DE for sequence data

Statistics for Genomic Data Analysis

Sequence data

- Last time, we saw that seqence data are *counts*
- DNA sample $\Longrightarrow$ *population of cDNA fragments*
- Each genomic feature $\Longrightarrow$ species for which the population size is to be estimated
- Sequencing a DNA sample $\Longrightarrow$ random sampling of each of these species
- *Aim :* to estimate the relative abundance of each species in the population

Poisson model

- If we assume :
  - each cDNA fragment has the *same chance* of being selected for sequencing
  - the fragments are selected independently
- Then : the number of read counts for a given genomic feature should follow a Poisson variation law across repeated sequence runs of the same cDNA sample
- The Poisson model implies that the mean equals the variance
- This relationship has been validated in an early RNA-Seq study using the same initial source of RNA distributed across multiple lanes of an Illumina GA sequencer

Single gene model

- DNA sample $\Longrightarrow$ 'library'
- Contains genes $1, \ldots, g, \ldots$
- For a given gene g in library i, $Y_{gi} =$ number of reads for gene g in library i
- $Y_{gi} \sim Bin(M, p_{gi})$, where p_{gi} is the proportion of the total number of sequences M in library i that are gene g
- M large, p_{gi} small $\Longrightarrow$ $Y_{gi} \sim Pois(\mu_{gi} = Mp_{gi})$ (approximately)

Technical vs. biological replicates

- For the Poisson model, the *variance* is equal to the *mean*
- With *technical replicates*, this relation holds fairly well
- With *biological replicates*, the variance is typically *larger* than expected using the Poisson model
- Last time, we looked at the *Negative Binomial* model as an extension to the Poisson model that allows for this extra-Poisson variability :

$$Y_{gi} \sim NegBin(\mu_{gi} = Mp_{gi},\ \phi_g)$$

- $Var(Y_{gi}) = \mu_{gi} + \phi_g\, \mu_{gi}$
- The (square of the) *coefficient of variation* is

$$CV^2(y_{gi}) = \frac{1}{\mu_{gi}} + \phi_g$$

DE with sequence data

- Many methods for identifying differential expression (DE) have been developed for microarrays
- (for example, the method we have used with `limma`
- $\Longrightarrow$ *could we use for sequence data ? ?*
- Problematic : data from microarrays (transformed fluorescence intensities) are *continuous*
- Possibilities for analysis :
    - *transform* data and use microarray methods
    - analyze data using models for counts

t-test for DE

- In the case of microarrays, we considered different possibilities for identifying DE genes
- Single gene models, contrasts k
  - $M =$ log fold change $\Longrightarrow$ does not take variability into account
  - ordinary $t = \dfrac{\hat{\beta}_g k}{s_g\ c}$ $\Longrightarrow$ can get artificially small s_g due to small df
  - common variance $t = \dfrac{\hat{\beta}_g k}{s_0\ c}$ $\Longrightarrow$ but not all genes have the same variance
  - moderated $t = \frac{\hat{\beta}_g k}{\tilde{s}_g\ u_{gk}}$ $\Longrightarrow$ ‘borrows information’ across genes

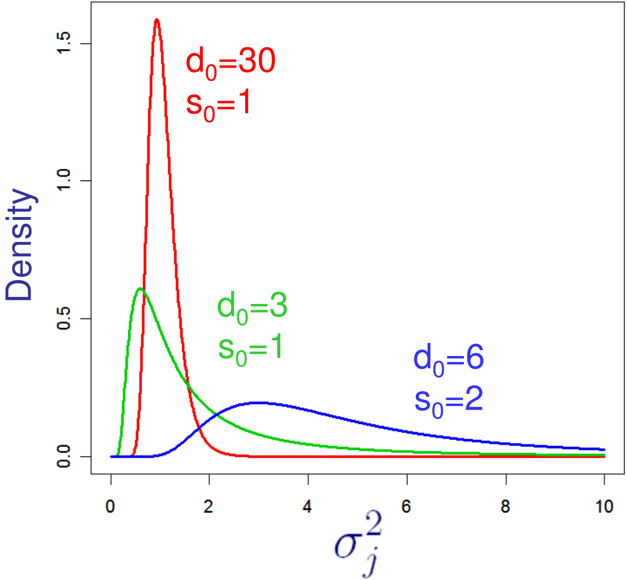
DE for count data

- *Idea :* use this same strategy in the case of *count data*
- One extreme : common dispersion parameter for every gene
- This assumption is very unrealistic
- Other extreme : estimate separate dispersion parameter *independently* for each gene
- This procedure gives poor estimates especially when the number of samples (libraries) is small
- 'Moderated' : *shrink* individual estimates toward a common parameter
- This problem is more challenging in this case :
    - The approach taken in `limma` is based on a *hierarchical model* – don't have that here
    - How to formulate statistical test (no t-distributions here)

Hierarchical model

- Linear model $E[\mathbf{Y}_g] = X\beta$; $Var(\mathbf{Y}_g) = W_g\sigma_g^2$
- $\hat{\beta}_{gj} \mid \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$
- $s_g^2 \mid \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$, where d_g is the residual df for the linear model for gene g
- Assume $P(\beta_{gj} \neq 0) = p_j$
- Prior $\frac{1}{\sigma^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}$
- Prior $\beta_{gj} \mid \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$
- *Posterior variance estimate :* $\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$
- $\Longrightarrow$ $\boxed{mod\ t = \frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{v_{gj}}}}$

Variance density examples



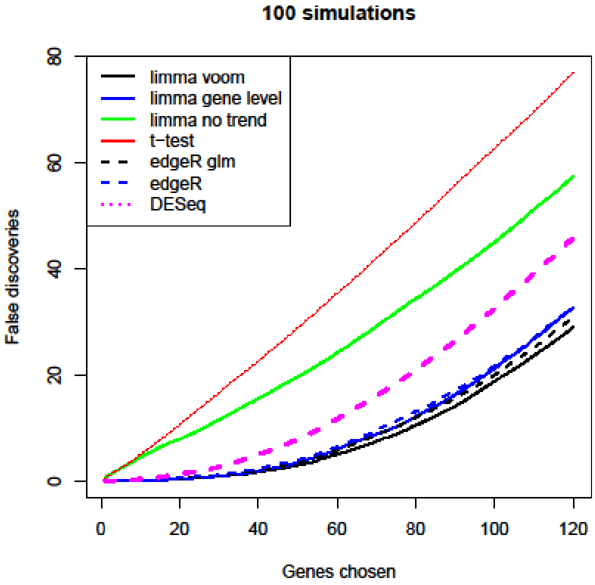
edgeR approach

- BioConductor package edgeRfor differential expression analysis of digital gene expression data
- edgeR estimates the genewise dispersions by *conditional maximum likelihood*, conditioning on the total count for that gene
- *Empirical Bayes* procedure is used to shrink the dispersions towards a consensus value $\Longrightarrow$ *borrowing information between genes*
- Differential expression is assessed for each gene using an *exact test* analogous to Fisher's exact test (but adapted for overdispersed data)

voom (from limma) approach

- The approach taken above was to *model* the count data, then analyze for DE according to that model
- A new, alternative approach is to *transform* the count data and use existing methods $\Longrightarrow$ voom function in limma
- In this approach, the idea is to transform RNA-Seq data so that they are ready for linear modeling
- You could then use limma as usual for assessing DE

DE methods comparison



Model mean-var relationship

On variance models for RNA-seq

- Mean-variance relationship is essentially *quadratic* for RNA-seq counts
- *Modeling the variation* is more important than getting the distribution right
- *Gene-specific variation* exists and must be accounted for

edgeR summary

- Fits an intuitive model
- The biological coefficient of variation (the biological variance divided by the mean expression) is interpretable
- Excellent statistical power
- It treats the dispersion as known (once estimated) and so test size can be a little liberal
- Can't estimate the optimal prior weight (the prior weight is used in the empirical Bayes shrinking of the dispersion estimates)
- Computationally challenging to program (e.g. fitting ≈ 30,000 GLMs, one per gene)

voom summary

- More ‘agnostic’ to the mean-variance relationship
- Does ‘natural’ (but *ad hoc*) fold change shrinkage
- Easily estimates the prior weight
- Holds test size since it tracks the uncertainty of the empirical Bayes estimates throughout the model
- Feeds into many existing `limma` tools
- Wins all comparisons with other methods (so far!)

BREAK

Examples `limma` and `edgeR`

- The procedure used in `edgeR` is analogous to the procedure used in `limma`
- Let's 'walk through' the process ...

About that exam...

- **Overall presentation :**
  - follow instructions regarding margins, point size, *etc.*
  - *plot labels* : increase using `plot` pars (`cex.axis`, *etc.*)
  - include figures as jpegs if your pdf file is too big
- **Intro/background :**
  - purpose of experiment/study and analysis
  - specify chip (*e.g.* Affymetrix U133A, or whatever chip) and number of probe sets (‘genes’)
- **Quality assessment :**
  - describe general approach/procedure : PLM, model fitting (robust regression), and briefly how the resulting quantities reflect data ‘quality’
  - pseudoimages of *weights* (or possibly residuals, if that ends up looking more informative)
  - NUSE plot (and possibly RLE if that adds information)

More about that exam...

- **Normalization :**
  - For Affy chips, use RMA – briefly describe model and result (a measure of gene expression)
- **DE :**
  - describe the model you are fitting, and define all parameters and notation
  - do not do a comparison of multiple testing procedures, choose a procedure and use that (most common in microarray studies to use B-H FDR ; do NOT use Bonferroni)
  - make sure that how you rank the genes is clear, and that it corresponds to the volcano plot (most common to use adjusted *p*-value for mod-*t*)

Even more...

- **Cluster analysis :**
  - clearly describe the distances and clustering algorithm you end up using
  - if you have both dendrogram and heatmap, include them as subfigures in the same figure
  - clearly state and interpret your findings
- **Conclusions :**
  - this can be brief, but should include any major findings, your comments, interpretations, recommendations
- **Gene list :**
  - on **1** single page ! ! ! !
  - make sure any values are *informative*
  - make ‘nicer’ table headings
- **R code :** must be *reproducible*