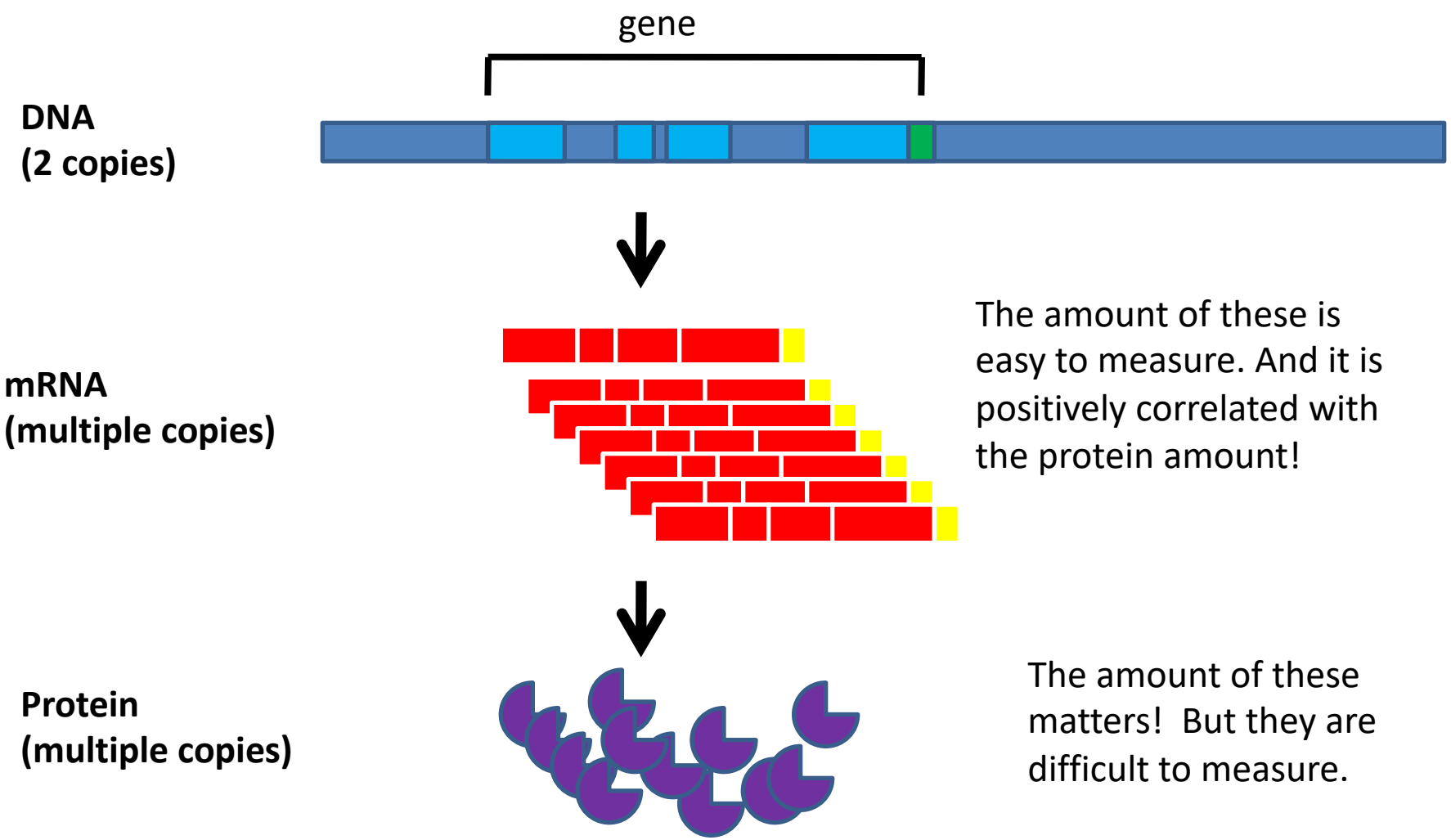
Introduction to RNA-seq data analyses

Outline

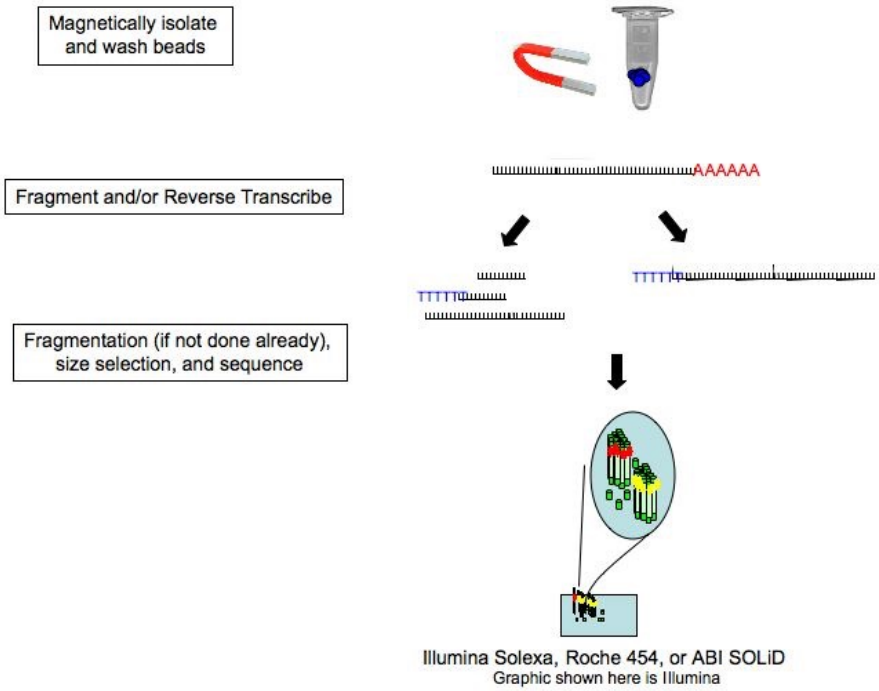
- Biological motivations
- Experimental procedures.
- Analyses of “bulk” RNA-seq data: methods, useful software tools, and Bioconductor packages.
  - Data summarization and normalization.
  - Differential expression.

Review: Gene expression levels are measured through their mRNA abundance.



RNA-seq experiment

1. Extract RNA from samples.
2. Generate cDNA.
3. Fragmentation (cut cDNA into small pieces), then select the fragments with certain lengths.
4. Sequence the fragmented cDNA.



Measuring mRNA abundance

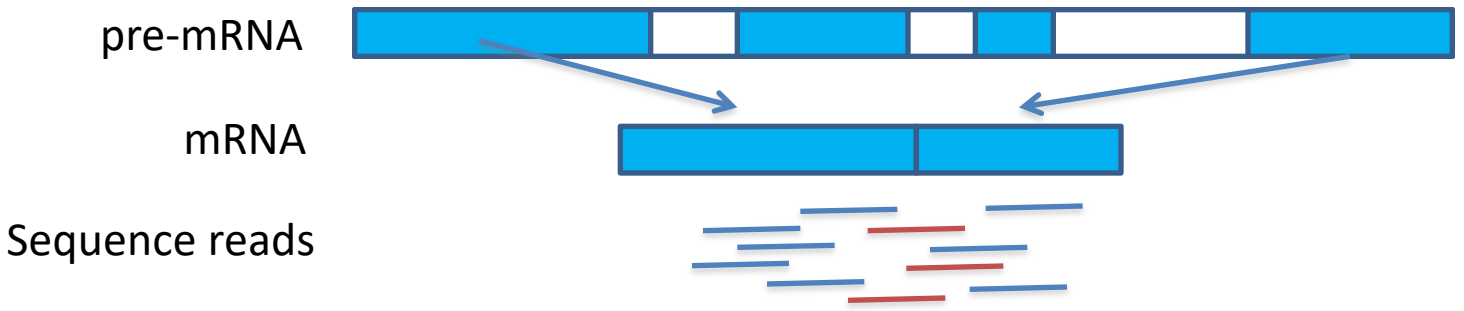
- Using microarray:
  - Probes are designed to target genes.
  - mRNA are converted to cDNA, labeled by dyes, hybridized to microarray (cDNA are attracted to probes with complementary sequences).
  - High gene expression -> more cDNA -> corresponding probes have higher intensity values (brighter).
- Using RNA-seq:
  - Sequence the cDNA, then align all reads.
  - High gene expression -> more cDNA -> more reads aligned to the genes.
  - Difference from microarrays: hybridization is replaced by sequencing.

Beyond gene expressions

- RNA-seq provide much more information than gene expression microarrays. In additional to gene expressions, it provides information for:
  - alternative splicing.
  - structural changes of genes: gene fusion.
  - new genes/exons.

Alternative splicing

- Definition: the same pre-mRNA produces different mRNA products, through joining different exons.
  - Locations where two exons join is called “junction”.
  - Can be detected and quantified using **exon arrays**, on which the probes are designed to target the junction regions.
  - From RNA-seq: look at “junction reads”, which are reads overlap two exons.



Structural modification of transcriptome

- Example: gene fusion.
  - Two or more separate genes are “fused” together to form a new gene.
  - Often associated with cancer.
- Cannot be detected from microarray.
- Can be detected from “Paired-end” RNA-seq:
  - If a pair of reads are very far apart on the reference genome, it suggests gene fusion. Because the DNA segments are selected based on the sizes, and there shouldn’t be long cDNA segments.

Other information from RNA-seq

- Finding new genes/exons:
  - Reads aligned to genomes with no gene/exon annotation are possibly from new gene/exon.
- mRNA processing efficiency (Bai *et al.* 2013, PNAS):
  - The percentage of reads mapped to introns could indicate the efficiency of mRNA processing (like splicing).
- RDD: RNA-DNA difference (Li *et al.* 2011, *Science*):
  - The transcription process (DNA -> RNA) is not perfect, based on different SNPs detected from DNA and RNA-seq data. This is very controversial!!

RNA-seq data analyses

- Raw data: sequence reads.
- First step: alignment to the reference genome.
- Information for different tasks:
  - expression: read counts in genes/exons.
  - alternative splicing: junction reads.
  - gene fusion: distances between paired reads.
- We will focus on expression analysis in this class.

Sequence alignment

- RNA-seq reads can be aligned using a general aligner such as bowtie.
- The reads should be all from **transcriptome** (set of genes).
- RNA-seq specific aligners are available
  - General: Tophat, STAR, HISAT
  - Pseudoaligner*:* Salmon, kallisto.
  - In Bioconductor: Rsubread, QuasR

Gene expression analysis

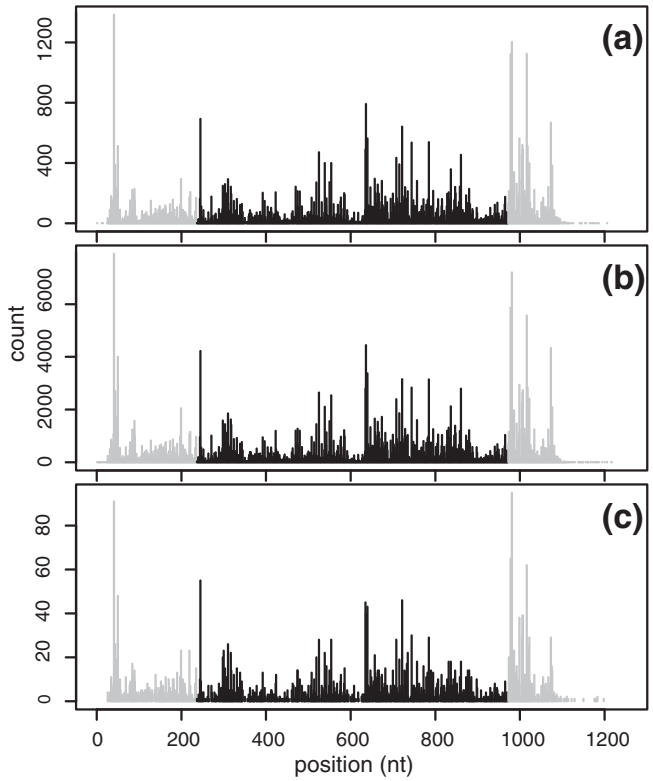
- Biological motivation is the same as gene expression microarrays: compare the expression of genes between different samples.
- Steps:
  - summarization: get a number for each gene to represent its expression level.
  - normalization: remove technical artifacts so that data from different samples are comparable.
  - differential expression detection: gene by gene statistical test.

Summarization of read counts

- From RNA-seq, the alignment result gives the chromosome/position of each aligned read.
- For a gene, there are reads aligned to the gene body. How to summarized them into a number for the expression?
- Easiest: simply count the number, then normalized by total number of reads in the experiment, and optionally gene lengths
  - RPKM (reads per kilo-bp per million reads). Mortazavi *et al.* (2008), *Nature Method*.
  - RPM/CPM/TPM (reads/counts/transcript per million reads): without normalization by gene length.

Artifacts in the read distribution

- The reads are NOT uniformly distributed within gene bodies. It affects by many things such as the sequence composition, chromatin structure, etc.



Li et al. (2010) *Genome Biology*

Weighted sum (Hensen et al. 2010 *NAR*)

- Discovered that reads from Illumina has a 7-bp motif at beginning: there are more reads started with certain 7-bp due to technical artifacts (the random priming bias).
- Down-weight the reads started with the motif.

$$w(h) = \frac{\frac{1}{6}\sum_{i=24}^{29} \hat{p}_{hep:i}(h)}{\frac{1}{2}(\hat{p}_{hep:1}(h)+\hat{p}_{hep:2}(h))}$$

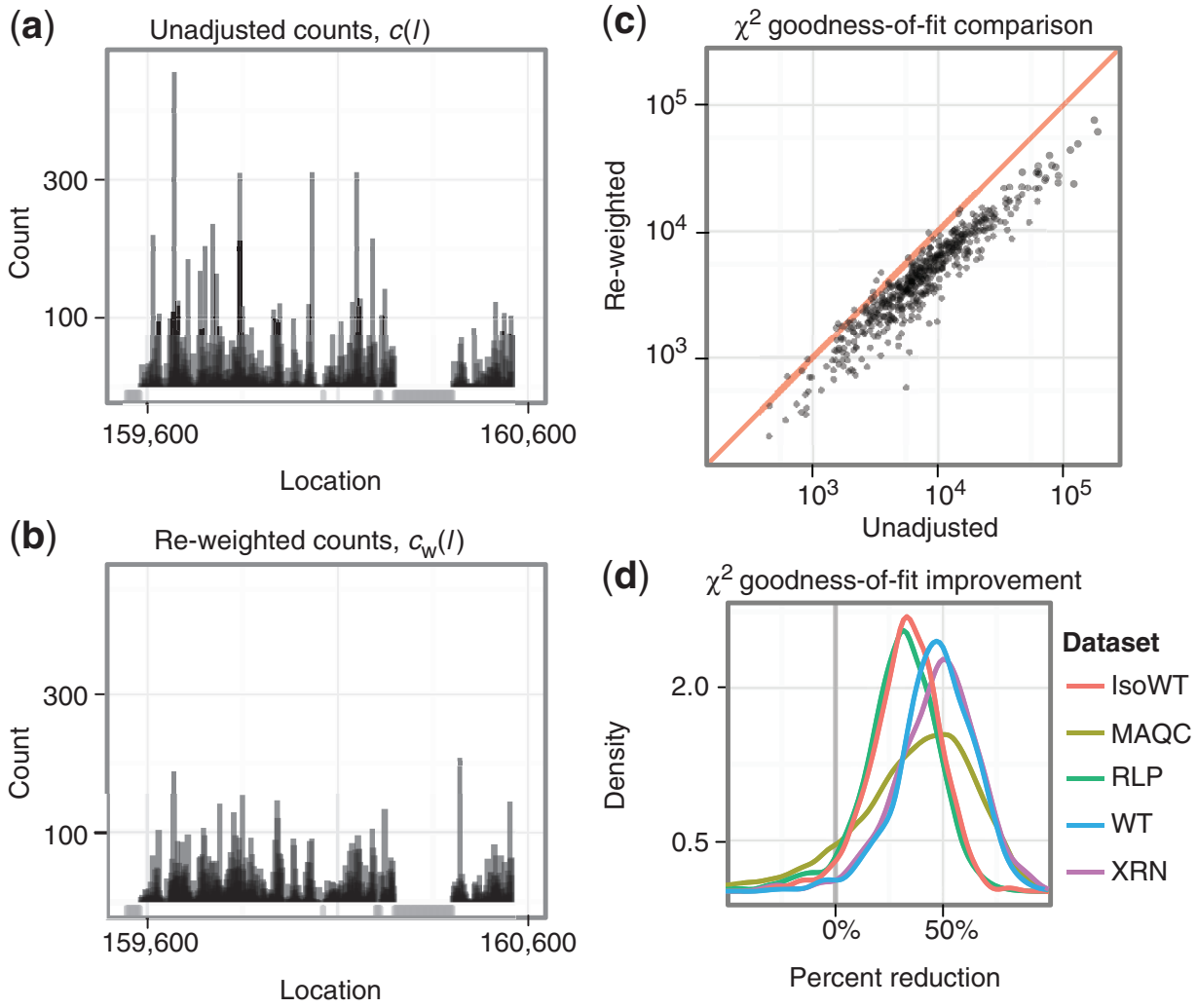
  - *w(h)*: weights for reads starting with heptamer *h*.
  - $\hat{p}_{hep:i}$: observed distribution of heptamers starting at position *i*.

Table 1. Data from a small genomic region in the sense strand of the YOL086C gene in *S. cerevisiae*

Strand	Location l	Heptamer $h(l)$	Count $c(l)$	Weight $w(h(l))$	Reweighted count $c_w(l)$
...					
−1	159792	TTGGTCG	17	1.39	23.6
−1	159793	TTTGGTC	17	0.25	4.3
−1	159794	TTTTGGT	65	0.31	20.4
−1	159795	GTTTTGG	72	0.32	23.3
−1	159796	CGTTTTG	10	1.66	16.6
...					

$c(l)$ denotes the number of mapped reads starting at a particular (stranded) location l and $h(l)$ is the unique heptamer associated with this location. $w(h(l))$ are weights such as in Equation (1) and $c_w(l) = c(l)w(h(l))$ are the location-specific reweighted counts. For this particular small genomic region, reweighting makes the counts more comparable between different locations. Data from the WT experiment.

Results: reweighting increase uniformity of read distribution within gene body



Model the read counts as a function of base compositions (Li et al. 2010 *GB*)

- Log-linear model: for nucleotide j of gene i,
  - n_{ij}: number of reads starting at this position.
  - μ_i: true expression of the gene, $v_i=log(\mu_i)$.
  - ω_{ij}: sequence biases at this position.
- Model: let $\mu_{ij}=\mu_i*\omega_{ij}$, assume $n_{ij}|\mu_{ij} \sim Poisson(\mu_{ij})$, and

$$\log(\mu_{ij}) = v_i + \alpha + \sum_{k=1}^{K} \sum_{h \in \{A,C,G\}} \beta_{kh} I(b_{ijk} = h)$$

- Non-linear model: MART (multiple additive regression trees).

Results

- Results from the (linear or non-linear) model are estimated gene expression.
- Comparing the correlation with microarray data, MART model is slightly better than using sum:

Table 4: Spearman's rank correlation coefficients in mouse embryoid bodies

Fold change bin	SCC by uniform model	SCC by our MART model	Relative improvement
(1.00, 1.09)	0.465	0.466	0.1%
(1.09, 1.19)	0.437	0.444	1.4%
(1.19, 1.33)	0.413	0.434	5.1%
(1.33, 1.53)	0.481	0.520	8.2%
(1.53, 4.82)	0.389	0.490	26.0%

SCC: Spearman's rank correlation coefficient.

More complicated likelihood approach

- Roberts *et al.* (2011) *GB:*
  - Denote the transcript abundance by ρ.
  - Focus on relative expressions: $\sum_{t \in T} \rho_t = 1$
  - Whole data likelihood:

$$L(\rho|F) = \left(\prod_{g \in G} \beta_g^{X_g} \right) \left(\prod_{g \in G} \left(\prod_{f \in F: f \in g} \sum_{t \in g} \gamma_t \cdot D(t,f) \cdot \frac{b(t, e_{5'}(t,f), e_{3'}(t,f))}{B(t, I_t(f))} \right) \right)$$

  - Maximize the likelihood using iterative methods and obtain the estimates of relative expression.

Summary

- Sequence reads are not uniformly distributed within gene body.
- The distribution is highly dependent on sequence compositions.
- Proposed methods didn’t provide convincing performance improvements over simple method like RPKM in downstream analysis.

Data normalization

- Data from different samples need to be normalized so that they are comparable.
- Most important – sequencing depth: sample with more total counts will have more counts in each gene on average.
- Easiest method: normalize by the total number of counts – RPM/TPM.

Data generative process

- The read counts from RNA-seq follow a sampling process.
- for gene *i*, *i=1, …, G,* let
  - the true expression (number of cDNA fragments) be μ_i.
  - gene length be L_i.
- 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 $\lambda_i = Np_i, \quad Y_i \mid \lambda_i \sim Poisson(\lambda_i)$
- 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 approximate the same, normalizing by N will be sufficient – this is what RPM 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!
  - There is no such problem in microarray, because the mRNAs don't "compete" with each other.
- 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.
- Bullard *et al. (*2010), *BMC Bioinformatics:*
  - use counts from house keeping genes.
  - use a certain quantile (75^{th}) for all counts.
- Anders et al. (2010), *Genome Biology*:
  - median of the ratios of observed counts.
- Robinson et al. (2010), *Genome Biology*: TMM (trimmed mean of M values).
  1. compute M (log fold changes) and A (log total counts) for all genes.
  2. Discard genes with extreme M and A values, and compute a weighted mean of M's for the rest of genes. The weights as the inverse of the approximate asymptotic variances.
  3. Underlying assumption is that most genes are not DE.

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

- Hansen et al. (*Biostatistics 2012*):
  - The gene-specific biases (from GC content, gene length, etc.) need to be considered.

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

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

true expression

biases, e.g., GC content

- A conditional quantile normalization (cqn) procedure is designed to estimate *h* and *f*, and then ϑ.

Summary

- RNA-seq normalization can be difficult.
- 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.

Differential expression

- Biological motivation: to find DE genes.
- Microarray methods are not directly applicable: continuous vs. count data, but ideas can be borrowed.
- Usually needs multiple replicates per sample, so that the means and variances can be evaluated.
- Test is carried gene by gene.

Data model for RNA-seq

- For a sample with replicates, the counts for gene *i* replicate *j* is often modeled by following hierarchical model:

$$Y_{ij} \mid \lambda_i \sim Poisson(\lambda_i), \lambda_i \sim Gamma(\alpha, \beta)$$

- Marginally, the Gamma-Poisson compound distribution is Negative binomial, thus

$$Y_{ij} \sim NB(\alpha, \beta)$$

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: $\phi = \frac{var-\mu}{\mu^2} \approx \frac{var}{\mu^2}$
- ϕ represents the “biological variance”.
- NB distribution can be parameterized by mean and dispersion.

Simple ideas for DE

- Transform data into continuous scale (e.g., by logarithm) then use microarray methods:
  - Doesn't work well 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).

DEseq (Anders *et al.* 2010, *GB*)

- Counts are assumed to follow NB, parameterized by mean and variance: $K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)$,
- The variance is the sum of shot noise and raw variance:

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}.$$

- The raw variance is a smooth function of the mean: assuming genes with same mean have the same variance.
- Hypothesis testing using exact test:

$$p_i = \frac{\sum_{a+b=k_{iS},\ p(a,b)\le p(k_{iA},k_{iB})} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)}.$$

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
```

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 the dispersion (ϕ) in NB, which controls the within group variances.
- There is no conjugate prior for ϕ so a shrinkage is not straightforward.
- Used a conditional weighted likelihood approach to establish an approximate EB estimator for ϕ.

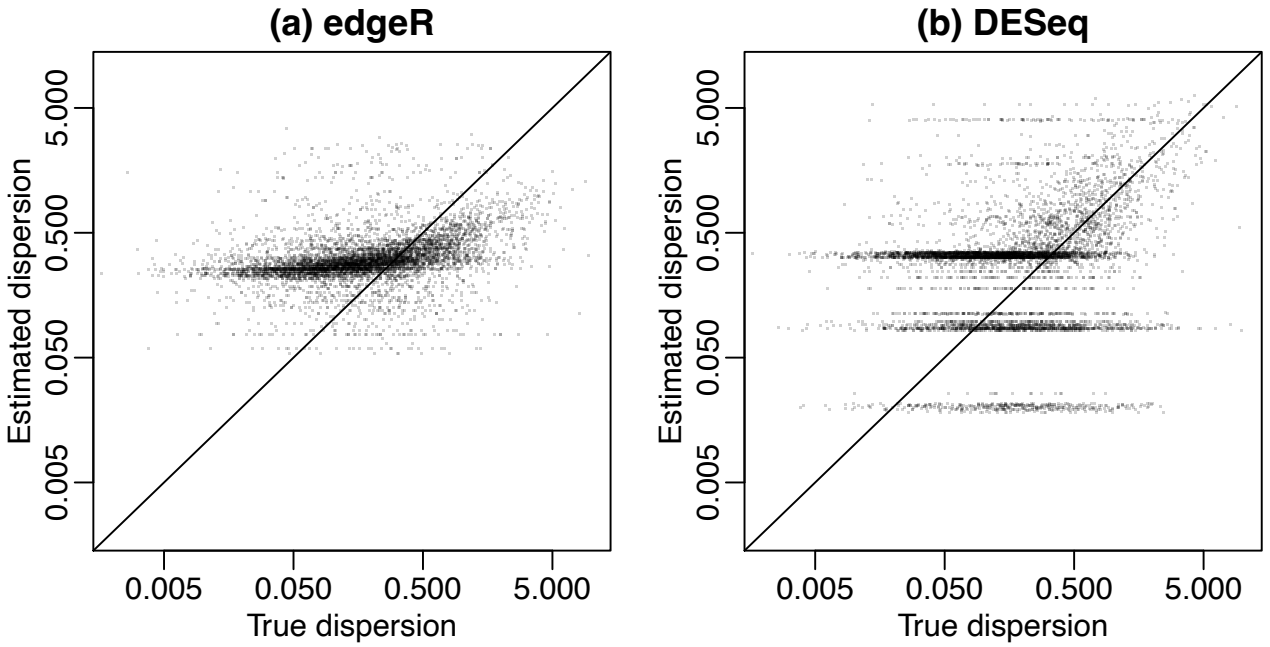
Bioconductor package edgeR

- 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),
          lib.size=colSums(data))
d = calcNormFactors(d)
d = estimateCommonDisp(d)
d = estimateTagwiseDisp(d, trend=TRUE)
fit.edgeR = exactTest(d)
pval.edgeR = fit.edgeR$table$p.value
```

DSS (Wu *et al.* 2012, *Biostatistics*)

- Found that the shrinkage from DESeq and edgeR are too strong.



A hierarchical model for the data

$$Y_{gi}|\theta_{gi} \sim \text{Poisson}(\theta_{gi} s_i)$$

$$\theta_{gi}|\phi_g \sim \text{Gamma}(\mu_{g,k(i)}, \phi_g)$$

$$\phi_g \sim \text{log-normal}(m_0, \tau^2)$$

- Y_{gi}: observed counts for gene *g*, sample *i*
- Θ_{gi}: unobserved true expression for gene *g*, sample *i*
- Φ_g: dispersion (related biological variance) for gene *g*.
- s_i: library size for sample *i*.

The posterior

- Negative binomial is parameterized by mean and dispersion, then the posterior for dispersion is:

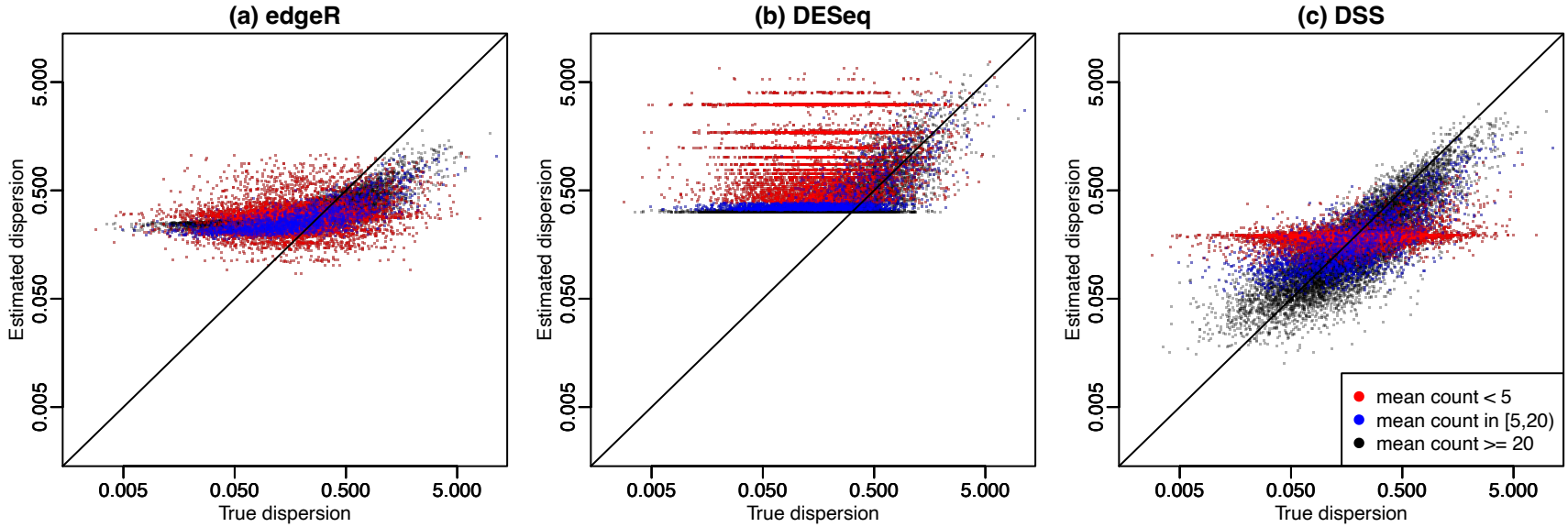
$$\begin{aligned}\log[p(\phi_g|Y_{gi},\nu_{gi},i=1,\ldots,n)] \propto & \sum_i \psi(\phi_g^{-1}+Y_{gi}) - n\psi(\phi_g^{-1}) - \phi_g^{-1}\sum_i \log(1+\nu_{gi}\phi_g) \\ & + \sum_i Y_{gi}[\log(\nu_{gi}\phi_g) - \log(1+\nu_{gi}\phi_g)] \\ & - \frac{[\log(\phi_g)-m_0]^2}{2\tau^2} - \log(\phi_g) - \log(\tau), \qquad (4.1)\end{aligned}$$

- Here, $\nu_{gi} = \mu_{g,k(i)} s_i$ is the expected value for Y_{gi}.
- It's a penalized likelihood to penalize dispersions far away from prior mean.

Testing and inference in two-group comparison

- Wald test: $t_g = \frac{\hat{\mu}_{g,1} - \hat{\mu}_{g,2}}{\sqrt{\hat{\sigma}^2_{g,1} + \hat{\sigma}^2_{g,2}}}$
  - With dispersions, variances can be computed according to NB distribution: $var = \mu + \mu^2\phi$
  - The variance for $\hat{\mu}_{g,1}$ is: $\hat{\sigma}^2_{g,1} \equiv \frac{1}{n_1^2}\left[\hat{\mu}_{g,1}\left(\sum_{j:k(j)=1}\frac{1}{s_j}\right) + n_1\hat{\mu}^2_{g,1}\tilde{\phi}_g\right]$
- Inferences: use normal P-values and local FDR.

Results: estimation of dispersions



DESeq2

- DESeq2 adopted similar approach as DSS, with additional shrinkage on the means.
- Provide comprehensive functionalities for DE.

```
Library(DESeq2)
countData <- matrix(1:100,ncol=4)
condition <- factor(c("A","A","B","B"))
dds <- DESeqDataSetFromMatrix(countData,
                  DataFrame(condition), ~ condition)
dds <- DESeq(dds)
results(dds)
```

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.
- There are other methods perform **transcript** level expression estimation and DE analysis: cufflink and cuffdiff.

Review

- RNA-seq provides information for:
  - expression.
  - Alternative splicing.
  - Structural variation, e.g., gene fusion.
- Statistical problems include:
  - Summarization.
  - Normalization.
  - differential expression testing.
  - Many others.