RNA-seq differential expression analysis

Mikhail Dozmorov

Spring 2018

Overview

- Models for count data
- Filtering low counts
- Multiple comparisons with low counts
- Exploratory analysis (Quality assessment)
- Data Preparation/Normalization
- Modeling and Moderating dispersion
- Linear Models for differential analysis
- Differential expression analysis

Differential expression statistics

- Sequence data differs from microarray data - they are counts
- We assume that in sample i, some percentage π_{ij} of the reads come from feature j
- We want to test whether π_{ij} varies with treatment

Differential expression statistics

- Count data has (at least) 3 sources of variability:
  - **Poisson** - due to the fact that each read either is or is not captures
  - **Biological** - due to sample differences (extra-Poisson variation)
  - **Systematic** - due to the "treatments"
- The goal of differential expression analysis is to identify systematic variability due to the "treatments"
- This is typically measured in standard deviations taking into account both the technical and biological variability

Models for count data

- $N(i)$ - sequencing depth (size factor) - the number of mapper reads in sample i
- We assume that in sample i, some percentage π_{ij} of the reads come from feature j
- When $N_i\pi_{ij}$ is small, the observed number of reads from feature j in sample i should come from Poisson distribution with mean $N_i\pi_{ij}$
- Technical replication confirms this
- But due to biological variability, π_{ij} varies among replicates

Marioni (2008) Genome Res (https://www.ncbi.nlm.nih.gov/pubmed/18550803)

Models for count data

- For Poisson data, *variance* $= N\pi =$ *mean*, i.e. $\sigma_i^2 = \mu_i$
- But due to biological variation, $\sigma_i^2 > \mu_i$
- A simple model is $\sigma_i^2 = \mu_i(1 + \phi_i \mu_i)$
- ϕ_i is called the dispersion

Data generative model for replicated RNA-seq

- Assume data are properly normalized.
- For a sample with M replicates, the counts for gene i replicate j is often modeled by following hierarchical model:
  $Y_{i,j}|\lambda_i \sim Poisson(\lambda_i), \lambda_i \sim Gamma(\alpha, \beta)$
- Marginally, the Gamma-Poisson compound distribution is Negative binomial. So the counts for a gene from multiple replicates is often modeled as Negative binomial: $Y_{i,j} \sim NB(\alpha, \beta)$.

A little more about the NB distribution

- NB is over-dispersed Poisson
  - Poisson: $var = \mu$
  - NB: $var = \mu + \mu^2\phi$
- Dispersion parameter ϕ approximates the squared coefficient of variation: $\phi = \frac{var-\mu}{\mu^2} \approx \frac{var}{\mu^2}$
- Dispersion ϕ represents the biological variance, so shrinkage should be done for ϕ
- NB distribution can be parameterized by mean and dispersion, but there's no conjugate prior for ϕ

Warnings about the use of negative binomial

Some transcript data does NOT fit the negative binomial model

- Transcripts which have zero counts in some samples and moderate counts in others from the SAME treatment
- Transcripts that are particularly prevalent in some tissues
- These that have high dispersion - possibly should be analyzed separately but not filtered

RPKM/FPKM and TPM are not supported statistically for differential expression analysis

Simple ideas for DE

- Transform data into continuous scale (e.g., by logarithm) then use microarray methods:
  - Troublesome for genes with low counts.
- For each gene, perform two group Poisson or NB test for equal means. But:
  - Number of replicates are usually small, asymptotic theories don't apply so the results are not reliable.
  - Like in microarray, information from all genes can be combined to improve inferences (e.g., variance shrinkage).

Differential expression analysis

- `limma` - Linear Models for Microarray Data
- `voom` - variance modeling at the observational level transformation. Uses the variance of genes to create weights for use in linear models
- After `voom` transformation, the RNA-seq data can be analyzed using `limma`

https://bioconductor.org/packages/release/bioc/html/limma.html

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29

https://gist.github.com/mdozmorov/fb7a1f40eb18699298442c3e77a0de02 - Differential expression analysis in RNA-seq, short
https://gist.github.com/stephenturner/e34e32b3d054bb850ae2 - Differential expression analysis in RNA-seq, long

voom

- Treat log-CPM analogous to log-intensity values from a microarray experiment, but log-CPM cannot be treated as having constant variances
- Estimate non-parametrically the mean-variance trend of the log-CPMs and to use this relationship to predict the variance of each log-cpm value
- The predicted variance is then encapsulated as an **inverse weight** for the log-cpm value
- When the weights are incorporated into a linear modeling procedure, the mean-variance relationship in the log-cpm values is effectively eliminated

voom

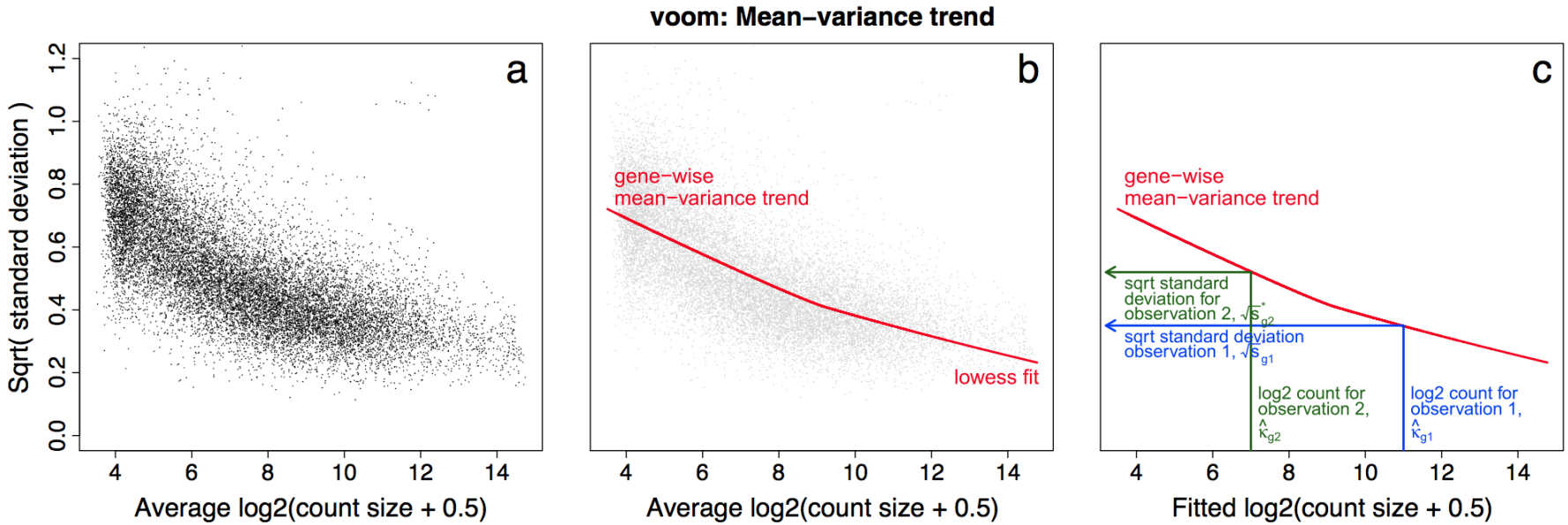


Figure 2 voom mean-variance modeling. (a) Gene-wise square-root residual standard deviations are plotted against average log-count. **(b)** A functional relation between gene-wise means and variances is given by a robust LOWESS fit to the points. **(c)** The mean-variance trend enables each observation to map to a square-root standard deviation value using its fitted value for log-count. LOWESS, locally weighted regression.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." Genome Biology 15, no. 2 (2014): R29. https://doi.org/10.1186/gb-2014-15-2-r29.

DEseq

- Counts are assumed to follow NB, parameterized by mean and variance $K_{ij} \sim NM(\mu_{ij}, \sigma^2_{ij})$
- The variance is the sum of shot noise and raw variance $\sigma^2_{ij} = \mu_{ij} + s^2_j \nu_{i,p(i)}$
- The raw variance is a smooth funciton of the mean: assumes that genes with same means will have the same variances
- "This assumption allows us to pool the data from genes with similar expression strength for the purpose of variance estimation"
- DESeq pulls the low dispersions towards the common value, but leaves the high dispersions, which is more conservative

Anders at. al. 2010, GB https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106

Bioconductor package DEseq

Inputs are:
- integer matrix for gene counts, rows for genes and columns for samples. - experimental design: samples for the columns.

```
library(DESeq)
conds=c(0,0,0,1,1,1)
cds=newCountDataSet(data, conds )
cds=estimateSizeFactors( cds )
cds=estimateVarianceFunctions( cds )
fit=nbinomTest( cds, 0, 1)
pval.DEseq=fit.DEseq$pval
```

https://bioconductor.org/packages/release/bioc/html/DESeq.html
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

The `rlog` transformation

- For RNA-seq, the variance across samples grows with the count mean.
- Highly expressed genes will dominate in clustering, PCA, due to the largest absolute differences between samples.
- log_2-transformation helps, but it "amplifies" the strong Poisson noise inherent to small count values.

The `rlog` transformation

- The *regularized-logarithm transformation* or *rlog*. `DESeq2::rlog`
- For genes with high counts, the rlog transformation will give similar result to the ordinary log2 transformation of normalized counts.
- For genes with lower counts, however, the values are shrunken towards the genes' averages across all samples.
- Using an empirical Bayesian prior on inter-sample differences in the form of a ridge penalty, the rlog-transformed data then becomes approximately homoskedastic, and can be used directly for computing distances between samples and making PCA plots.

edgeR

- From a series of papers by Robinson et al. (the same group developed limma): 2007 Bioinformatics, 2008 Biostatistics, 2010 Bioinformatics.
- Empirical Bayes ideas to "shrink" gene-specific estimations and get better estimates for variances.
- The parameter to shrink is over-dispersion (ϕ) in NB, which controls the within group variances.
- There is no conjugate prior so a shrinkage is not straightforward.
- Used a conditional weighted likelihood approach to establish an approximate EB estimator for ϕ.

Bioconductor package edgeR

- Inputs are the same as DEseq: an integer matrix for counts and column labels for design.

```
library(edgeR)
d = DGEList(counts=data, group=c(0,0,0,1,1,1))
d = calcNormFactors(d)
d = estimateCommonDisp(d)
d = estimateTagwiseDisp(d, trend=TRUE)
fit.edgeR = exactTest(d)
pval.edgeR = fit.edgeR$table$p.value
```

https://bioconductor.org/packages/release/bioc/html/edgeR.html
https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616

Filtering low expressing genes

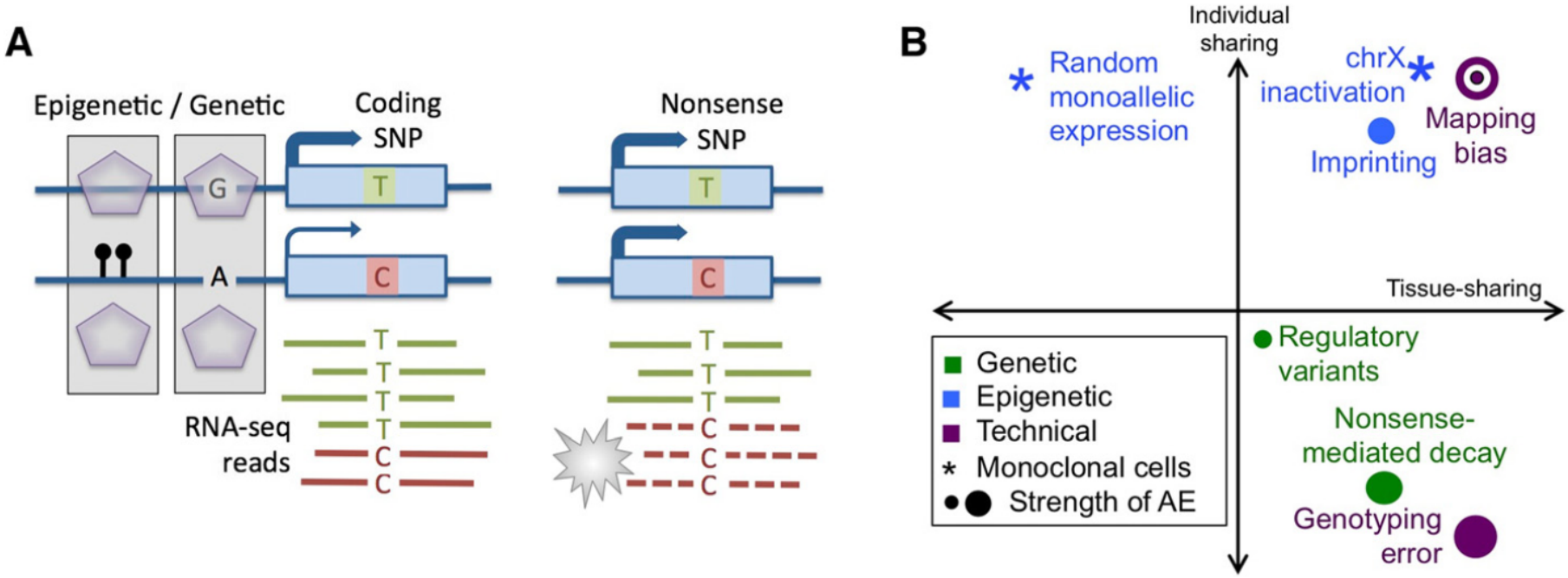
- The model for each gene does not depend on the counts
- On the other hand, the power of statistical tests for count data depends on the mean counts
- For this reason, we may prefilter the data to eliminate features with very low total read counts before going differential expression analysis
- If we have more samples or more reads per sample or aggregate features we can increase the power to detect differential expression

Summary for DE test

- The methods we talked about are based on the gene counts.
- DESeq and edgeR are the most popular software for that.
- Both use Negative Binomial distribution
- Differ in estimation of the dispersion parameter
- There are other methods performing transcript level expression estimation and DE analysis: **cufflink** and **cuffdiff**.

Allele-specific expression analysis

- Human genome is diploid - paternal and maternal DNA may differ
- Gene expression may differ depending on allele



- Problems: small insertion/deletion analysis is challenging

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. "Tools and Best Practices for Data Processing in Allelic Expression Analysis." Genome Biology 16 (September 17, 2015): 195. https://doi.org/10.1186/s13059-015-0762-6.

Allele-specific expression analysis

Application	Publication	Software / Package	AE Statisitcal Test	Input Data
QTL Detection	de Geijn et al., 2014 *	WASP	Beta-binomial	Population Data: Genotypes, RNA-Seq
QTL Detection	Kumasaka et al., 2015 *	RASQUAL	Beta-binomial	Population Data: Genotypes, RNA-Seq
Nonsense Mediated Decay	Pirinent et al., 2015	MAMBA	Binomial	Site Level Read Counts
Imprinting	Baran et al., 2015	N/A	Beta-binomial	Site Level Read Counts
AE in F1 Individuals	Pandey et al., 2013	Allim	Binomial	Parental Genotypes, F1 RNA-Seq
Gene Level AE	Romanel et al., 2015	ASEQ	Fisher Exact Test (DNA vs RNA read counts)	DNA-Seq, RNA-Seq
Gene Level AE	Mayba et al., 2014	MBASED	Beta-binomial	Site Level Read Counts
Gene Level AE	Skelly et al., 2011	N/A	Beta-binomial	DNA-Seq, RNA-Seq
Site Level AE	Rozowsky et al., 2011	Allele-Seq	Binomial	Trio Genotype Data, RNA-Seq
GUI for AE Analysis	Soderlund et al., 2014	Allele Workbench	Binomial	Sample Genotype, RNA-Seq
Site Level Read Counts	This Publication	GATK ASEReadCounter	None Performed	Sample Genotype, RNA-Seq

https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0762-6/MediaObjects/13059_2015_762_MOESM10_ESM.xlsx