RNA sequencing (RNA-Seq) part 2: expression level estimation and differential expression analysis

Vincent Plagnol

UCL Genetics Institute

Outline

1. Estimating gene level and normalisation issues
2. Obtaining isoform specific expression estimates
3. Differential expression
   - What does "differential expression" mean?
   - Using read count to test for differential expression
   - Exon level analysis using DEXSeq
   - Integrated isoform discovery and differential expression

Tools for differential expression analysis

- DEseq: differential expression analysis from RNAseq data at the gene level.
  - Key feature: this is a R package part of bioconductor.
- DEXseq: differential expression analysis from RNAseq data at the exon level.
  - Bioconductor package
  - Detecting differential usage of exons from RNA-Seq data.
- Cufflinks: Transcript assembly, differential expression, and differential regulation for RNA-Seq
  - Cufflinks web page
  - Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Trapnell et al, Nature Protocols 2012

Additional reference papers

- Mapping and quantifying mammalian transcriptomes by RNA-Seq, Mortazavi et al, 2008
- Understanding mechanisms underlying human gene expression variation with RNA sequencing, PIckrell et al, Nature 2010
- Noisy Splicing Drives mRNA Isoform Diversity in Human Cells, Pickrell et al, 2010
- A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Dillies et al, Briefing in Bioinformatics, 2012

Our starting point: a table of counts

- I assume that we know how to map reads, i.e. generate a matrix of counts that look like

	Sample 1	Sample 2	...	Sample N
Feature 1	r_{11}	r_{21}	...	r_{N1}
Feature 2	r_{12}	r_{22}	...	r_{N2}
...				
Feature K	r_{1K}	r_{2K}	...	r_{NK}

- These features can be genes, but they can also be isoforms, or even paternal/maternal genes.
- It all depends on the mappin strategy that was used.

Outline

The simplest question: how much of that gene is present in my sample?

- This question is already a challenge, much more difficult than it appears.
- All libraries are different, and comparison between samples is not obvious.
- **How do we ensure our estimates are comparable across features and libraries and are on a standard scale?**
- Given the number of reads r_{ij} mapping to a feature j for sample i, we are looking for a normalization factor ρ_{ij} that makes the quantity r_{ij}/ρ_{ij} meaningful.

Strategy 1: normalizing using reference/housekeeping genes

- This is a strategy often used by scientists running qPCR experiments.
  - Take your favorite highly expressed gene k and use $r_{ik} = \rho_{ij}$.
  - This certainly controls for the depth of sequencing.
- An obvious issue is that the r_{ik} quantity may vary between samples, making it a poor reference.
- Also this allows comparison between samples but not between genes.
  - Why? Because transcripts that are twice longer will have twice the number of reads for the same level of expression.
- It looks very flawed. But can you see a situation where this is actually useful? (without looking at the next slide!)

But strategy 1 can in fact be very useful

- There is at least one situation where this approach is used routinely and everybody is (and should be) happy with it.
- Take 2 genes, with 2 different isoforms.
  - We like to say that isoform 1 represents x% of the expression of isoform 2.
  - This is in fact exactly the proposed strategy 1.
- In that situation, another isoform of the same gene is probably a very good normalization factor.

Strategy 2: the RPKM approach

- What factors influence the number of reads mapping to a gene/exon? The first one is the library size:
  - If I generate 10 times more reads, I expect 10 times more reads mapping to my sequence.
- The second obvious factor is feature length, as discussed above:
  - If the sequence is twice longer with the same level of expression, then twice more reads will map to it.
- This suggests the following normalization:

$$\mu_{ij} = \frac{r_{ij}}{l_j \times N_i}$$

- Usually l_j is in kb and N_i in million of reads, we are dealing with Reads per Kb and per Million reads, hence RPKM.
- Of note, for paired end data, we care more about fragment count than read count, hence the use of FPKM (same thing, but by fragment).

Strategy 2: the RPKM approach

- RPKM are very useful, and are widely used.
- This is a rather obvious idea, which was first used this paper.
- A question is what does the total number of reads N_i really mean?
  1. Is this really the total number of reads generated (most likely post QC)?
  2. Or the number of reads mapping to all exons genome-wide?
- Option 2 is probably more useful
  - For example, if one use option 1 and compares a poly(A) prep with a total RNA prep, RPKMs are necessarily much lower owing to the vast amount of reads not mapping to exons.

Making the RPKM strategy fail

- Take 2 libraries from two very different tissues, but same total number of reads.
- Assume that the read count is the same for all genes in library 1 (say count 50).
- In library 2, half of the genes have count 100, the others have 0 read.
- In that situation one may conclude:
  - That there is a 2 fold increase of expression in library 1 for shared genes.
  - Or that the expression is the same for the shared genes.
- Which interpretation is right depends whether the total RNA output is the same.
- If so, RPKM is correct to see a 2 fold increase.

Normalization strategy 3, implemented in TMM and edgeR

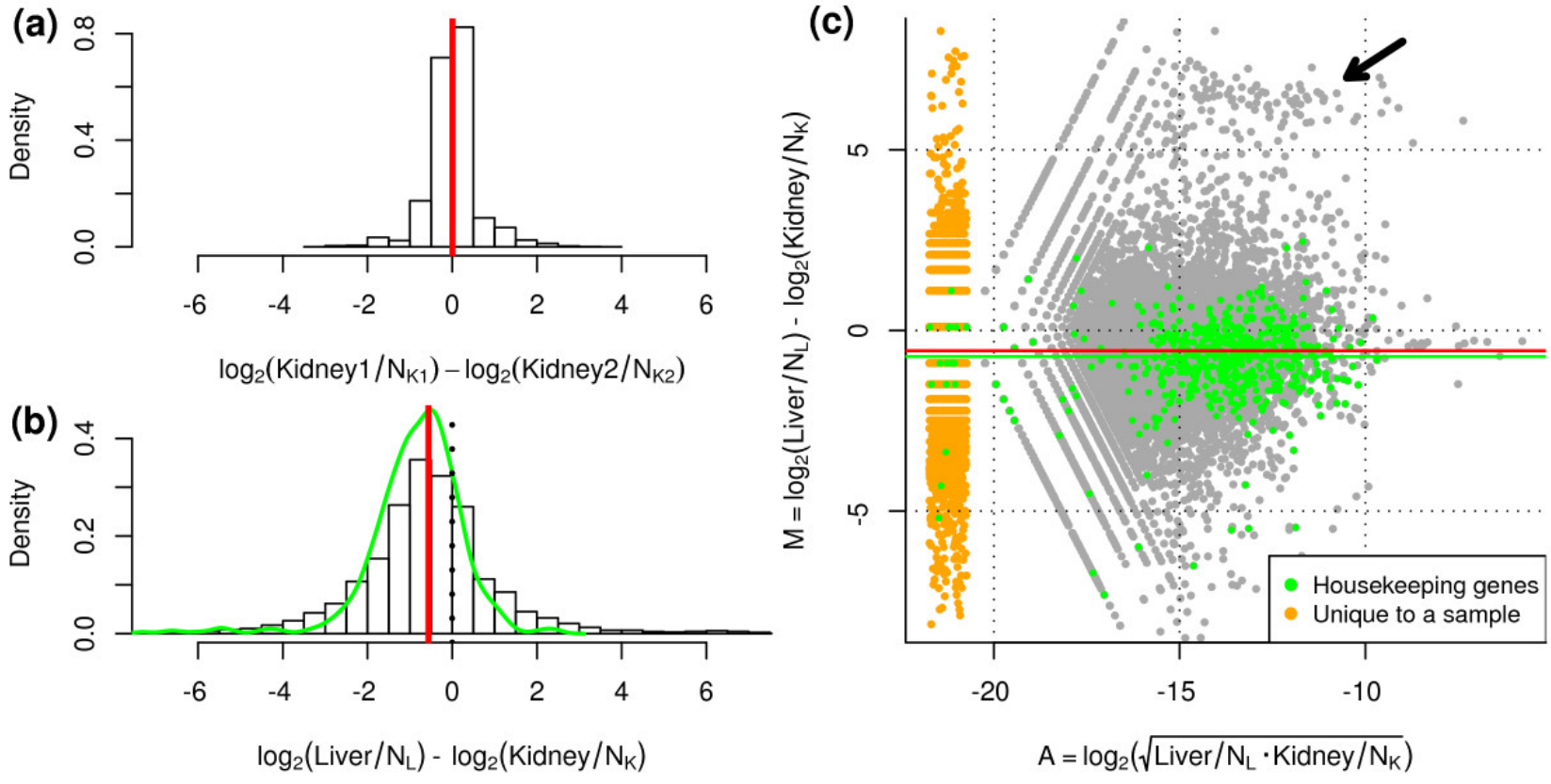
- Identify a subset of unexceptional genes by computing the fold change and total count.

$$M_g^{(i,j)} = \log \frac{r_{ig}}{N_i} - \log \frac{r_{jg}}{N_j}$$

$$A_g^{(i,j)} = \log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j}$$

- Pick the genes after excluding the ones that fall in the x% quantile for either of these measures (the boring ones).
- Normalize based on these.
- This is a reminder of strategy 1, we don't use one gene, but we do not use all the genes either.

An application: liver, kidney dataset



Robinson, Oshlack, Genome Biology 2010

Normalization strategy 4, implemented in DESeq

- The idea is similar, look at all genes and see to what extent the read count differ from the average:

$$d_{ig} = \log r_{ig} - \mathrm{average} \log r_{ig}$$

- Take the median of this quantity, then bring it back to the natural scale:

$$S_i = exp(\mathrm{median} d_{ig})$$

- S_i is a correction term for the the library size, and instead use

$$\tilde{N}_i = \frac{N_i}{s_i}$$

- This is our adjusted library size, also robust because of the use of the median.

Summary

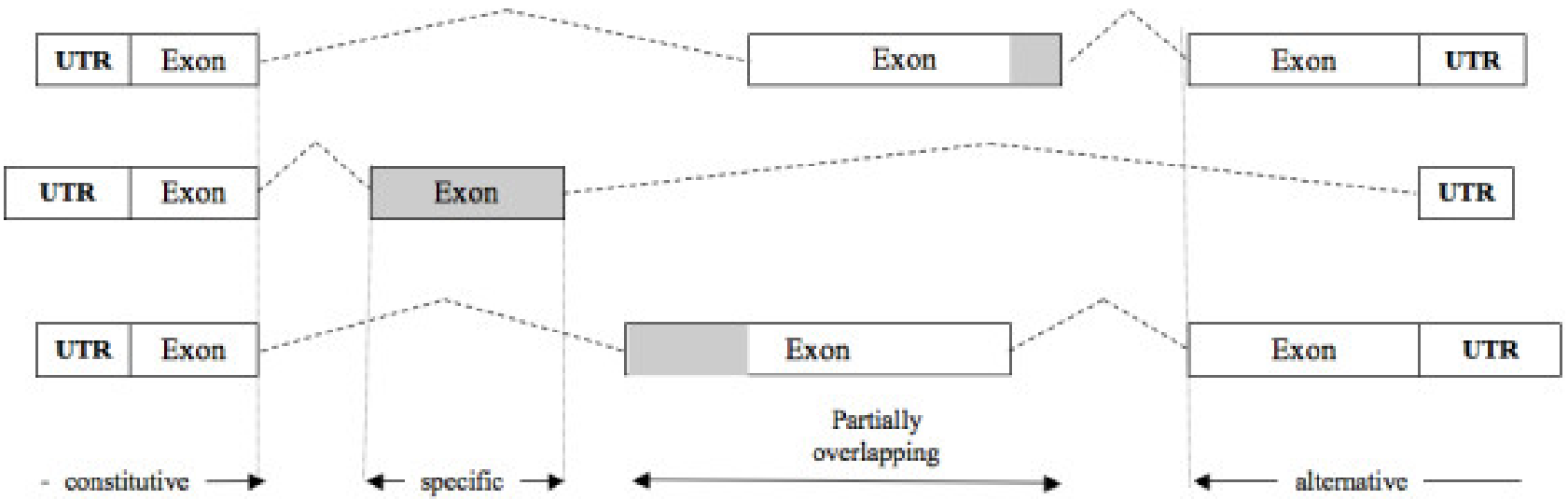
- There is no such thing as one size fits all.
- Different normalizations make different assumptions which can explain differences in outcome of the tests.
- From the Briefing in Bioinformatics paper: "Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis. Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data."

Outline

Outline

Aligning reads to variable isoforms

- Some transcriptome alignment strategy skip the whole genome alignment and map instead to various isoforms.
- Most reads map to all or most isoforms, so this is not a trivial problem.
- Mathematically speaking this is a missing data problem.



Leoni et al, Genome Biology 2011

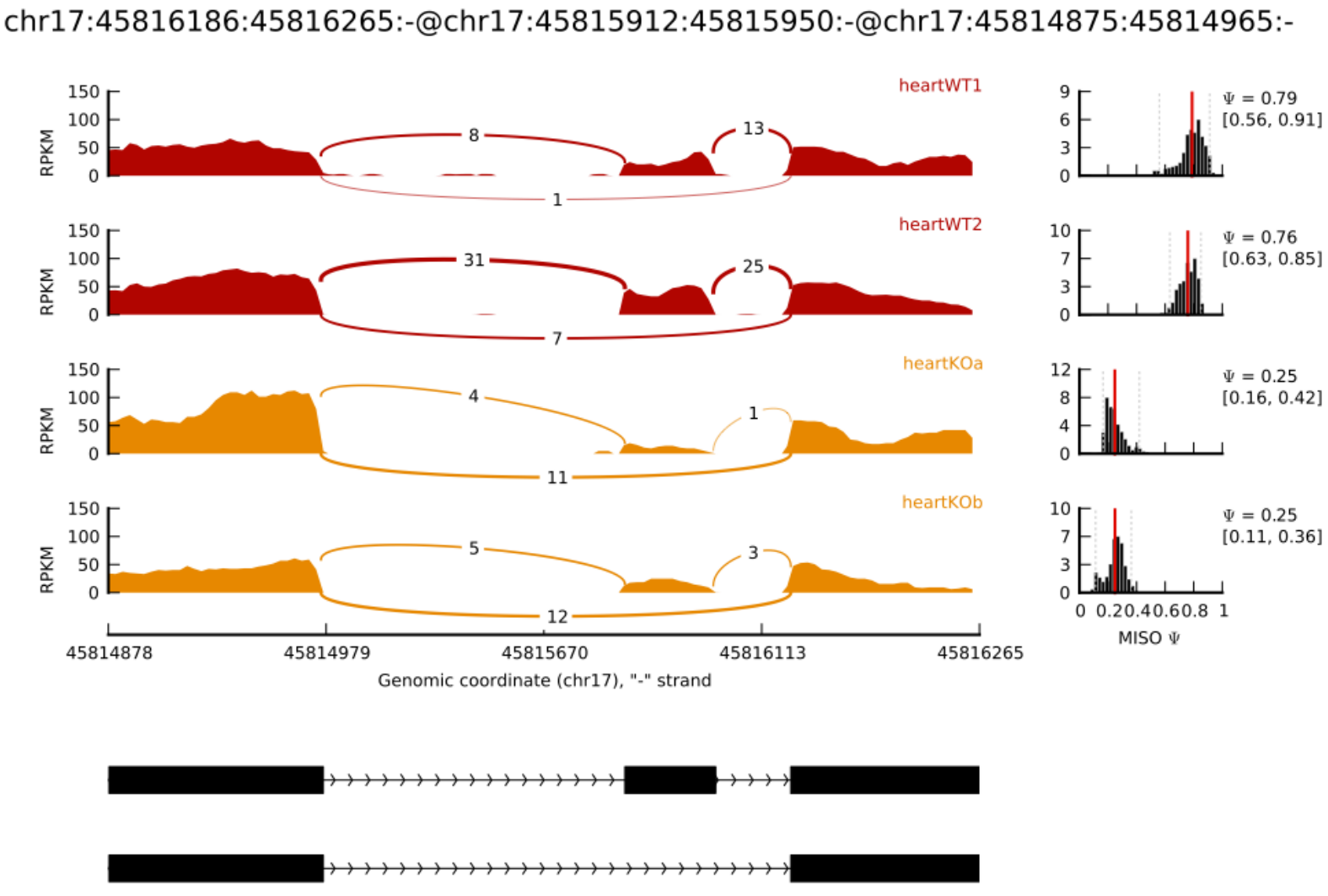
Dealing with genome aligned RNA-Seq data

- If you aligned your reads with tools like `STAR` or `tophat`, you have done the computer science part, but statistics have not yet happened.
- I assume that we know what isoforms exist (no novel isoform assembly at this point).
- One needs to relate the aligned reads to the available isoforms.
- This is a difficult statistical and computational missing data problem.
- It can be dealt with using an expectation-maximization (EM) algorithm or a Bayesian approach.

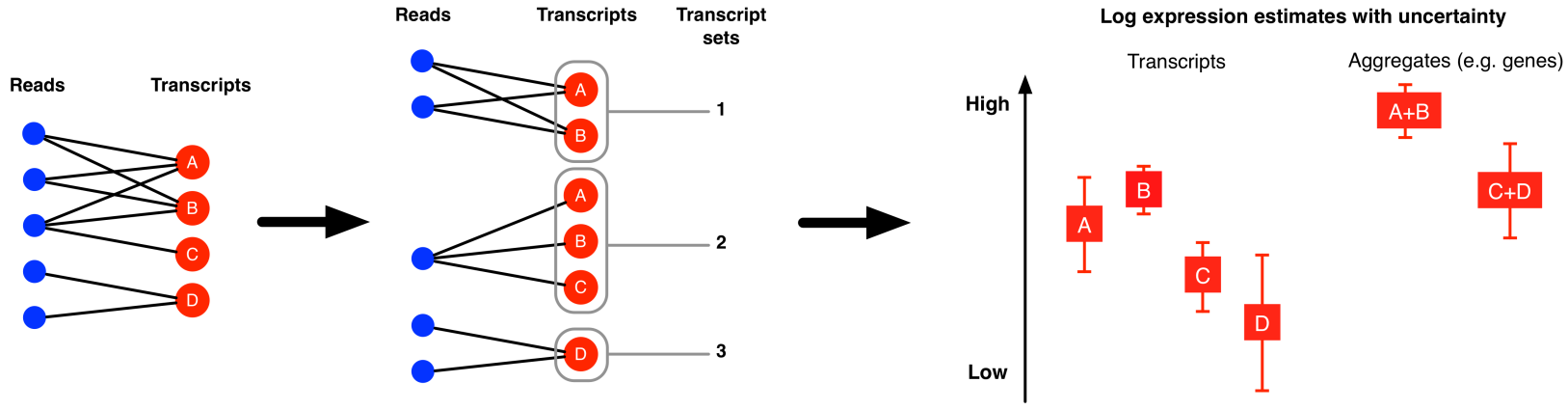
Various tools are available for this

- iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data
- TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference.
- MISO: Probabilistic analysis and design of RNA-Seq experiments for identifying isoform regulation

An illustration with MISO and Sashimi plots

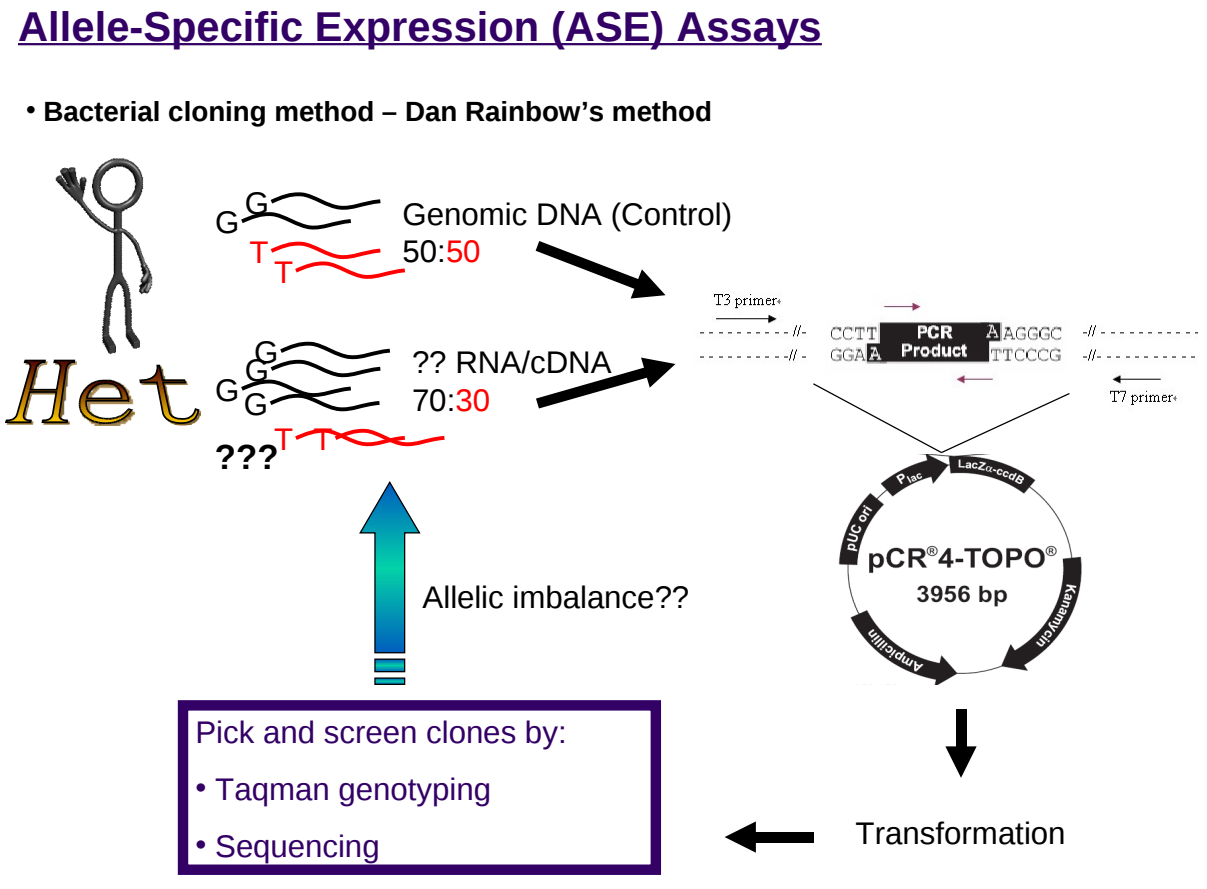


Another useful alternative: MMSEQ



- Expression levels are inferred for each transcript using the mmseq program by modelling mappings of reads or read pairs (fragments) to sets of transcripts.
- These transcripts can be based on reference, custom or haplotype-specific sequences. The latter allows haplotype-specific analysis, which is useful in studies of allelic imbalance.

Allele specific expression



It is very hard to make it work properly.

Outline

Outline

What does differential expression analysis mean?

- In its simplest form, one can take 2 samples and ask whether the expression level varies between the 2 for a gene of interest.
- But this can easily be mistaking when dealing with RNA-Seq data.
- Take, for example, two samples with 10M library depth and RPKM 200 and 210 on a 10 kb transcript.
- The number of mapping reads is 20,000 and 21,000.

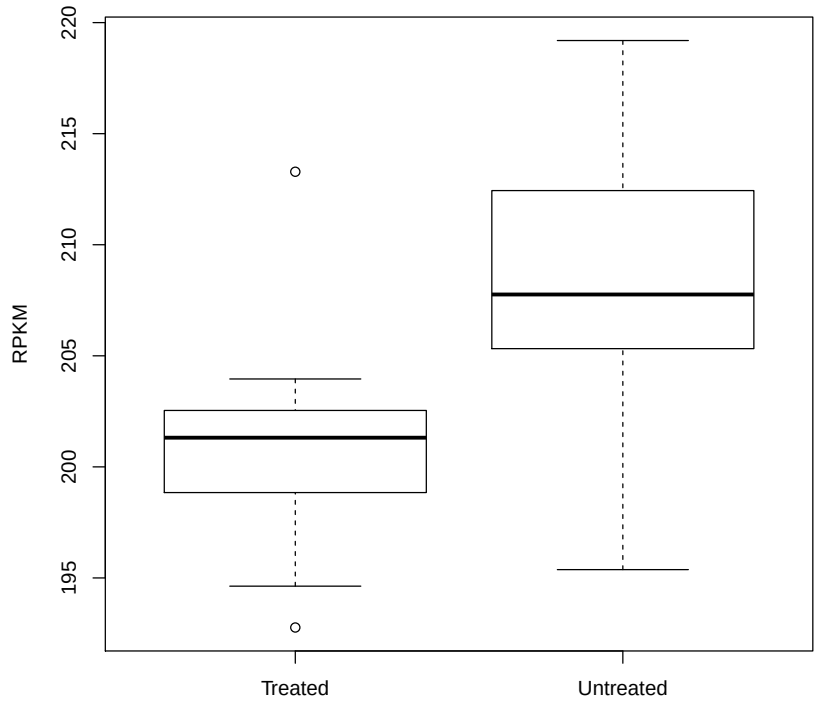
How do we assess the significance of these numbers?

- Using Poisson approximation, we can have either 2 Binomial with different probabilities or a single Binomial with the average probability.
- A likelihood ratio test in R looks like this:

```
ll <- dbinom(size = rep(10^7, 4),
  p = 10*c(200,210,205,205)/10^6,
  x = c(20000, 21000, 20000,21000),
  log = TRUE)
llike <- sum(ll[1:2]) - sum(ll2[3:4])
p.val <- pchisq(q = 2*llike,
   df = 1, lower.tail = FALSE)
```

- The resulting P-value is quite significant ($p = 10^{-7}$), which seems in contrast with RPKM varying between 200 and 210 (5% reduction).

What does differential expression analysis mean?



- Three replicates per category is minimum number of replicates you want to be dealing with.

Outline

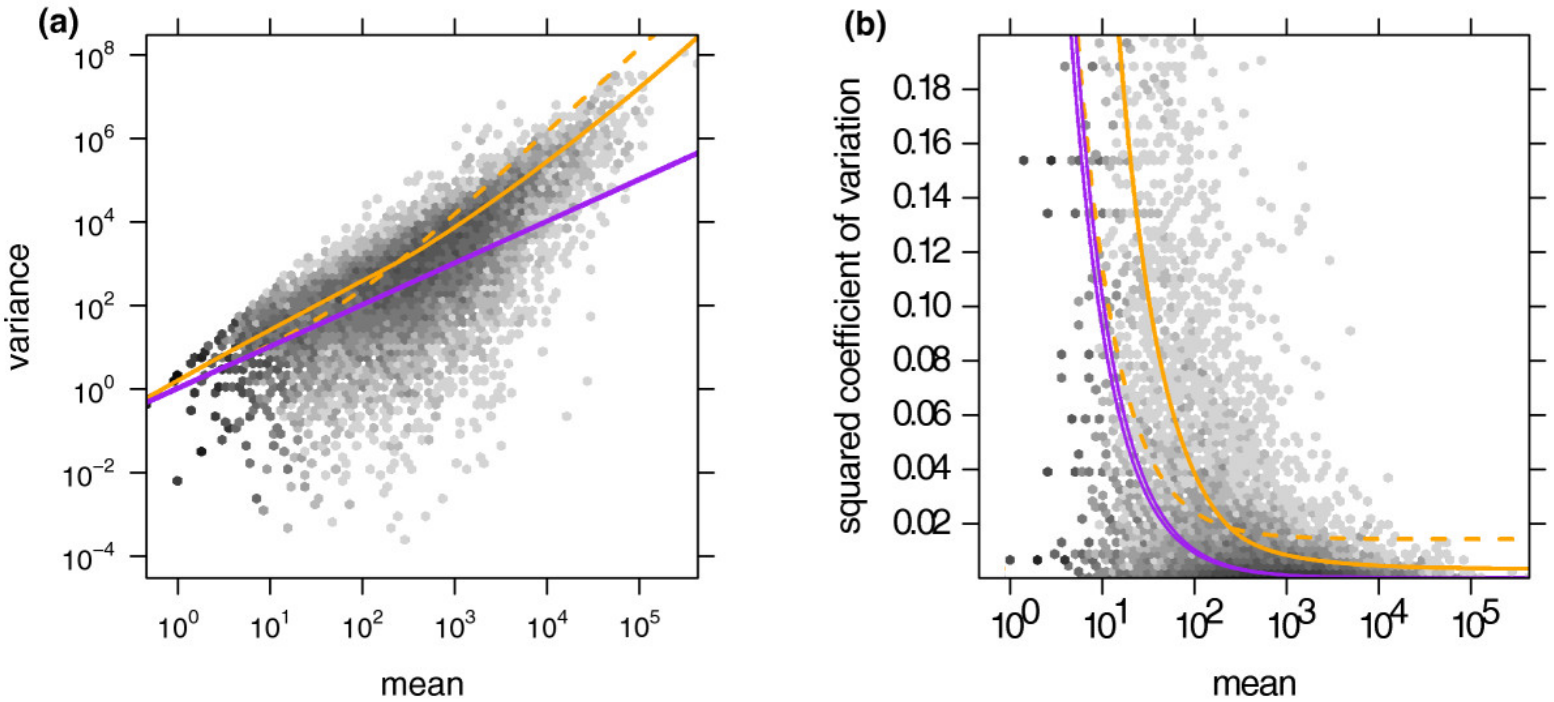
What to do with isoform specific estimates

- If you used a "mapping to transcriptome" strategy, you should have generated point estimates with confidence interval.
- Tools such as edgeR and DESeq do not incorporate the variance due to read mapping uncertainty.
- Because read mapping uncertainty is key across isoforms, these tools are not ideal for DE detection in such conditions.
- EBSeq, an empirical Bayesian DE analysis tool, does some version of this but I have not tried to use it.

The statistics behind `DESeq`

- Generate a table of counts by parsing the BAM file, combined with gene definitions based on a GTF file.
- Normalize the count data using the procedure described above.
- Fit a statistical model to the read count that incorporate the extra variance compared to a simple binomial model.
- Perform a differential testing approach based on the fit of the model and the data, and return a P-value per gene.

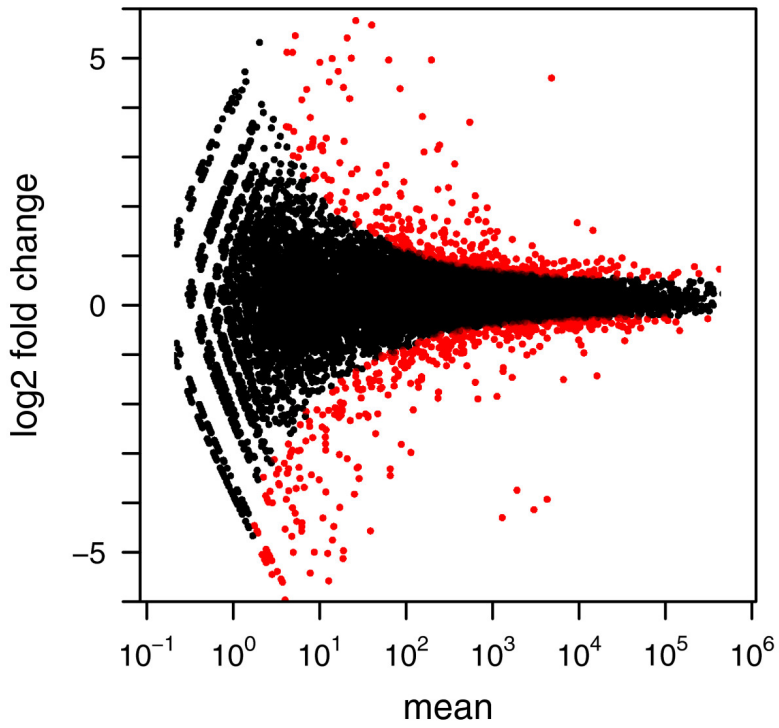
Overdispersion in count data



Dependence of the variance on the mean for condition A in the fly RNA-Seq data. (a) The scatter plot shows the common-scale sample variances plotted against the common-scale means. The purple lines show the variance implied by the Poisson distribution for each of the two samples, (b) Same data as in (a), with the y-axis rescaled.

Andes and Huber, Genome Biology 2010

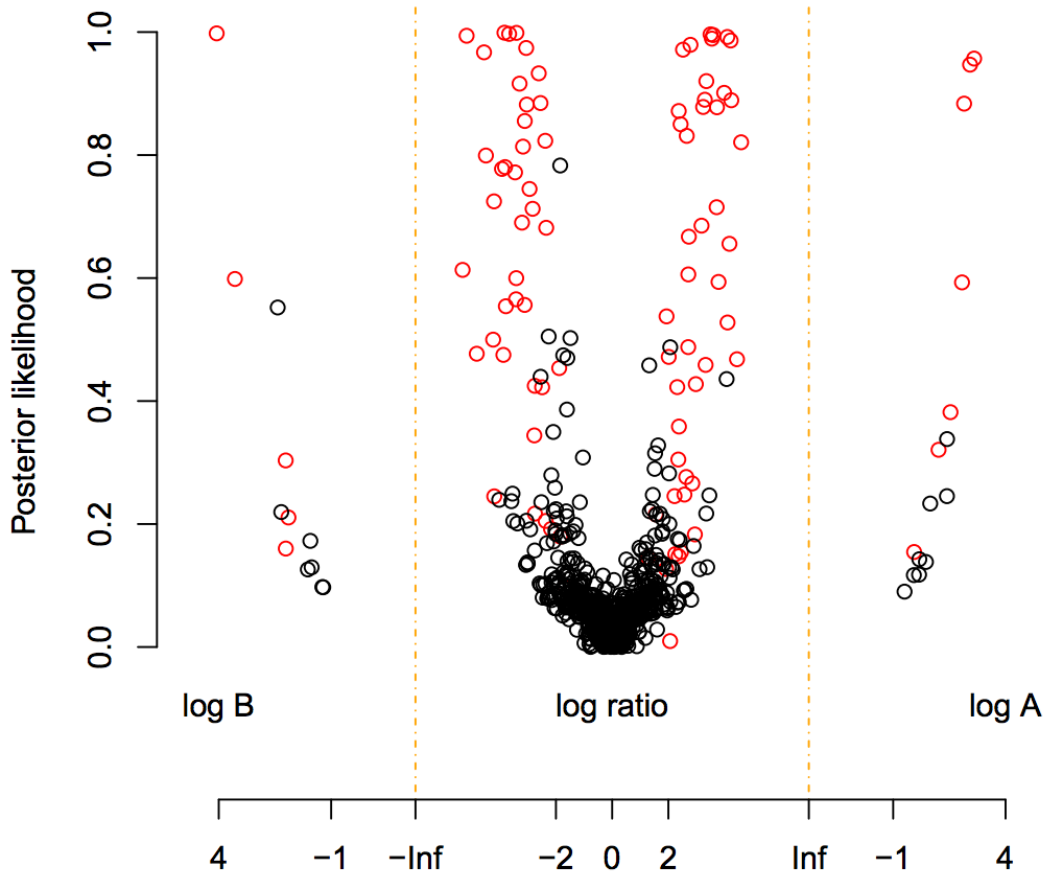
Standard MA plot for DE analysis



Testing for differential expression between conditions A and B: Scatter plot of log2 ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used.

Andes and Huber, Genome Biology 2010

If you do not fancy P-values, BaySeq is available



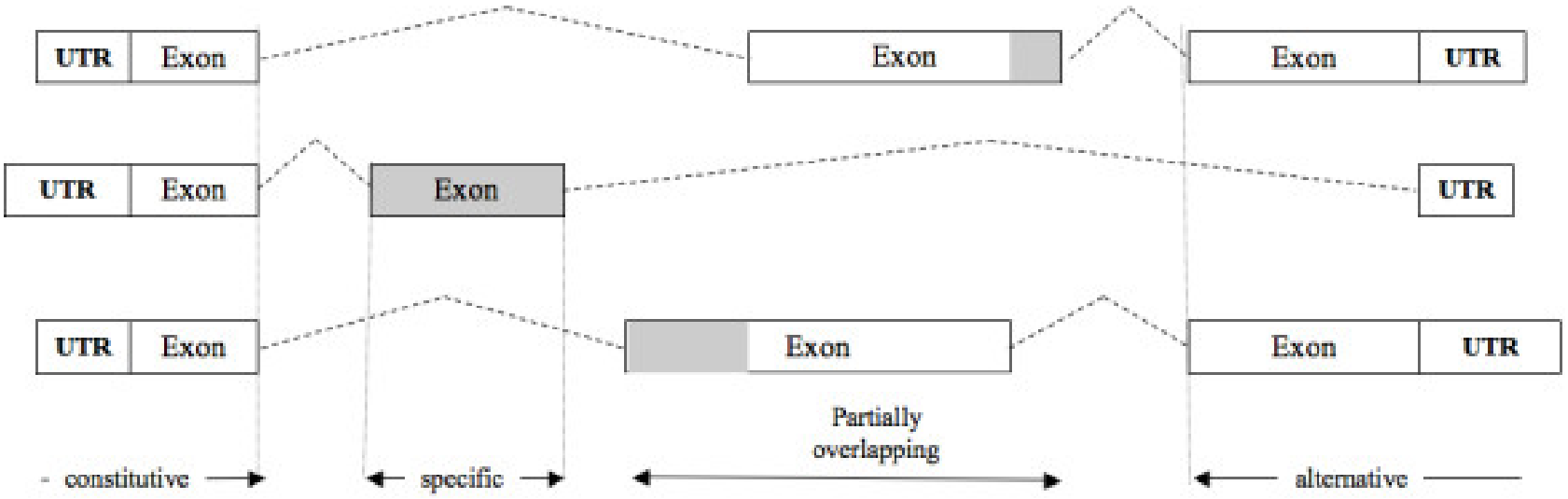
from BaySeq vignette

Outline

The statistics behind `DEXSeq`

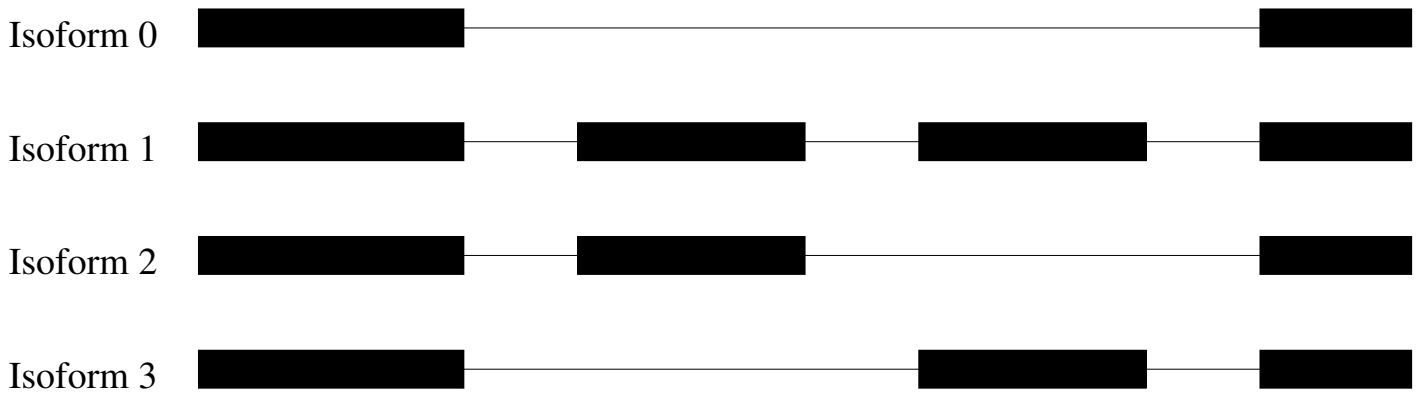
- DEXSeq takes a different approach from isoform level estimation.
- It cuts each gene into chunks and tests whether, in the "treated" group, the level of expression differs from the "untreated" group.
- This is closely related to the isoform level estimation problem, but statistically the treatment is quite different.
- Like DESeq it is a frequentist approach, i.e. returns P-values.

This approach is relatively simple to describe



Leoni et al, Genome Biology 2011

But it may fail in some rare and unusual cases



Think of group 1 with 50% isoform 0 plus 50% isoform 1 and group 2 with 50% each of isoform 2 and 3.

The statistics behind DEXSeq

- The basic model, which assumes no differential effect, looks like this:

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S$$

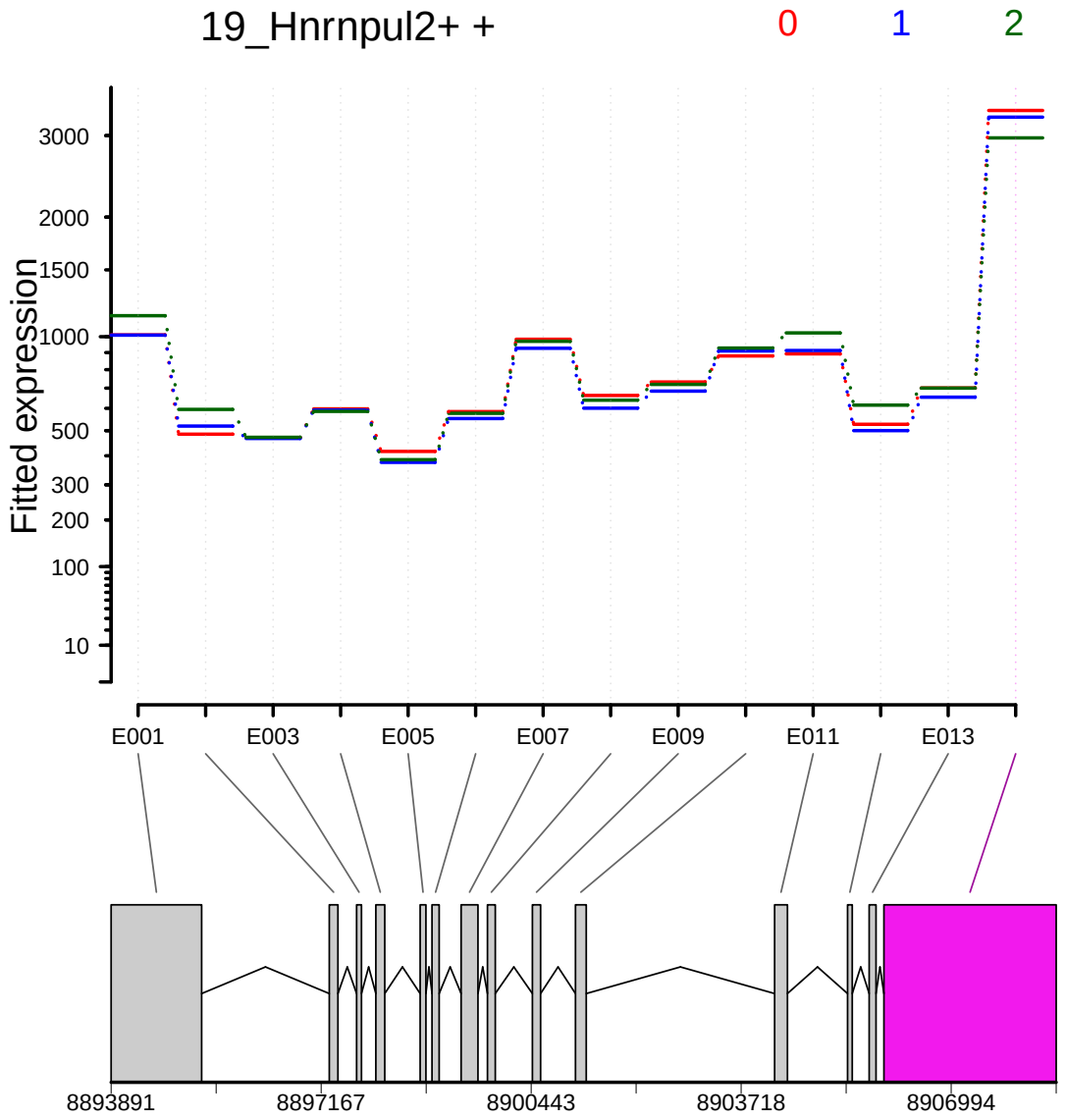
  where i refers to the gene, l refers to the exon, and j to the sample.

- This model is then compared to a richer model for each exon l':

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S + \beta_{i\rho_j l}^{EC} \delta_{ll'}$$

- In this case, $\delta_{ll'}$ is non-zero only for the exon of interest l' and a different interaction term is estimated for each condition ρ_j.
- The point is to obtain information about splicing, rather than the confounding effect of the overall expression of the gene.

An example of DEXSeq output

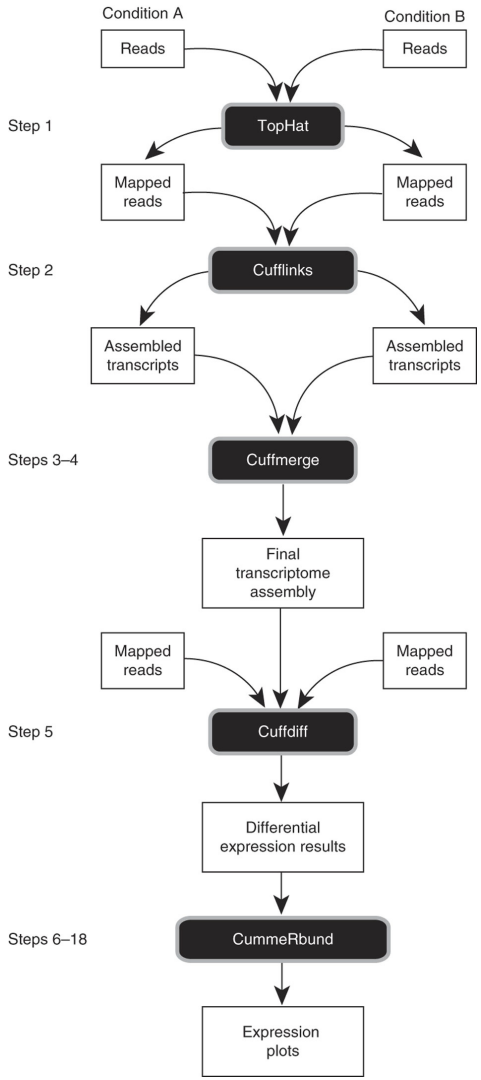


An example of DEXSeq output

- DEXSeq will typically return one P-value per exon.
- This will also include a dispersion factor that measures the noise in the data.
- The first category will be used as a baseline and log2 ratio is shown for the other groups.
- It treats categories as discrete rather than numeric variables.
  - I suppose this is a feature to include in future releases.
- It uses multi-core libraries in R which is very useful to speed up the process.

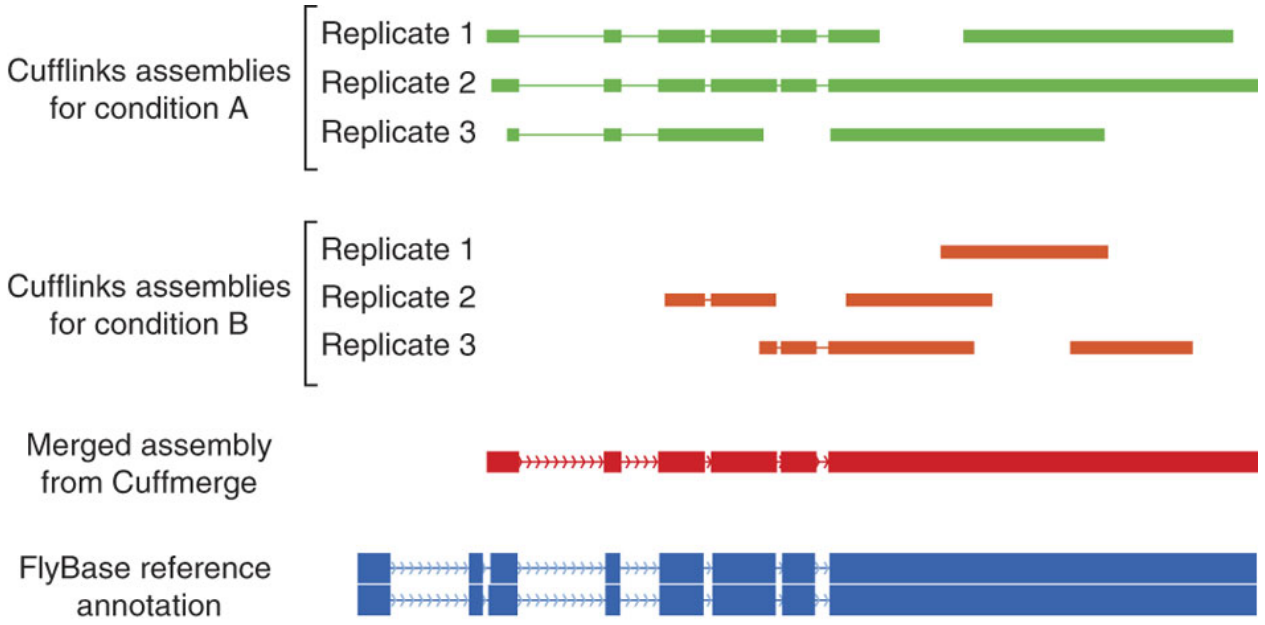
Outline

The integrated tophat/cuffdiff/cuffmerge tools



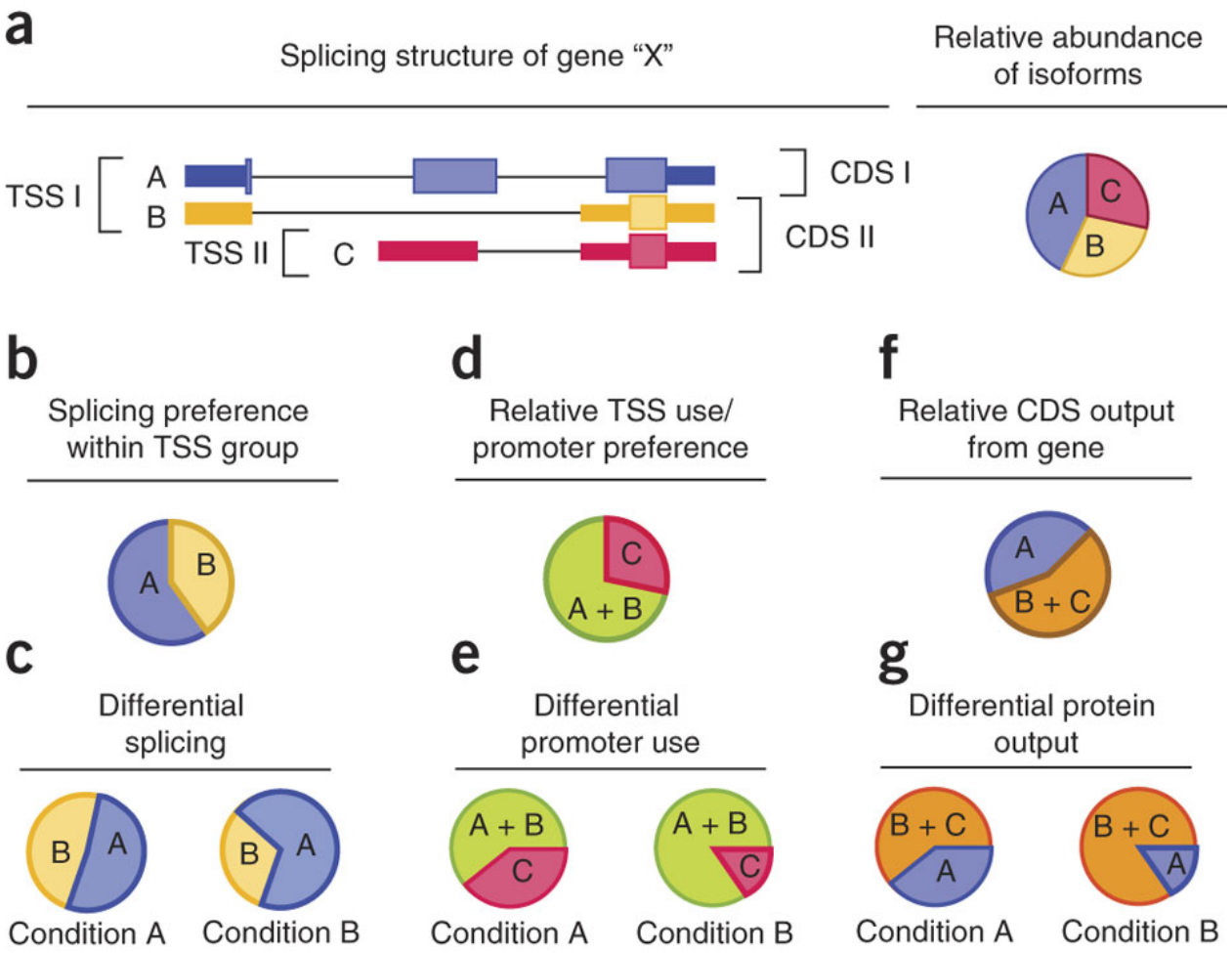
Trapnell et al, Nature Protocols 2012

Is cuffmerge useful?



Trapnell et al, Nature Protocols 2012

More sophisticated differential expression testing



Trapnell et al, Nature Protocols 2012

Overall Summary

- Gene based expression analysis is largely sorted, generates very clean and usable data with RNA-Seq.
- Differential expression analysis using isoform level data is another level of complexity.
  - We will probably need longer reads to properly sort that problem.
- I recommend spending time to reads the following blog post by Stephen Turner.
- DESeq and DEXSeq have so far proven to be the most reliable tools in my hands.