GC-Content Normalization for RNA-Seq Data (Risso, Schwartz, Sherlock, and Dudoit, BMC Bioinformatics 2011)

GC-content normalization: an introduction

* Raw counts are not directly comparable 1) between genes within a lane or 2) between lanes assaying the same library (called “replicate lanes”)
* Read count proportional to gene’s length (longer genes will be more likely to have more reads) and its transcript abundance. Read counts can vary between replicate lanes due to the total number of reads produced in a given lane (differences in sequencing depth). Read counts can depend on GC content and mappability (GC-rich and GC-poor fragments underrepresented in RNA-seq)
* GC content effects tend to be lane-specific (different libraries?)- WHY?(library preparation? differences in sequencing depth?) IMPACT WHEN YOU HAVE MASTER-MIXES. WHAT ABOUT DIFFERENT FLOW CELLS? DIFFERENT ALL SMALL RELATIVE TO DIFFERENCES BETWEEN SPECIES?

- Good news for our data: same library preparation date

Within-lane normalization- want to account for the dependence of read counts on GC content

* The longer the gene, the higher the read count for a given expression level (more significant DE statistics for longer genes even when considering per-base read counts)
* Benjamini and Speed- read counts are most affected by the content for the actual DNA fragments from the sequence library (rather than the sequenced reads themselves
* Hansen et al.- base-specific read biases associated with random hexamer priming in Illumina’s standard library preparation protocol
* Pickrell et al. – sample-specificity of the GC-content effect. Dependence of gene counts on GC-content may vary between lanes (belief beforehand: for a given gene, the GC-content effect was the same across samples and hence would cancel out when considering DE statistics)

Solutions:

* Regression normalization: Log(gene-level read counts) (Yj) are regressed on GC-content (Xj) using the loess robust local regression method and normalized expression measures are obtained by shifting the residuals to recover the scale of the raw counts:

Y’ = yj – y j-hat + T summary statistic (median)

* Global-scaling normalization: Genes are stratified into K equally-sized bins based on GC-content and the normalized expression measures are defined as

Y’ = yj – T (yj: j’ set k(j)) + T (y1,…yj)

Ex: on the original (unlogged) scale, the normalized count for a particular gene could be its raw count divided by the ratio of the median count in its GC-bin to the overall median count of all genes

* Full-quantile normalization [Figure 3 of paper suggests this decreases dependency the most]: genes are binned by GC-content (same as global-scaling normalization). The quantiles of the read count distributions are then matched between GC-bins, by sorting counts within bins and then taking the median of quantiles across bins

Between-lane normalization- trying to correct for sequencing depth differences

* Apply a between-lane normalization procedure immediately after the within lane normalization. This means correct for GC content and then for library size (e.g. TMM)
* Inherently make count distributions more similar between lanes (risk of “dampening” true differential expression

Measuring effectiveness of decreasing the dependency between the between GC-content and number of read counts

Theta (parameter that we are interested in estimating)- expression fold-change estimatation

Bias of an estimator- difference between an estimator’s expectation and the true value of the parameter being estimated

MSE- the average of the squares of the errors. MSE = Var(theta hat) + (Bias(theta hat, theta) )^2

* FILTER FIRST? Paper notes that the GC-content effect is unimodal 🡪 our data does not reflect this. Impacted by the fact that Julien did the filtering out of lowly expressed genes after the GC content normalization step?
* Paper says the GC-content bias is even more of an issue when comparing read counts between species
* REMOVING H1H. Consistently clusters with human livers
* Motivating examples and what is time series in general
* Considerations for time series data (autocorrelation) and importance of modeling variances accurately
* Covariance matrices
* Modeling variances: Fixed versus random effects, mixed models
* Assessing model fit
* Implimentation of mixed models in limma
* Grouping genes based on similar expression patterns (Cormotif)