Bioinformatics Worksho

Lab 5B: Dummy RNA-seq Data Analysis

Maoying,Wu

ricket@sjtu.edu.cn

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Overview

In this session we will cover 2 out of the 3 main computational challenges of RNA-seq data analysis for counting applications:

1. Read mapping: Placing short reads in the genome
2. Transcriptome reconstruction: Finding the regions that originated the reads.
3. Expression quatification
   - Assigning scores to regions
   - Finding regions that are differentially represented between two or more samples.

Outline

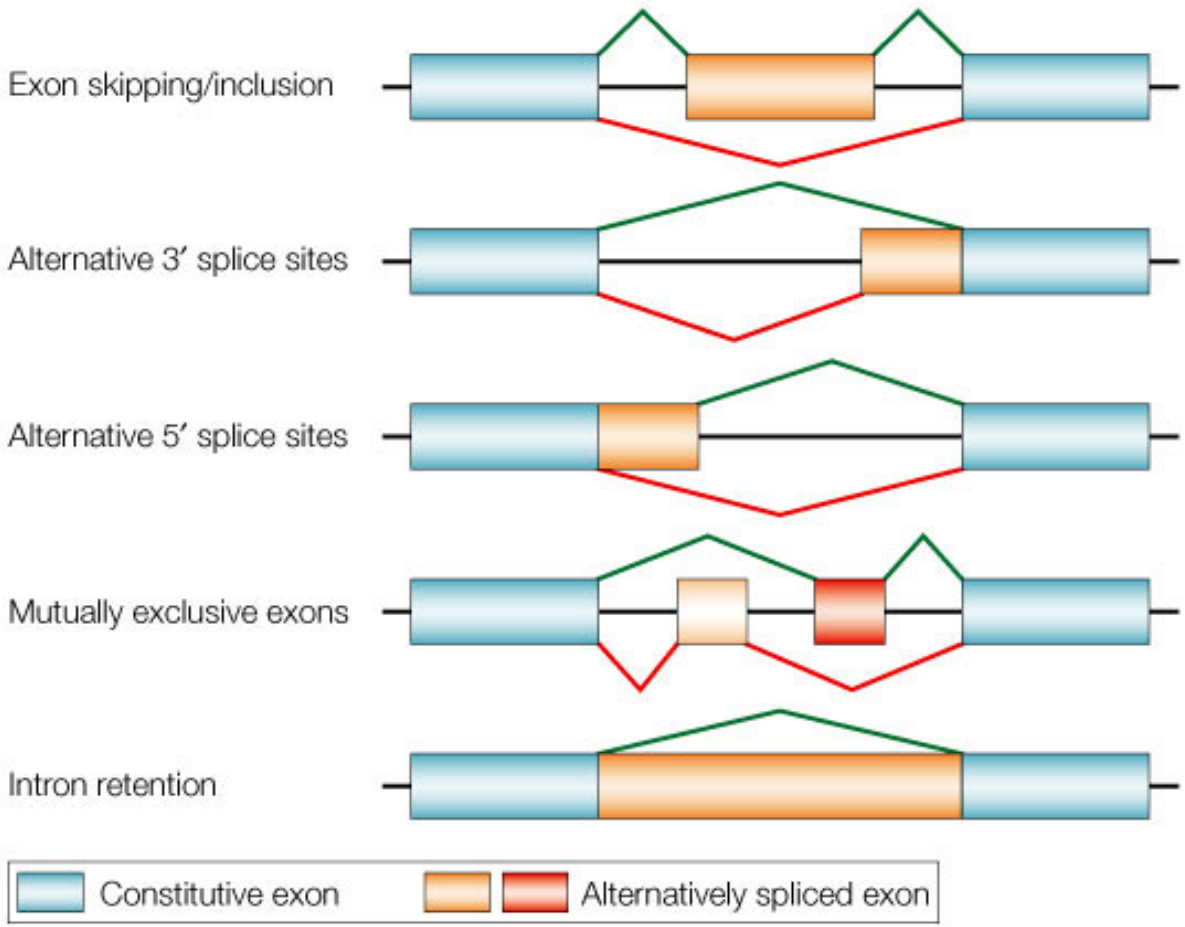
1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

Outline

Transcriptome: RNAs

- Total RNAs
  - Poly-adenylated (coding) RNAs, "mRNAs"
  - Short non-coding RNAs (ncRNAs), "small RNAs"
  - Long non-coding RNAs, "lncRNA"
  - Ribosome RNAs, "rRNAs"
- RNA Enrichment
  - PolyA-capture
  - Ribominus kit

RNA splicing: From pre-mRNA to mature RNA



Defining the alternative isoforms as well as their respective expression across tissues is critical for understanding biology.

Outline

1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

Transcriptome technologies

- High-throughput technologies
  - ESTs or cDNA sequencing (Sanger)
  - cDNA Microarray (Probes)
  - RNA Sequencing (NGS)
- Low-throughput technologies
  - qRT-PCR

Comparisons of the transcriptome technologies

Technology	Tiling microarray	EST sequencing	RNA-seq
Technology specification			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	5-100 bp	Single base	Single base
Throughput	High	Low	High
Require reference genomes?	Yes	No	In some case
Background noise	High	Low	Low
Application			
Sequence mapping	Yes	Limited for gene expression	Yes
Dynamic range to quantify	Up to a few-hundred fold	Not practical	$> 8,000$-fold
Isoform detection	Limited	Yes	Yes
Allele-specific expression	Limited	Yes	Yes
Practical issues			
Amount of RNA	High	High	Low
Cost	High	High	Relative low

Transcriptome technologies: Bring-home messages

- Microarray and ESTs have provided the proxy to capture the expressions of the transcripts (妯℃ 淇″). However, a prior knowledge of the genome of interest is a prerequisite.
- With RNA-Seq, you can get a fuller picture (panorama) over the whole transcriptome even if you don't have much knowledge of the transcriptome of interest. However, appropriate choices should be made for further analysis (板 淇″).
- qRT-PCR is a low-throughput technology used to validate the findings by the other high-throughput technologies.

Outline

1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Short read mapping
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

RNA-Seq: Standard protocol

standard protocol

(1) Extraction of RNA
(2) mRNA purification
(3) Reverse transcription of RNA to cDNA
(4) Fragmentation of RNAs
(5) Ligation of adapters
(5) Size selection
(6) PCR amplification (15 rounds or so)
(7) Injection into flowcell

This produces reads from poly-Aed RNA without strand information.

Variants

- Ribominus instead of polyA purification
- Strand-specific
- small RNA sequencing (direct ligation of adaptors to RNA)
- oligo(dT) priming instead of random hexamer priming

RNA-seq: Library preparation

Choosing a method

- What species?
- How good is your total RNA
  - intact, RIN
- How clean is your total RNA?
  - Genomic or non-target contamination
- How much total RNA do you have?
  - 1-200ng, 1-5μg, $> 10\mu$g
- What do you want to analyze?
  - mRNA, siRNA, miRNA, isoforms, ...

Example: PolyA selection protocol

How it works

- Life Technologies DynalBeads mRNA system
- Captures polyadenylated RNAs only
- Once polyA RNAs selected, RNAs fragmented, cDNAs generated with random primers, ds-cDNA, standard library preparation

Features/Limitations

- Requires 10-20μg total RNAs
- Total RNA must be of high integrity (瀹 锛 瑙 d 澶
  - Intact, RIN $>$ 8.0

Example: Ribosomal Depletion

How it works

- Illumina/Epicentre Robo-Zero systems
  - Human, mouse, rat, and bacterial
- Depletes most ribosomal RNAs, be certain to check species compatibility. Some rRNAs are not compatible (i.e. 5S-RNA fragments sometimes not removed)
- Allows for sequencing of mRNAs and non-coding RNAs. Very short RNAs (miRNAs) are mostly lost during library preparation
- Random primers =¿ ds-cDNA standard library preparation.

Features/Limitations

- Requires 1-5μg total RNA
- System allows for use of degraded RNA (LCM, FFPE)
- Genomic contamination appears to compress differentials
- Yields broadest range of RNA species

Example: Small RNA protocol

How it works

- TruSeq Small RNA Sample Prep System
- Allows for sequencing of small RNAs, particularlly mature miRNAs
- Adapters are ligated to total RNA sequentially, cDNA synthesis, library amplification, size selection by acrylamide gel cut.

Features/Limitations

- Requires 1μg total RNA
- Total RNA must be of high integrity (ie. RIN > 8.0)

RNA-Seq sequencers

RNA-seq: Sequencing

Sequencing FAQs

- How deep should I sequence?

- You need to balance between efficiency and cost.

RNA-seq: Sequencing

Sequencing FAQs

- How deep should I sequence?
- single-end or paired-end: $1x50$ or $2x100$

- You need to balance between efficiency and cost.
- For gene expression, 1x50 and 2x100 produce highly similar results.
- For novel isoform detection, 2x100 is preferential due to its longer reads.

RNA-seq: Sequencing

Sequencing FAQs

- How deep should I sequence?
- single-end or paired-end: 1x50 or 2x100
- Do I need technical replicates?

- You need to balance between efficiency and cost.
- For gene expression, 1x50 and 2x100 produce highly similar results.
- For novel isoform detection, 2x100 is preferential due to its longer reads.
- Biological replicates are far more valuable than technical replicates.

RNA-seq experiment issues: Bring-home messages

1. Be careful with your RNA sample quality: RIN.

RNA-seq experiment issues: Bring-home messages

1. Be careful with your RNA sample quality: RIN.
2. For your RNA types of interest, you need to select different library preparation techniques:
   - Different RNA enrichment methods
   - Different PCR primers

RNA-seq experiment issues: Bring-home messages

1. Be careful with your RNA sample quality: RIN.
2. For your RNA types of interest, you need to select different library preparation techniques:
   - Different RNA enrichment methods
   - Different PCR primers
3. Choose your sequencers according to the cost, efficiency and required sequencing depth, single-end or paired-end.
4. Use more biological replicates instead of technical replicates.

Outline

1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Short read mapping
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

RNA-Seq analysis workflow

1. Alignment of RNA reads to reference
   - Reference can be transcriptome or genome.
2. Count reads overlapping with annotation features of interest
   - counts for exonic regions, counts per exonic model
3. Normalization
   - Correction for sequencing depth and compositional bias.
4. Differential expression analysis (DEA)
   - Identification of significantly differentially expressed genes
   - Identification of strongly expressed genes
5. Special application
   - Splice variant discovery (semi-quantitative), isoform discovery, strand-specific expression, etc.
6. Clustering and classification analysis
   - Idenfification of genes with similar expression profiles.
7. Enrichment analysis of annotations
   - Functional analysis of obtained genes.

Outline

Two major algorithms for mapping

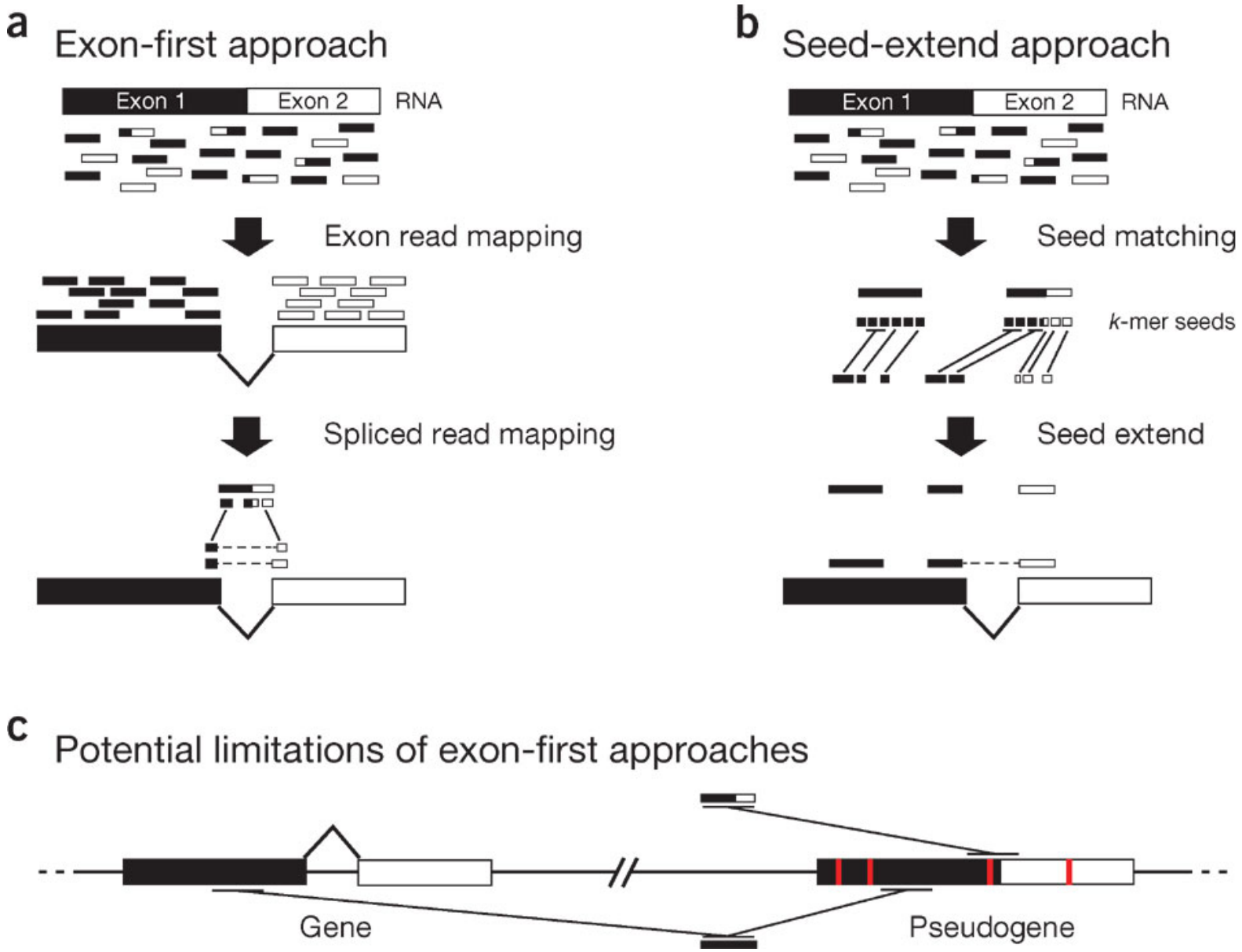
- **Unspliced read aligners** align reads to a reference without any large gaps.
  - Hash-based methods: MAQ and Stampy, appropriate for quantification of allele-specific gene expression.
  - Burrow-Wheeler transform (BWT) methods: BWA and Bowtie, appropriate for exact mapping

Two major algorithms for mapping

- **Unspliced read aligners** align reads to a reference without any large gaps.
  - Hash-based methods: MAQ and Stampy, appropriate for quantification of allele-specific gene expression.
  - Burrow-Wheeler transform (BWT) methods: BWA and Bowtie, appropriate for exact mapping

- **Spliced aligners** align the reads to the entire genome.
  - Exon-first approach: MapSplice, SpliceMap, and TopHat treats unspliced-mapped reads and unmapped reads separately.
  - Seed-extend approach: GSNAP and QPALMA breaks reads into short seeds for aligning to genome, which is followed by more sensitive extension and merging.

Spliced Mapping



Sequence alignment map (SAM/BAM) format

- SAM is a TAB-delimited alignment format consisting of a header section (lines starting with @) and an alignment section with 12 columns.
- BAM is the compressed, indexed and binary version of the SAM format.
- The below example contains the following:
  (1) bases in lower cases are clipped from the alignment;
  (2) reads $r001/1$ and $r001/2$ constitute a read pair;
  (3) $r003$ is a chimeric read;
  (4) $r004$ represents a split alignment.

SAM example

Alignment

```
Coor        12345678901234  5678901234567890123456789012345
ref         AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1           TTAGATAAAGGATA*CTG
+r002           aaaAGATAA*GGATA
+r003         gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

SAM format

```
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG   *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA      *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA         * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC         *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC               * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT           * NM:i:1
```

Outline

1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Short read mapping
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

RPKM (FPKM)

- Reads Per Kilobase Per Million Mapped Reads (RPKM)

$$\frac{10^9 G}{L \times S}$$

  where,
  - G: number of reads mapping to the GoI
  - S: total number of reads mapping to all the gene models for the sample
  - L: total length of the gene model for the GoI in bp.

- FPKM (fragments per kilobase per million mapped reads) is the paired-end version of RPKM.

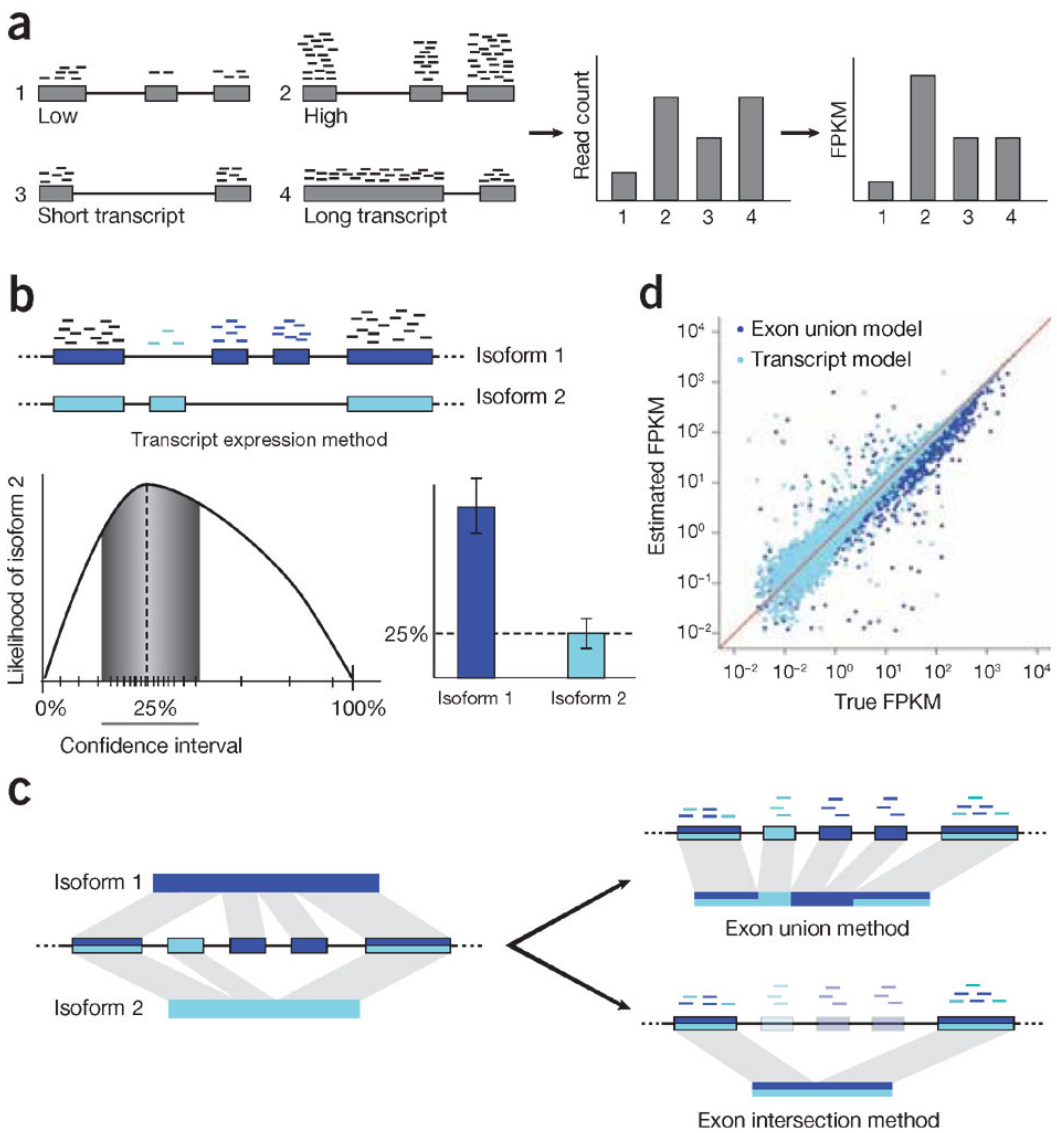
Be careful with RPKM (FPKM) values

- The more we sequence, the more reads we expect from each gene. This is the most relevant correction of this method.
- Longer transcript are expected to generate more reads. The latter is only relevant for comparisons among different genes which we rarely perform!
- RPKM/FPKM are not suitable for direct statistical testing:

library	**total #reads**	**#reads in treatment A**	**#reads in treatment B**
1	10^6	10	5
2	10^8	1000	500

- Thus, RPKM/FPKM are useful for reporting expression values, but not for statistical testing!

Gene expression quantification



Simple quatification methods

- **Exon-intersection** method, which counts reads mapped to its constitutive exons
- **Exon-union** method, which counts all reads mapped to any exon in any of the gene's isoforms.
- The **exon-intersection** method is analogous to expression microarrays, which typically probe expression signal in constitutive regions of each gene.

Cons

- exon-union model underestimates expression for alternative spliced genes;
- exon-intersection can reduce power for differential expression analysis.

Transcript-level expression tools

- Alexa-seq: Gene expression by constitutive exons
- ERANGE: Gene expression by using all exons
- Scripture: Gene expression by constitutive exons
- Cufflinks: Transcript deconvolution by solving the maximum likelihood problem.
- MISO: Transcript deconvolution by solving the maximum likelihood problem.
- RSEM: Transcript deconvolution by solving the maximum likelihood problem.

Outline

Why Normalization

Normalization is essential to ensure that the expression estimates are:

- **comparable among genes, transcripts, isoforms, etc**
- **comparable across samples, groups**
- **on a human-interpretable manner**

Normalization can control for the following aspects:

- Different RNA amounts (library size)
- Different Reverse Transcription efficiencies
- Different sequencing depth/error rates, etc.

Normalization is an essential step for a valid differential expression analysis:

- between transcripts within a sample
- between sample groups

Mathematics: Basic Poisson model

Number of reads from gene g in library i can be captured by a Poisson model (Marioni et al. 2008):

$$r_{ig} \sim \text{Poisson}(k_{ig}\mu_{ig}).$$

where μ_{ig} is the expression level of RNA in the library i and k_{ig} is a normalization factor:

$$\hat{\mu_{ig}} = \frac{r_{ig}}{k_{ig}}$$

- Normalization is all about deciding how to set k_{ig} such that the estimates of μ_{ig} are comparable between genes and across libraries.
- The number of reads r_{ig} is roughly proportional to
  - the effective length of the gene ℓ_g
  - the total number of reads in the library, N_i
- Therefore, if we set $k_{ig} = 10^{-3} \cdot \ell_g \cdot N_i \cdot 10^{-6}$, the units of $\hat{\mu}_{ig}$ are the so-called RPKM.

Scale normalization methods

- Total count (TC)
- Median (Med)
- Upper Quartile (UQ)(Bullard 2010, BMC Bioinformatics)
- Trimmed Means of Ms (TMM)(Robinson 2010, Genome Biol, `edgeR`)
- Geometric mean(Anders 2010, Genome Biol, `DESeq`)
- RPKM(Motazavi 2008, Nature Methods)
- Conditional quantile normalization (CQN)(Hansen 2010, Nucleic Acids Res.)
- Quantile (Q)(`normalizeQuantiles()` `limma`)

Langmead *et al.* (2010, Genome Biology) shows that it may be a good idea to use a gene-specific normalization factor.

Trimmed mean of M values (TMM) normalization

- RPKM normalization implicitly assumes that total RNA $\sum_g \mu_{ig}\ell_g$ is the same for all libraries.
- Poisson model is an approximation of Binomial model: $r_{ig} \sim \text{Binomial}(N_i, \frac{\mu_{ig}\ell_g}{\sum_j \mu_{ij}\ell_j})$
- However, sometimes this assumption does not hold.
- A better assumption: only the total expression for a **core gene set** G is similar:

$$\sum_{g\in G} \mu_{ig}\ell_g = \sum_{g\in G} \mu_{jg}\ell_g$$

- When this assumption does not hold, the naive MLE needs to be adjusted:
  - Calculate scaling factor for sample j relative to reference sample i:

$$\sum_{g\in G} \frac{r_{ig}}{N_i} \approx S^{i,j} \sum_{g\in G} \frac{r_{jg}}{N_j}.$$

  - Adjust the MLEs for sample j for all genes:

$$\hat{\mu}_{jg} = \frac{r_{jg}}{k_j g} = \frac{r_{jg}}{10^{-9} N_j \ell_g} \cdot S^{i,j}$$

TMM normalization (edgeR)

Then how to choose the subset G used to calculate $S^{(i,j)}$?

- For each pair of sample (i, j), compute the log-FoldChange (M) using the normalized counts
$$M_g^{(i,j)} = \log \frac{r_{ig}}{N_i} - \log \frac{r_j g}{N_j}$$
- and the mean of the log normalized counts (A):
$$A_g^{(i,j)} = \frac{1}{2}\left[\log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j}\right].$$
- Set G to remained genes after trimming upper and lower $x\%$ of the $\{A_g\}$ and $\{M_g\}$, say 5%.
- Compute the weighted mean of $\{M_g^{(i,j)}\}_{g \in G}$
- Let $S^{(i,j)}$ be the exponential of this mean.
- Adjust $\hat{\mu}_{jg}$ by a factor of $S^{(i,j)}$ for all genes g in library j.
- This means that library i is used as the reference.

TMM method corrects for RNA composition bias

TMM (Trimmed Mean of M Values) by Robinson (2010)

- Many normalization methods perform poorly on samples with extreme composition bias (e.g., in one sample a large number of reads comes from rRNAs while in another they are removed more efficiently).
- Most scaling-based methods, including RPKM and CPM, will underestimate the expression of lowly-expressed genes in the presence of extremely abundant RNAs.
- The TMM methods tries to correct for this kind of bias.
- Method implemented in `edgeR` (Robinson, 2010).

Median log deviation normalization (DESeq)

An alternative approach for cross-sample normalization:

- For each gene g in sample i, calculate deviation of $\log r_{ig}$ from the mean $\log r_{ig}$ over all libraries: $d_{ig} = \log r_{ig} - \sum_i \log r_{ig}/I$.
- Calculate median over all genes: $\log S^{(i)} = \text{median}_g(d_{ig})$.
- Adjust $\hat{\mu}_{ig}$ by a factor of $S^{(i)}$ for all gene g.

edgeR and DESeq are both robust across genes (weighted mean of core gene set vs. median of all genes).

Normalization: Bring-home messages

- Both $S^{(i,j)}$ and $S_{(i)}$ are used for library-level normalization. However, TMM is inferred from normalized counts (r_{ig}/N_i), while MLD is inferred from raw counts ($r_i g$).
- In other words, we have only account for factors that affect all genes in a library similarly.
- However, there are factors affecting different genes differently.
- Recall normalization equation:

$$\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}$$

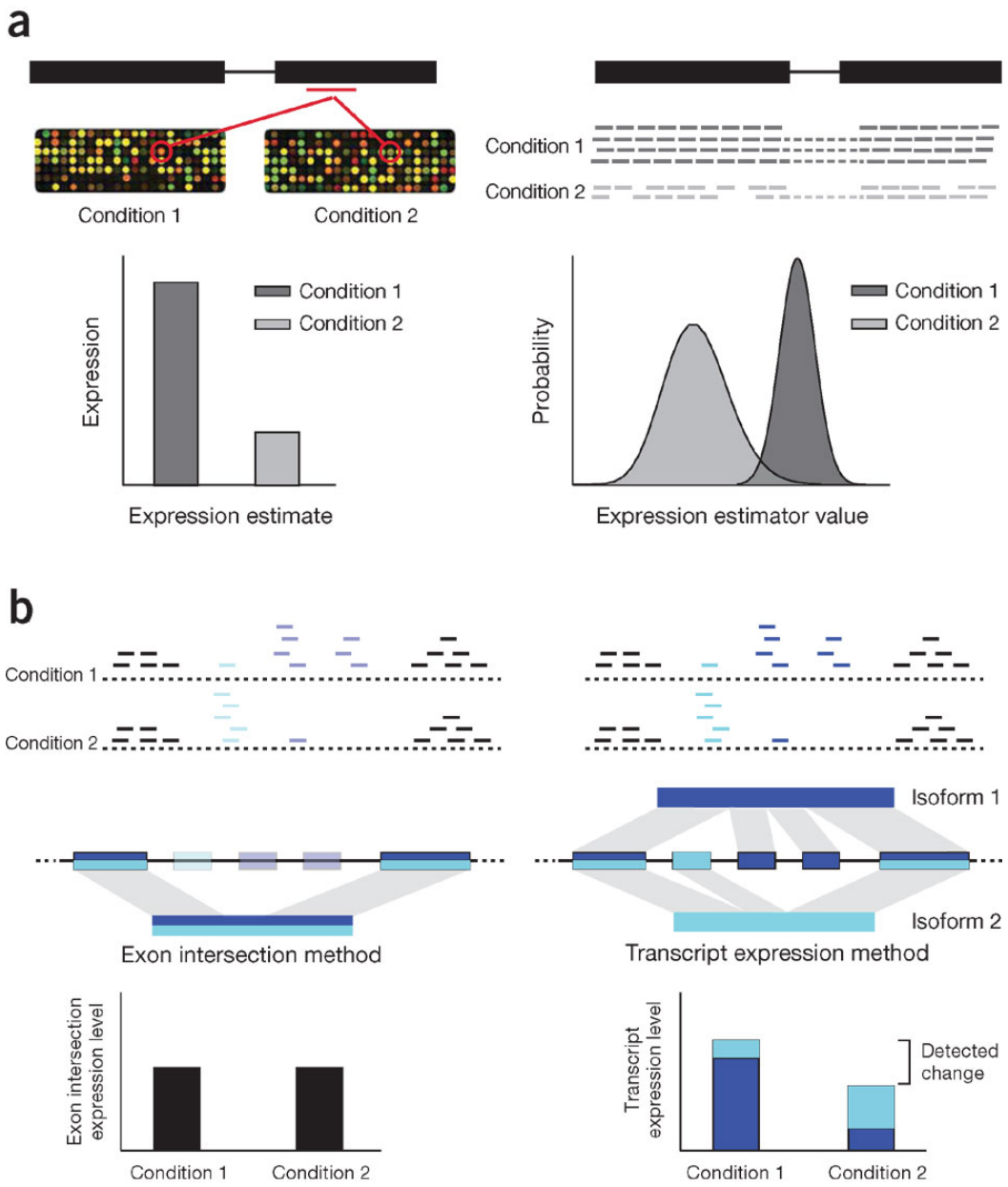
- Consider the decomposition of $k_{ig} = k k_i k_g$:
  - k: global scaling to get more convenient units, e.g., 10^{-9}.
  - k_i: library-specific normalization factors, e.g., $\tilde{N}_i = N_i/S^{(i)}$.
  - k_g: gene-specific normalization factors, e.g., ℓ_g (gene lengths).

Outline

Differential expression analysis (DEA)

- Parametric approaches
  - Poisson - approximation of binomial
  - Negative binomial (NB) distribution
- Nonparametric approaches
  - Rank-based approach
  - Permutation-based approach

Differential expression analysis



Analysis of Differentlly Expressed Genes (DEGs)

- The count data is discrete, and right-skewed. Therefore, log-normal model is not appropriate.
- The biological replicates are typically few. Thus we can not apply the rank-based or permutation-based methods.
- Sequencing depth varies among samples. Hence, normalization should be conducted to make them comparable.

Models of RNA-Seq count data

Let $t = 1, \ldots, T$ be the set of transcripts in the sample i. For each transcript t in sample i, use ℓ_t to denote its length and ρ_{ti} its original relative abundance. Thus we have number of different reads $\tilde{\ell}_t = \ell_t - m + 1$ for transcript t, where m is the length of read. Hence, the probability that a read comes from some transcript t in sample i, can be formulated as

$$\pi_{ti} = \frac{\rho_{ti}\tilde{\ell}_t}{\sum_{r=1}^{T} \rho_{ri}\tilde{l}_r}.$$

The sequencing process can be modeled as a simple random sampling. Hence the number of reads originated from t, namely the counts, can be modeled as

$$N_{ti} \sim \mathrm{Bin}(R_i, \pi_{ti})$$

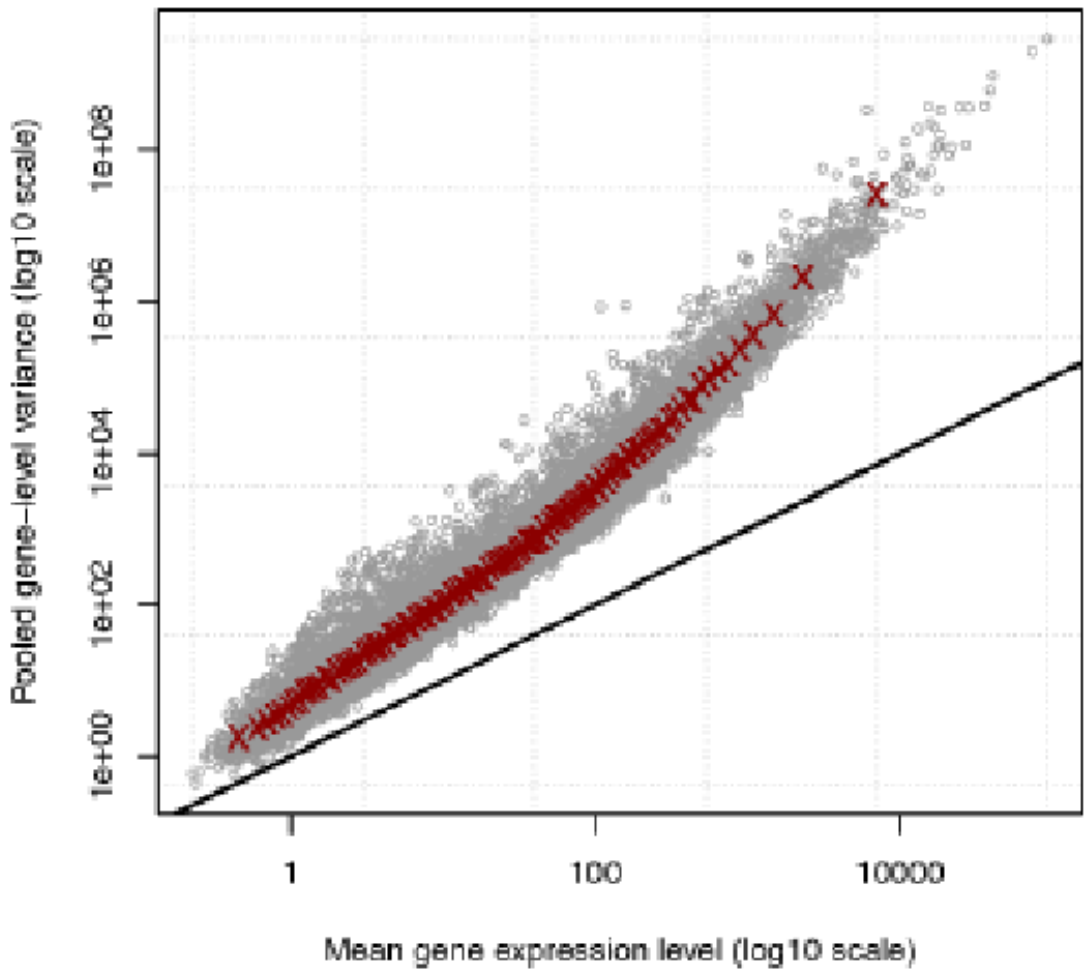
As R_i is very large and π_{ti} is approximately 0, this distribution can be approximated by a Poisson distribution with $\lambda_{ti} = R_i\pi_{ti}$:

$$N_{ti} \sim \mathrm{Poisson}(\lambda_{ti})$$

Overdispersion (1)

- There exists biological/measurement variation between samples/libraries.
- This will increase the variance of the total count dispersion.
- Failure to recognize this, will lead to underestimation of the variance.
- Underestimation of variance, will lead to many false positives.

Overdispersion (2)



Modeling overdispersion

- For every sample: $N_i \sim \mathrm{Bin}(n, p_i)$
- Equivalent (n large, p_i small): $N_i \sim \mathrm{Pois}(np_i)$
- Count variation: N_i/n varies around p_i
- Biological variation: p_i is different for each sample.
- Assumption: p_is are drawn from a population distribution.
- Overdispersion: Biological variation increases the dispersion (=variance).
- Test: if p_i differ on average between the groups.

solution: Gamma-Poisson or negative Binomial distribution, where the Poisson rate parameter is a mixture of gamma random variables with fixed coefficient of variation (绯绘).

Negative Binomial Distribution

The λ_{ti} in Poisson model corresponds to both the mean μ_t and the variance of the distribution, which captures the variation between technical replicates (measurement error due to technologies), but fails to capture the variation between biological replicates (variation among samples belonging to the same group).
For biological replicates, the variance is larger than the mean, and the count data are "over-dispersed", which can be handled by negative Binomial distribution with mean μ_t and variance depending on the chosen parametrization of $var(\lambda_{ti})$:

$$var(\lambda_{ti}) = \mu_t(1 + \phi\mu_t^{\alpha-1})$$

with $var(\lambda_{ti}) = \phi\mu_t^{\alpha}$.
We often set $\alpha = 2$, it becomes the most popular negative Binomial model of RNA-Seq counts with two parameters ϕ and μ_t:

$$N_{ti} \sim \mathrm{NB}(\mu_t, \phi)$$

Negative Binomial: Summary

$$N_{ti} \sim \mathrm{NB}(\mu_t, \phi)$$

- ϕ is the "overdispersion" parameter to account for the variance that cannot be explained by the Poisson model.
- When $\phi = 0$, the NB model reduces to the Poisson model.
- In summary, NB can be modeled as a Gamma mixture of Poisson distribution:
  - the technical variation is Poisson,
  - the Poisson means differ between biological replicates according to a Gamma distribution.

Software for RNA-Seq DEG analysis

- `edgeR` (Robinson, 2010)
- `DESeq/DESeq2` (Anders, 2010)
- `DEXSeq` (Anders, 2012)
- `limmaVoom`
- `Cuffdiff/Cuffdiff2` (Trapnell, 2013)
- `PoissonSeq`
- `baySeq`
- ...

Count matrix: RNA-seq Expression Data

- Each row denotes the counts for gene $g = 1, \ldots, G$ across samples.
- Each column denotes the counts for sample $i = 1, \ldots, N$ across all the genes.
- We refer to the set of read counts for a sample as the library.
- The total number of reads for a sample is called the library size, N_i.
- The number of reads mapped to gene g for sample i, y_{gi} can be modeled as:

$$E(y_{gi}) = \mu_{gi} = N_i \lambda_{gi}$$

where N_i is the library size for sample i, and λ_{gi} is the expected proportion of reads mapped to gene g in sample i.

- When we compare two groups of individuals, wildtype W, and mutant M, the null hypothesis and alternative hypothesis should be:

$$H_0 : \lambda_g^W = \lambda_g^M, H_1 : \lambda_g^W \neq \lambda_g^M$$

edgeR: Dispersion estimator

- Pearson's (pseudo-likelihood) estimator sets the `average Pearson goodness of fit statistics` to its (asymptotic) expected value, which may under-estimate the dispersion when number of libraries is small.
- Quasi-likelihood estimator sets the `average residual deviance statistic` to its (asymptotic) expected values, which may over-estimate the dispersion when number of libraries is small.
- Cox-Reid estimator maximizes the Cox-Reid adjusted profile likelihood, which can be the best choice for estimating the dispersion.

edgeR: Dispersion estimation

- edgeR estimate the common dispersion parameter by conditioning ϕ_i on the sum of counts and maximizing the common likelihood:

$$\ell_c(\phi) = \sum \ell_i(\phi_i)$$

- edgeR models gene-specific dispersion by abundance and shrinks individual likelihoods towards the common likelihood:

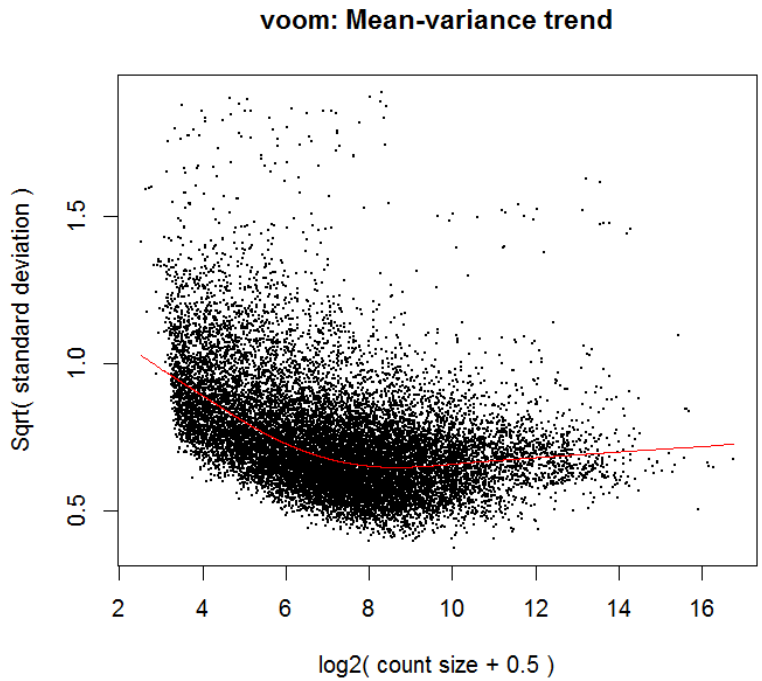
$$WL(\phi_i) = \ell_i(\phi_i) + \alpha \ell_c(\phi_i)$$

edgeR: Differential Expression Testing

- Two-sample test: Exact test replacing hypergeometric probabilites with NB-derived probabilities
- Multifactorial test: Generalized linear models (GLMs) with log-likelihood ratio test (LLRT).

limma as an alternative

- Rationale: if n is large, biological variance dominates the count variation.
- Treat the data the same as continuous (microarray) data by log-transform: $\log_2(N + 1/2)$
- Correct by lowess to suppress mean-variance relationship.



- Testing are done in linear model.

DEA: Bring-home messages

- Under the circumstance of small sample size, use the parametric models (either zero-inflated Poission or negative Binomial models (preferred)).
- If the sample size is large enough, apply the nonparametric methods (rank-based or permutation-based) to the data.
- Multiple testing correction should be applied for the selection of significant genes (Benjamini-Hochberg (BH) correction, i.e. FDR).

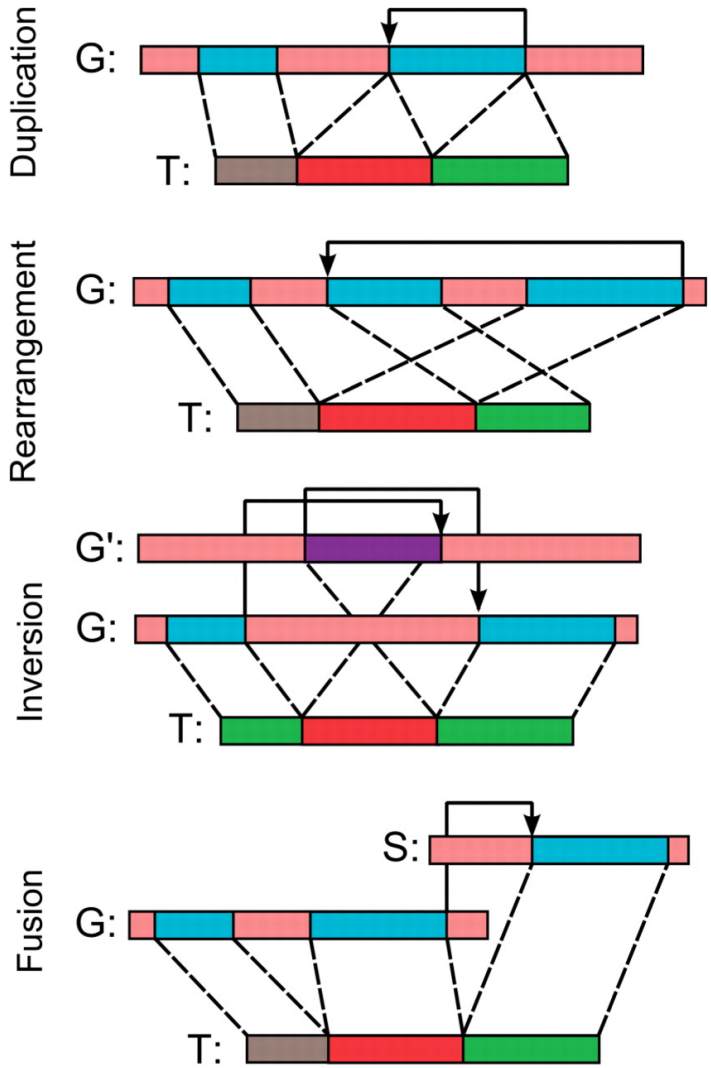
Outline

1. Transcriptomes
   - RNAs and alternative splicing
   - Low-throughput & high-throughput technologies
2. RNA-Seq technology
   - Library preparation techniques
   - Sequencing technology
3. RNA-Seq data analysis
   - Short read mapping
   - Expression quantification
   - Expression normalization
   - Differential expression analysis (DEAs)
4. Other applications of RNA-Seq

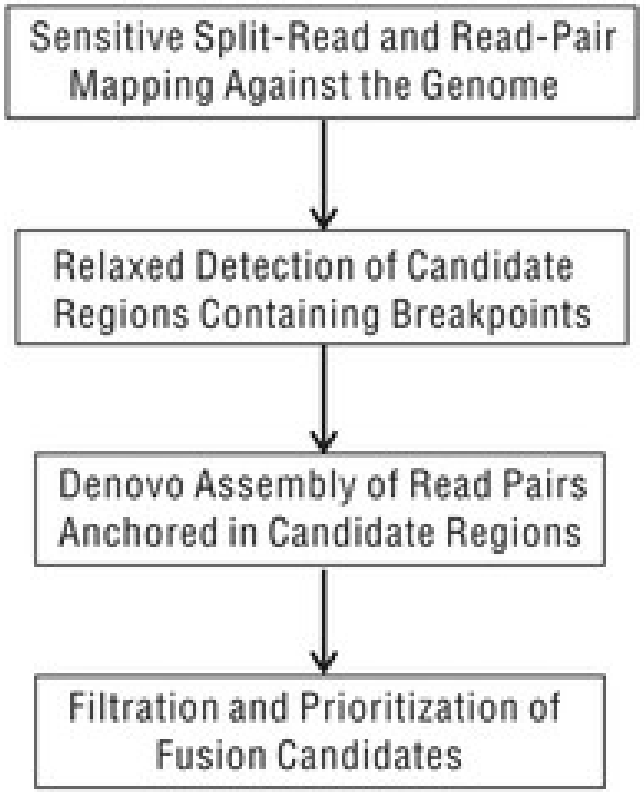
Allele-specific expression

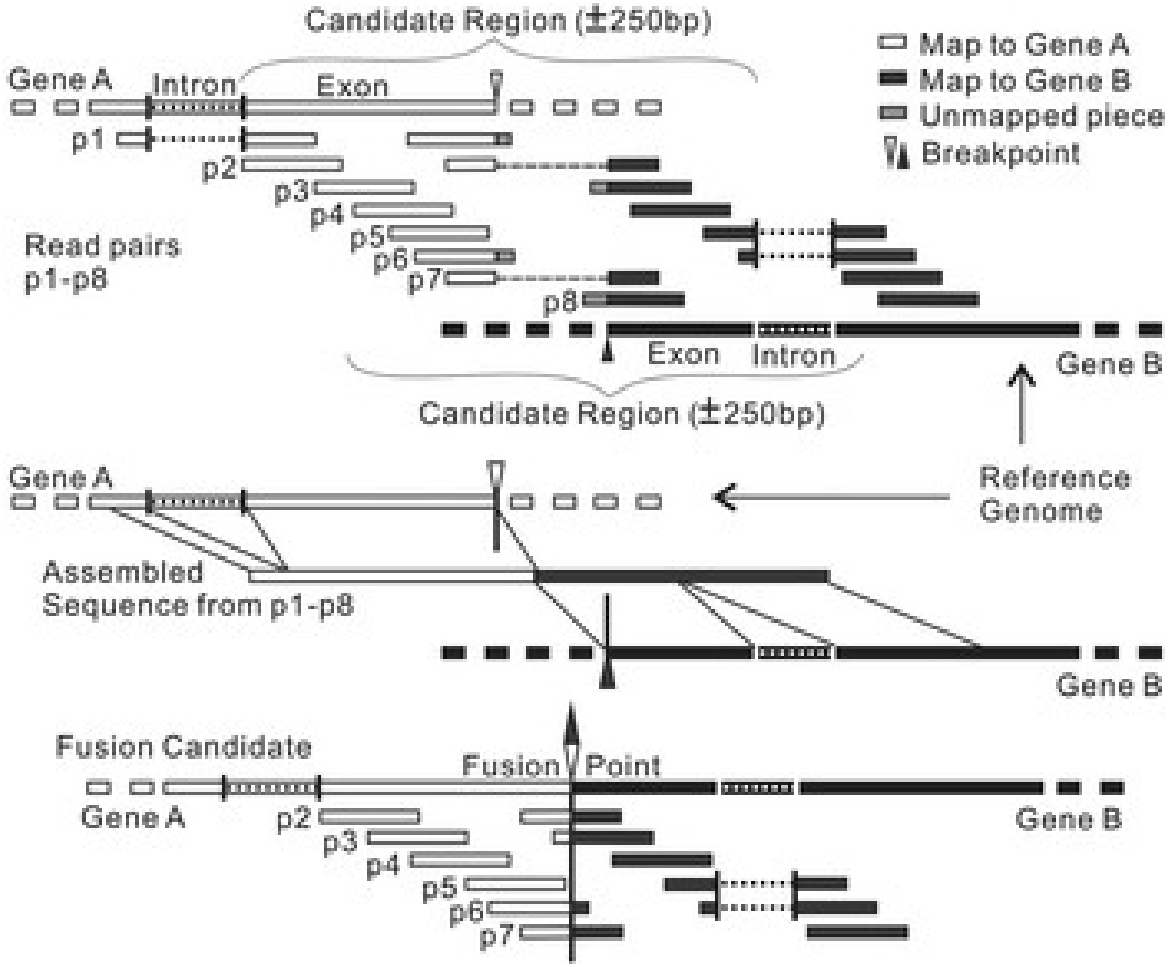
- Gene expression is a complex trait that is influenced by
  - *cis*- and *trans*-acting genetic and epigenetic variation
  - Environmental factors
- In principle, alle-specific approach can eliminate environmental or *trans*-acting effects
  - Two alleles serve as within-sample control

Structural variants discovery



Gene-fusion detection

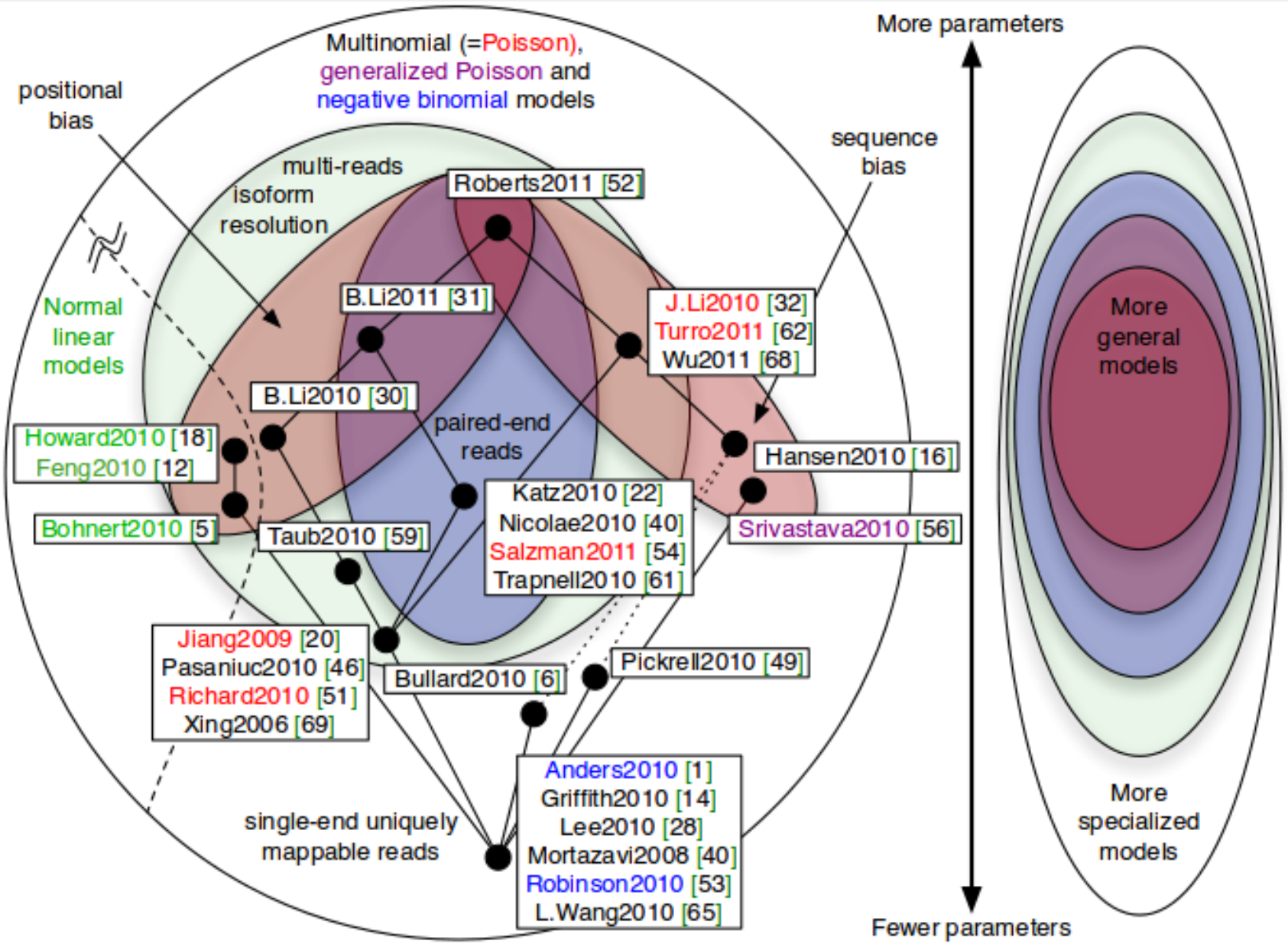




Future Developments

- Existing computational tools need to evolve to meet new demands for improved sequencing technologies.
  - Longer reads
- Specialized RNA-seq libraries are used for annotating the $5'$ start site and $3'$ ends of transcripts.
- The ongoing cycle of improvements in technology, both laboratorily as well as compuatationally, will continue to expand the possibilities of RNA-seq, making this technology applicable to an increasing variety of biological problems.

Models for RNA-Seq Data



Questions?