems - does not take into account the composition of RNA population being sequenced

Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." Genome Biology 11, no. 10 (2010): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Full-quantile normalization

- The quantiles of the distributions of the gene-level read counts are matched across samples
  - Use counts from cousekeeping genes
  - Use a certain quantile (75th) for all counts
- Implemented in `EDASeq::betweenLaneNormalization`

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. "Evaluation of Statistical Methods for Normalization and Differential Expression in MRNA-Seq Experiments." BMC Bioinformatics 11, no. 1 (2010): 94. https://doi.org/10.1186/1471-2105-11-94.

TMM: Trimmed mean of M values

- Compute M(*log fold changes*) and A(*log total counts*) for all genes
- Discard genes with extreme M and A values (30% and 5%), and compute a weighted mean of M's for the rest of genes. The weights as the inverse of the approximate asymptotic variances
- Underlying assumption is that most genes are not DE
- Implemented in `edgeR::calcNormFactors`

Robinson, Mark D., and Alicia Oshlack. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." Genome Biology 11, no. 3 (2010): R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Gene-specific normalization – each gene has a different normalization factor

- The gene-specific biases (from GC content, gene length, etc.) need to be considered. Model the observed counts $Y_{g,i}$ for gene g in sample i

$$Y_{g,i}|\mu_{g,i} \sim Poisson(\mu_{g,i})$$

$$\mu_{g,i} = exp h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) + log(m_i)$$

- $h_i(\theta_{g,i})$ - function that captures non-linearity of counts distribution across samples (technical variability)
- $f_{i,j}(X_{g,j})$ - sample-dependent biases, e.g., GC content
- m_i - sequencing depth
- Estimate h and f and θ using conditional quantile normalization

Hansen et al. 2012 (Biostatistics) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3297825/

voom - making RNA-seq counts look like microarray measures

- log-counts per million - capture relative changes in expression
- Model the coefficient of variation (CV) of RNA-seq counts as a decreasing function of count size.

$$CV^2 = 1/\lambda + \phi$$

- λ - the expected size of the count
- ϕ - biological variation
- Captures mean-variance trend for lower counts
- Used as weights in `limma` model

Law et.al. 2014 GB https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29

Summary

- RNA-seq normalization is difficult!
- Still an open statistical problem.
- The goal is to find a proper “baseline” to normalize data to.
- Single factor methods provide comparable results.
- Gene-specific normalization is promising, but be careful of over-fitting.