General considerations for RNA-seq quantification for differential expression

or how to count.

Ian Dworkin
@IanDworkin

Overview of lecture and tutorial today

- Today
  - The absolute basics of experimental design.
  - Overview of RNAseq pipelines.
  - What features should we count (genes, transcripts, exons, kmers).
  - How to count (reads).
  - Dealing with multiple comparisons
  - Avoiding “accidental” p-hacking (and why it can be so insidious in genomic analysis because of variations on pipelines for mapping, types of features, how to count and DE seq approaches).
- Tutorial
  - Mapping reads with a splice aware aligner (tophat)
  - Counting with a simple tool (HTSeq)

Caveats

- There are whole courses on proper experimental design. Great books too.

- For experimental design I highly recommend:
  - Quinn & Keough: Experimental Design and data analysis for biologists.

    http://www.amazon.com/Experimental-Design-Data-Analysis-Biologists/dp/0521009766/

Goals

I am not planning on trying to provide any sort of overview of statistical methods for genomic data. Instead I am going to provide a few short ideas to think about.

Statistics (like bioinformatics) is a rapidly developing area, in particular with respect to genomics. Rarely is it clear what the "right way" to analyze your data is.

Instead I hope to aid you in using some common sense when thinking about your experiments for using high throughput sequencing.

A simple truth:
There is no technology nor statistical wizardry that can save a poorly planned experiment. The only truly failed experiment is a poorly planned one.

To consult the statistician after an experiment is finished is often merely to ask them to conduct a post mortem examination. They) can perhaps say what the experiment died of.

Ronald Fisher

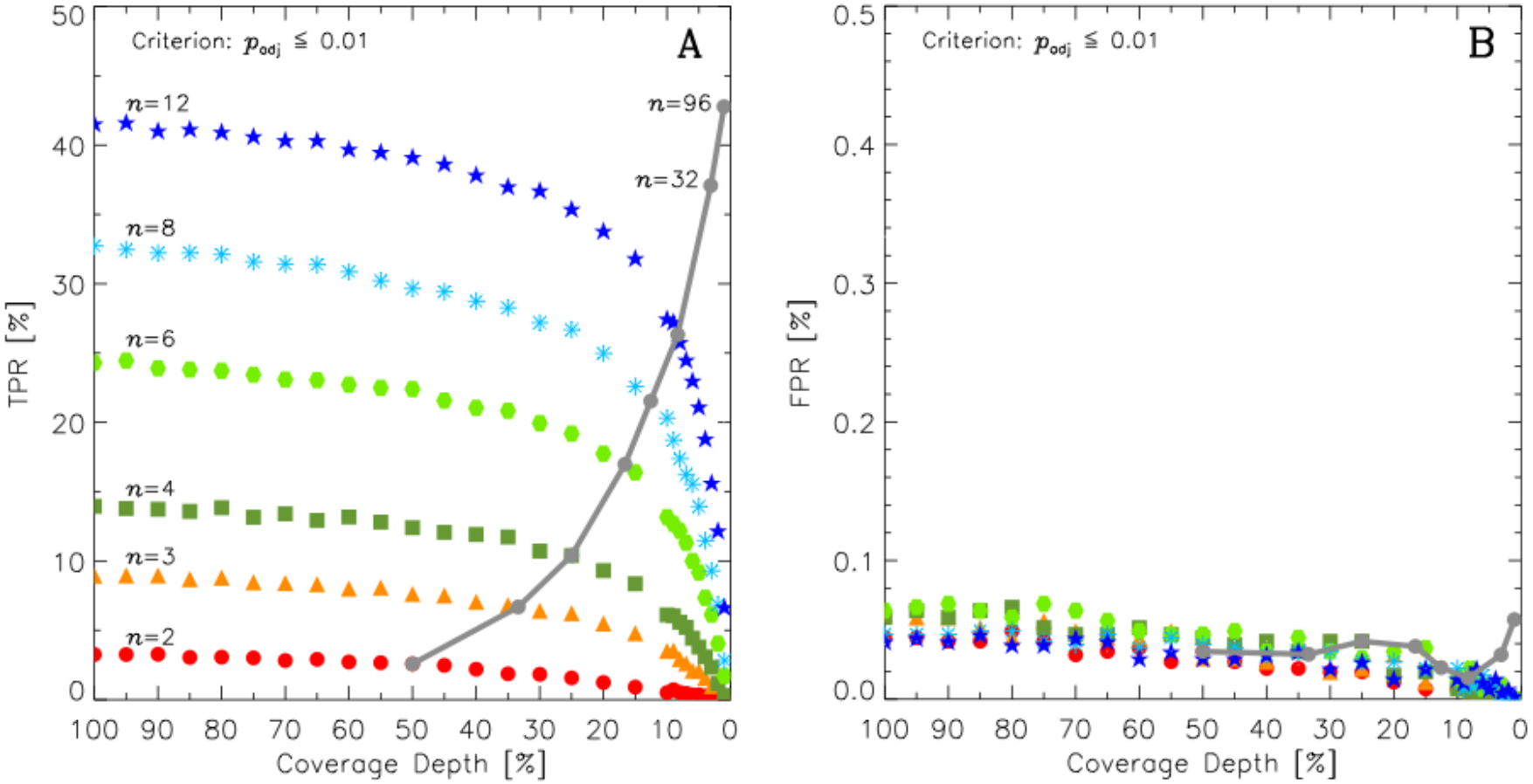
The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication (as much as you can afford) is extremely important. To robustly identify differentially expressed (DE) genes requires statistical powers.
    - (note: this is not how many reads you have for a gene within a sample, but how many biologically/statistically independent samples per treatment).
  - Technical replication does not help with statistical power (i.e. don't split a single sample and run as two libraries).

Biological replication gives far more statistical power than increased sequencing depth within a biological sample!!!!

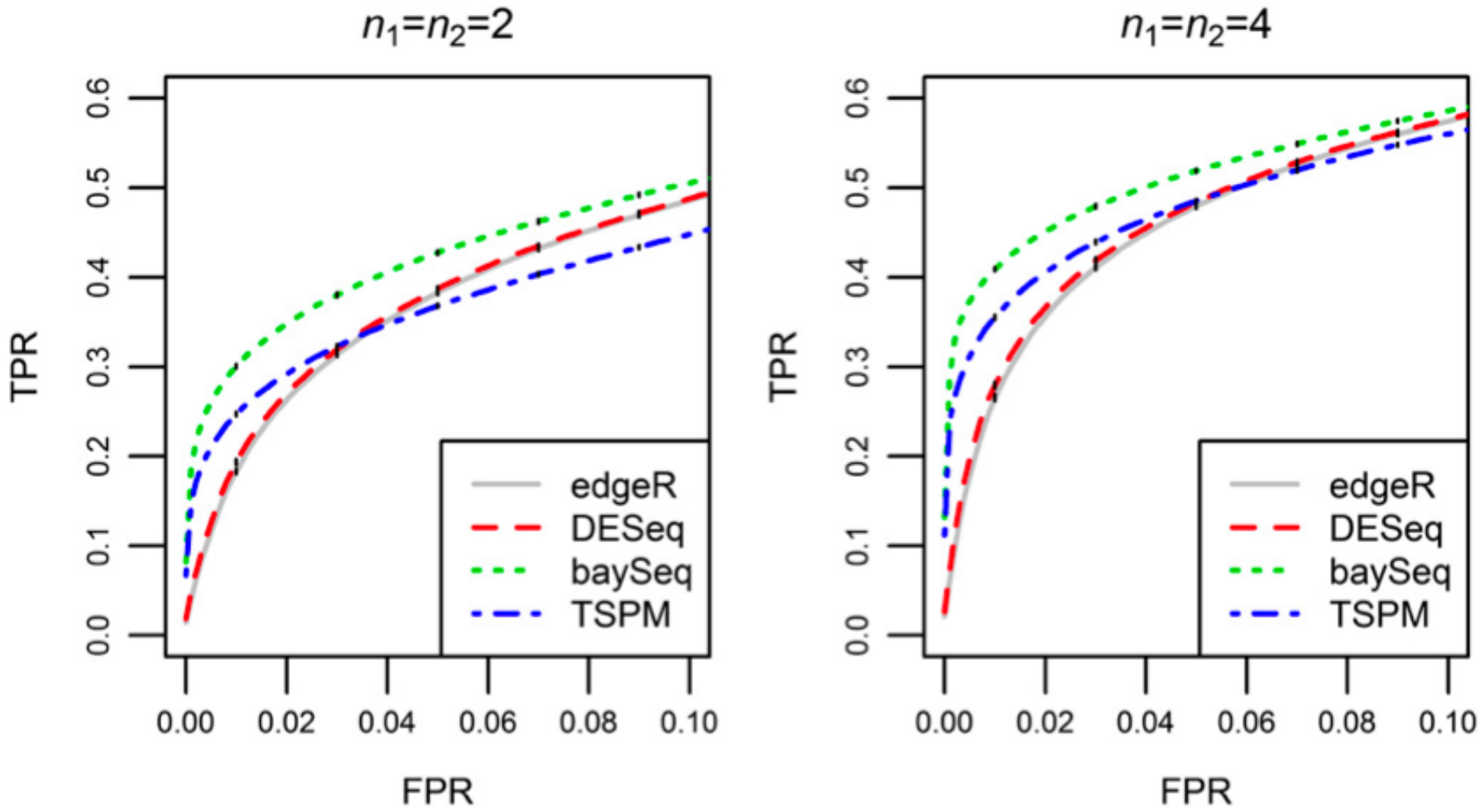
- Sequencing (and library prep) costs are still sufficiently expensive that most experiments use small numbers of biological replicates.
- Given the additional costs of library costs (~225$/ sample at our facility), many folks go for increased depth instead of more samples.
- For a given level of sequencing depth (total) for a treatment, it is far better to go for more biological replicates, each at lower sequencing depth (rather than fewer replicated at higher sequencing depth).

Biological replication gives far more statistical power than increased sequencing depth within a biological sample!!!!



Robles et al. 2012

How do the methods compare in simulation?



Kvam et al. 2012

The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication.
  - Design your experiment to avoid ***confounding*** your different treatments (sex, nutrition) with each other or with technical variables (lane within a flow cell, between flow cell variation).
    - Make diagrams/tables of your experimental design, or use a randomized design.

The basics of experimental design

- There are a few basic points to always keep in mind:
  - Biological replication.
  - Design experiment to avoid *confounding* variables.
  - Sample individuals (within treatment) randomly!

Useful references

Paul L. Auer and R.W. Doerge 2010. Statistical Design and Analysis of RNA-Seq Data. Genetics. 10.1534/genetics.110.114983
PMID: 20439781

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments BMC Bioinformatics, 11, 94. doi:10.1186/1471-2105-11-94

Designing your experiment before you start.

Sampling

Replication

Blocking

Randomization

Over all we are going to be thinking about how to **avoid Confounding** sources of variation in the data.

All of these are larger topics that are part of **Experimental Design**.

Sampling

Sampling

Replication

Blocking

Randomization

Sampling design is all about making sure that when you "pick" (sample) observations, you do so in a **random** and **unbiased** manner.

Proper sampling aims to control for unknown sources of variation that influence the outcome of your experiments.

This seems reasonable, and often intuitive to most experimental biologists, but it can be very insidious.
Whiteboard…

Sampling

Sampling

Replication

Blocking

Randomization

Biological replicates Not technical ones.

- There is little purpose in using technical replication (i.e. same sample, multiple library preps) from a given biological sample UNLESS part of your question revolves around it.

- Focus on biological variability. While you are confounding some sources of technical and biological variability, we already know a lot about the former, and little about the latter (in particular for your system).

Replication

Sampling

R**eplication**

Blocking

Randomization

Imagine you have an experiment with one factor (sex), with two treatment levels (males and females).

You want to look for sex specific differences in the brains of your critters based on transcriptional profiling, so you decide to use RNA-seq.

Perhaps you have a limited budget so you decide to run one sample of male brains, and one sample of female brains, each in one lane of a flow cell.

What (useful) information can you get out of this?

Not much (but there may be some). Why?

Replication

Sampling

R**eplication**

Blocking

Randomization

Why?

No replication. How will you know if the differences you observe are due to differences in males and females, random (biological) differences between individuals, or technical variation due to RNA extraction, processing or running the samples on different lanes.

All of these sources of variation are confounded, and there are no particularly good ways of separating them out.

But there are lots of sources of variation, so how do we account for these?

Replication

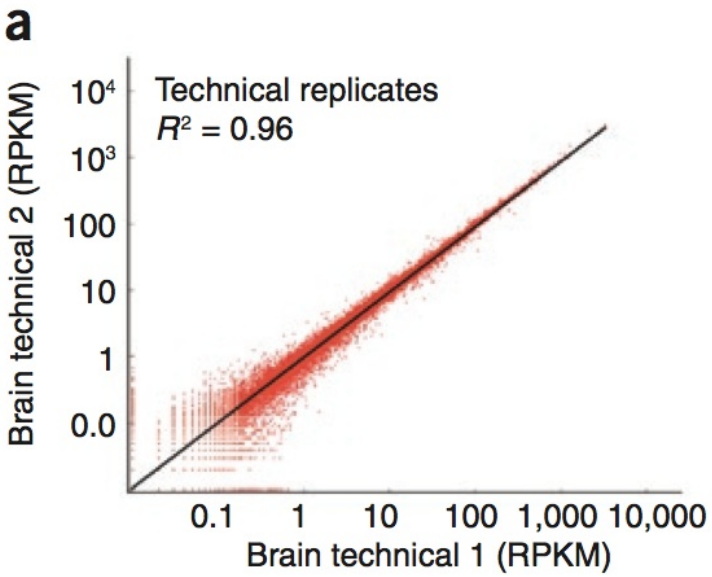
To date, several studies have suggested that “technical” replicates for RNA-seq show very little variation/ high correlation.

Sampling

Replication

Blocking

Randomization



Mortazavi et al. 2008

How might such a statement be misleading about variation?

Replication

Sampling

Replication

Blocking

Randomization

This study looked at a single source of technical variation.

Running exactly the same sample on two different lanes on a flow cell.

This completely ignores other sources of "technical variation"
variation due to RNA purification
variation due to fragmentation, labeling, etc..
lane to lane variation
flow cell to flow cell variation

All of these may be important (although unlikely interesting) sources of variation…

However…..

Replication

Sampling

Replication

Blocking

Randomization

Many studies have ignored the BIOLOGICAL SOURCES of VARIATION between replicates. In most cases biological variation between samples (from the same treatment) are generally far more variable than technical sources of variation.

While it would be nice to be able to partition various sources of technical variation (such as labeling, RNA extraction), it often too expensive to perform such a design (see white board).

IF you have limited resources, it is generally far better to have biological replication (independent biological samples for a given treatment) than technical replication.

Does these lead to confounded sources of variation?

Blocking

Sampling

Replication

Blocking

Randomization

Blocks in experimental design represent some factor (usually something not of major interest) that can strongly influence your outcomes. More importantly it is a factor which you can use to group other factors that you are interested in.

For instance in agriculture there is often plot to plot variation. You may not be interested in the plot themselves but in the variety of crops you are growing.

But what would happen if you grew all of strain 1 on plot 1 and all of strain 2 on plot 2?

Whiteboard.

These plots would represent blocking levels

Blocking

Sampling

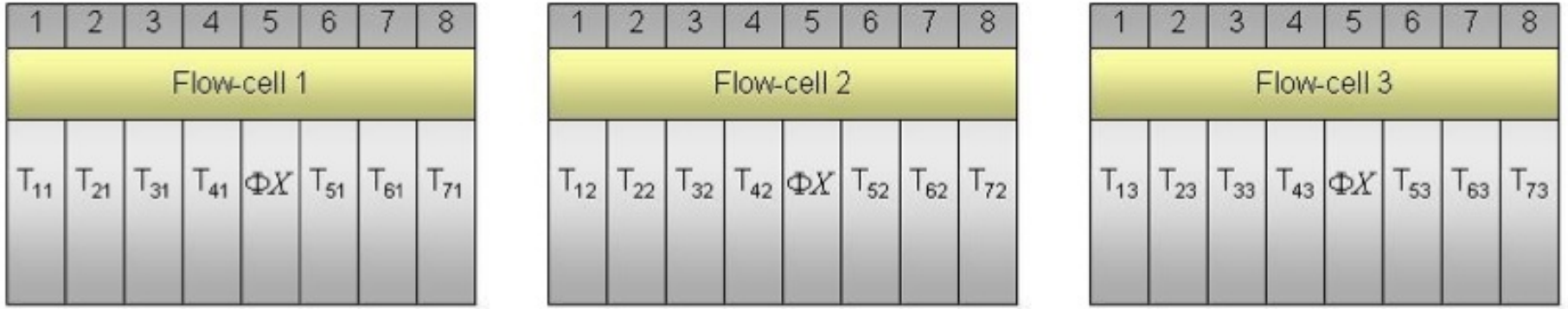
Replication

Blocking

Randomization

In genomic studies the major blocking levels are often the slide/chip for microarrays (i.e. two samples /slide for 2 color arrays, 16 arrays/slide for Illumina arrays).

For GAII/HiSeq RNA-seq data the major blocking effect is the flow cell itself and lanes within the flow cell.



Auer and Doerge 2010

Blocking

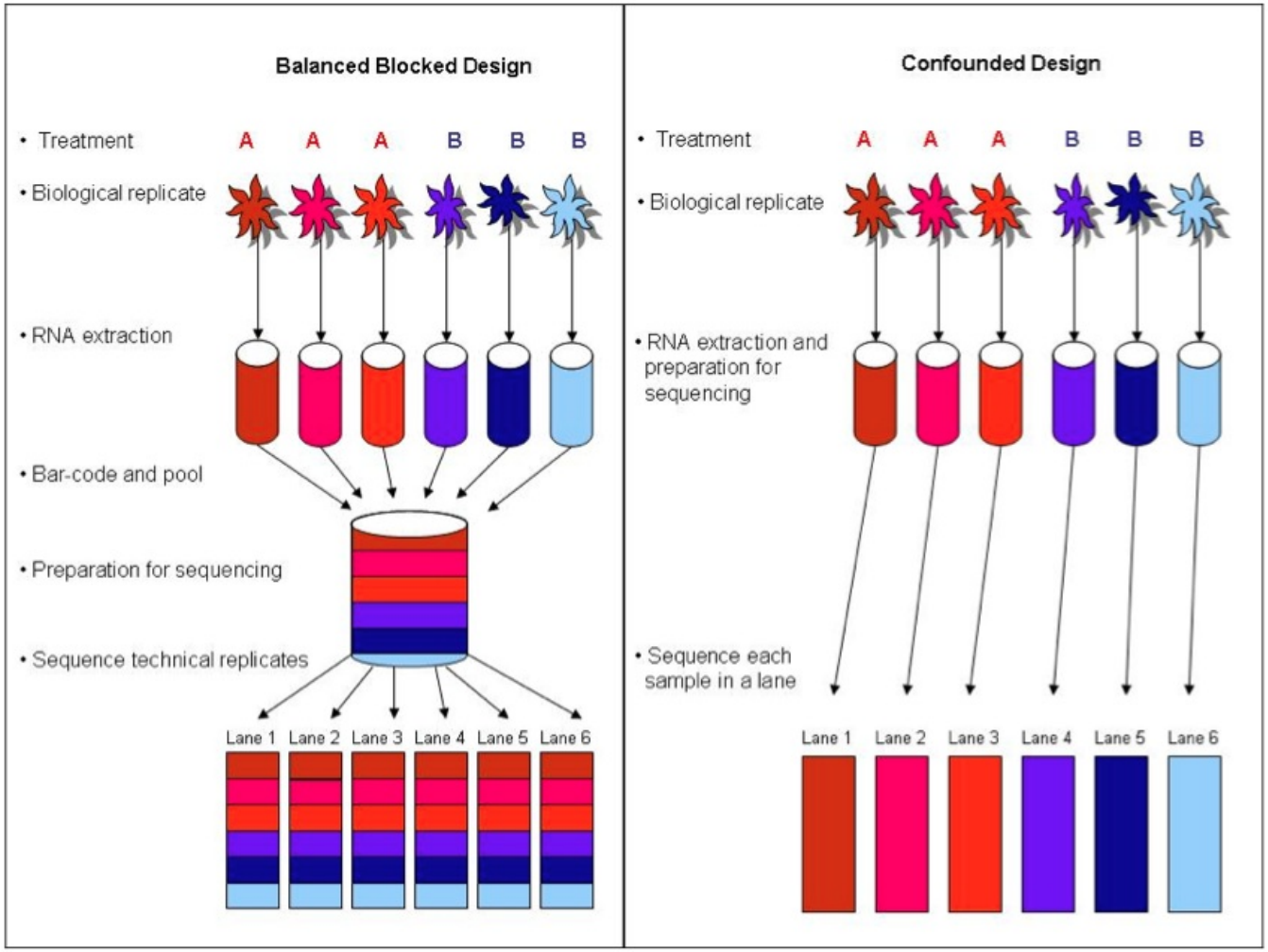
Sampling

Replication

Blocking

Randomization

Incorporating lanes as a blocking effect



Auer and Doerge 2010

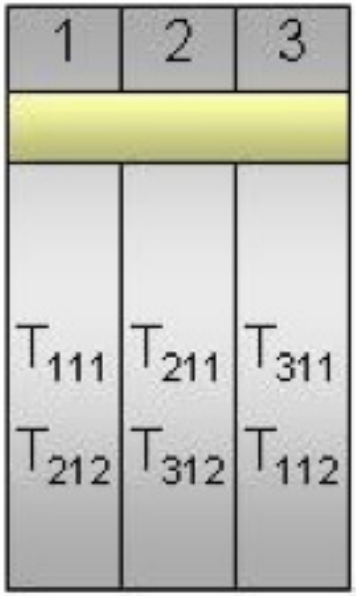
Blocking designs

Sampling

Replication

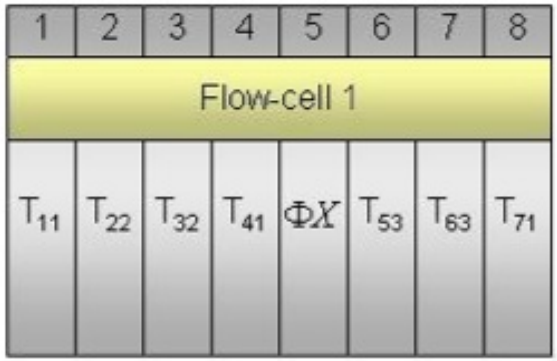
Blocking

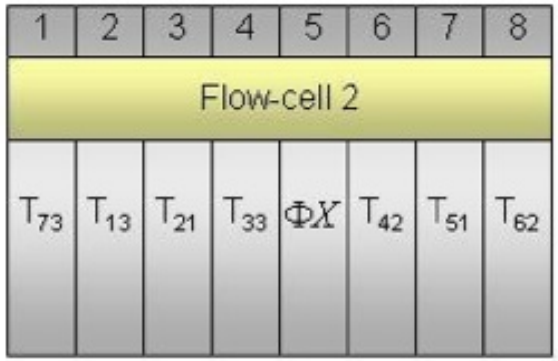
Randomization

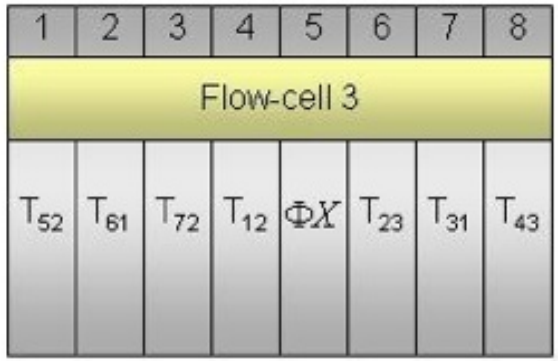


Balanced **I**ncomplete **B**locking **D**esign (BIBD)

Let's dissect these subscripts.







Balanced for treatments across flow cells.. Randomized for location

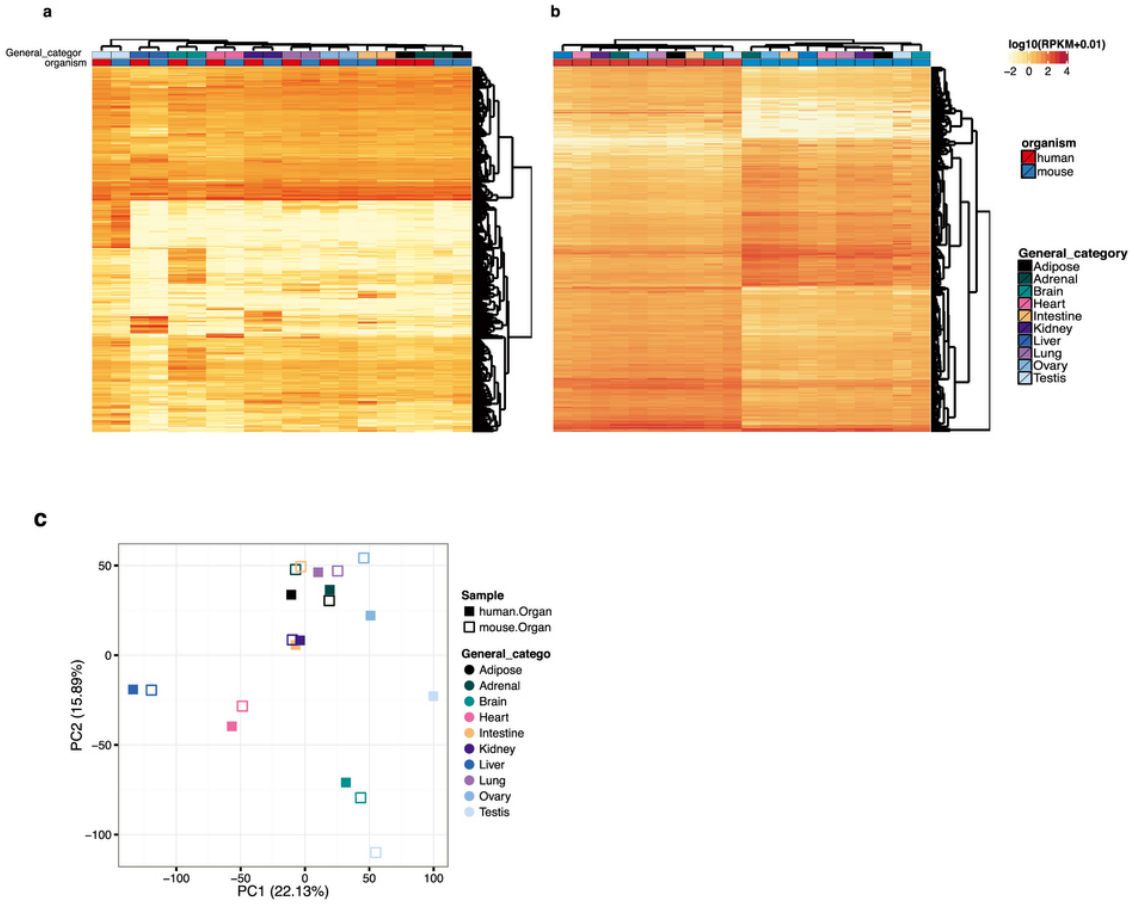
Auer and Doerge 2010

What standard technical issues should you consider for blocking:

- Flow Cell
- Lane
- Adaptors
- Library prep
- Same instrument
- People!
- RNA extraction/purification

What happens when you fail to block (or replicate)?

In a recent analysis of the mod-encode data, RNAseq data suggested that clustering (for gene expression) more by species than by tissue. This was an unusual finding.



Yue F, Cheng Y, Breschi A, et al.: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364

Lin S, Lin Y, Nery JR, et al.: Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014; 111(48): 17224–17229

A new re-analysis demonstrated some potentially serious issues with the experimental design

Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: indexed, http://f1000r.es/5ez] F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

Figure 1. Study design for :
Yue F, Cheng Y, Breschi A, et al.: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364
Lin S, Lin Y, Nery JR, et al.: Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014; 111(48): 17224–17229

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

Gilad Y and Mizrahi-Man O 2015 [v1; ref status: awaiting peer review, http://f1000r.es/5ez]
F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

F1000Research

Using RNAseq

- Transcriptome assembly.
- Improving genome assembly/annotation.
- SNP discovery (large genomes)
- Transcript discovery (variants for Transcription start site, alternative splicing, etc..)
- Quantification of (alternative transcripts)
- Differential expression analysis across treatments.

Using RNAseq: differential expression

- Differential expression of what?

Using RNAseq: differential expression

- Differential expression of what?

- Differential expression at the level of “genes”
- Allele specific expression
- Quantification of alternative transcripts

Your primary goals of your experiment should guide your design.

- The exact details (*# biological samples*, sample depth, read_length, strand specificity) of how you perform your experiment needs to be guided by your primary goal.

- Unless you have all the $$, no single design can capture all of the variability.

Your goals matter

- For instance: If your primary interest in discovery of new transcripts, sampling deeply within a sample is probably best.

- For differential expression analyses, you will almost never have the ability to perform Differential expression analysis on very rare transcripts, so it is rarely useful to generate more than 15-20 million read pairs.

Are single_ended reads ever useful?

- In my experience (plants and animals), almost never.
- My primary organism (*Drosophila melanogaster*) is one of the best annotated and experimentally validated genomes.
- Even still, we get surprising ambiguity for reads 75bp and shorter, which mostly goes away with PE.
- Hopefully less of a problem now (as most people are doing 100 -150 bp+).

What was once thought to be separate goals are now clearly recognized as intertwined.

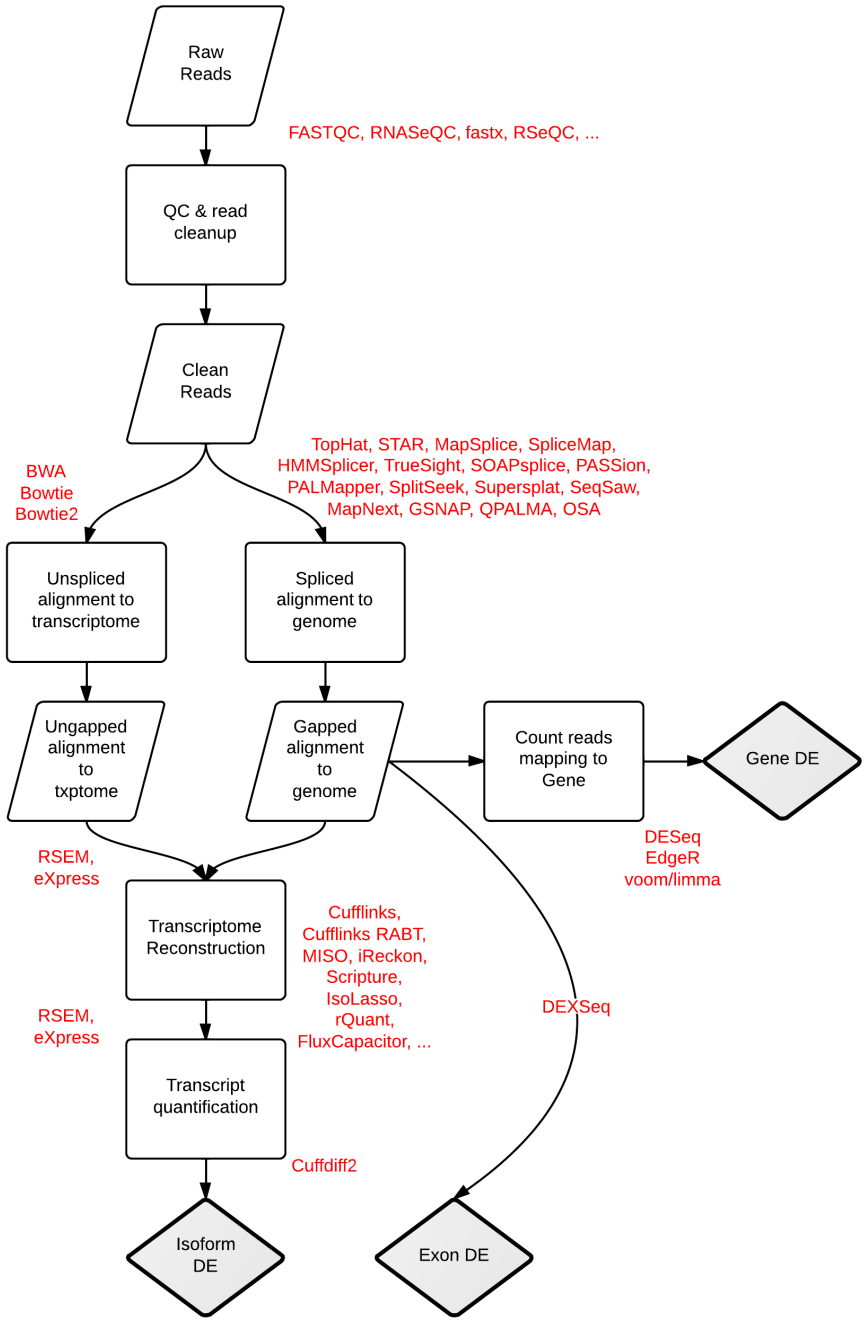
- Early work for RNA-Seq tried to “mirror” the type of gene level analysis used in microarrays.
- However, RNA-seq has demonstrated how important it is to take into account alternative transcripts, even when attempting to get “gene level” measures.

How do we put together a useful pipeline for RNAseq

- What are the steps we need to consider?

How do we put together a useful pipeline for RNAseq

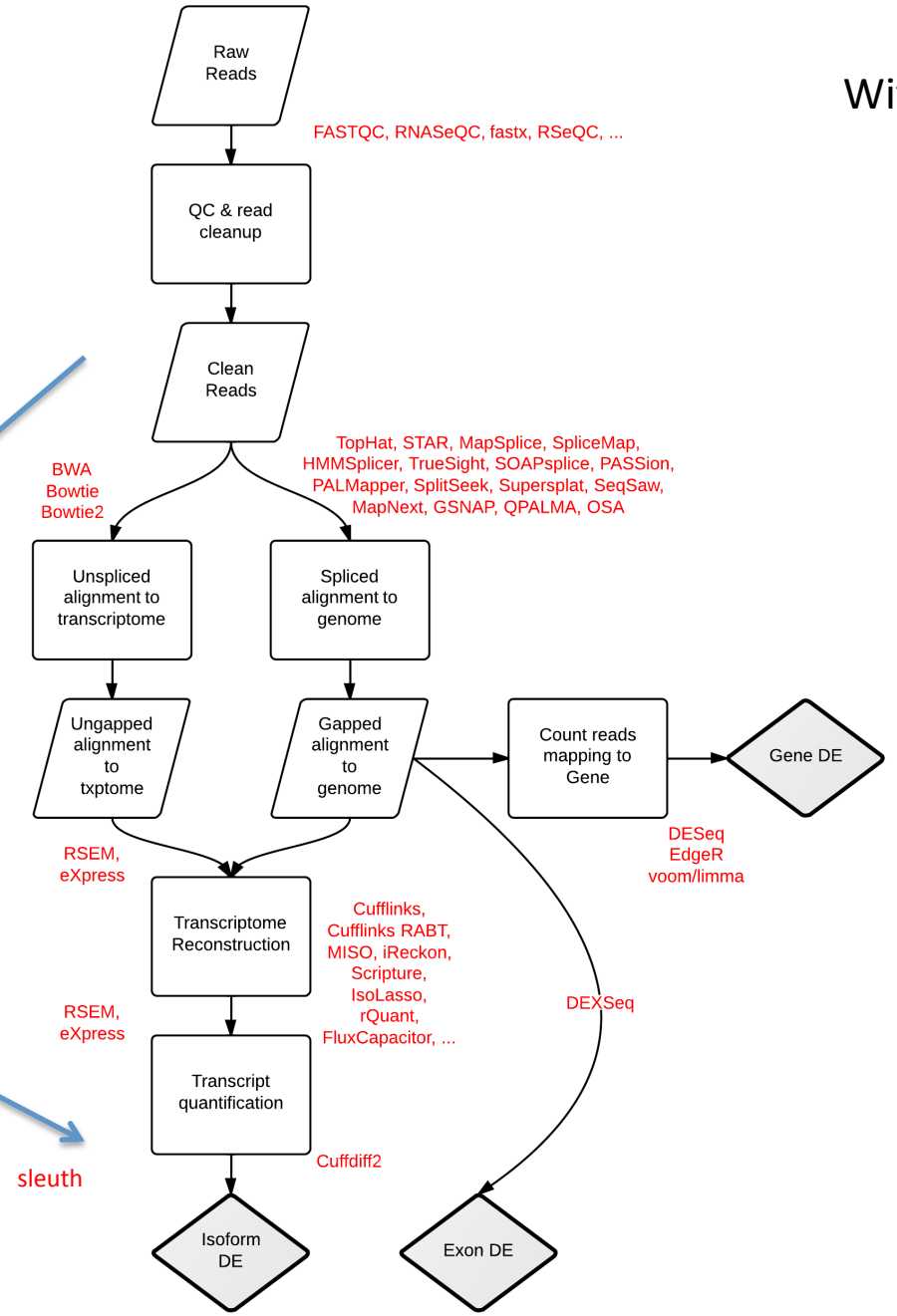
- What are the steps we need to consider?
- Genome/transcriptome assembly.
- Trimming/error correction of reads.
- Mapping reads to genome/transcriptome.
- Deal with alternative transcripts (new transcriptome)?
- Remap & count reads.
- Differential expression.



RNA-seq Workflows and Tools. Stephen Turner. Figshare. http://dx.doi.org/10.6084/m9.figshare.662782

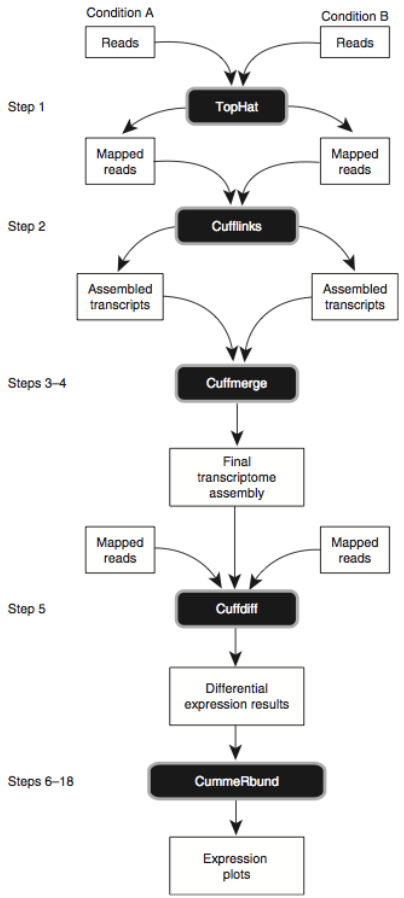
With some modifications…

Pseudo-alignment to transcriptome.
(Sailfish, salmon, RNA-Skim, kallisto)



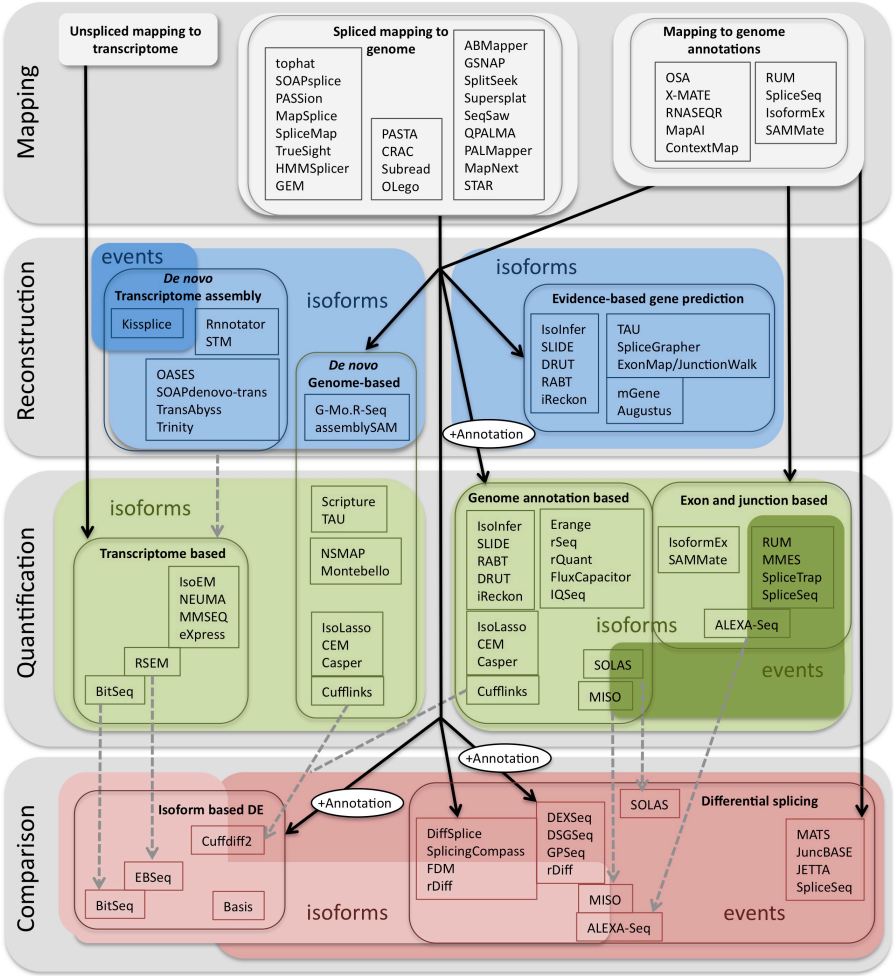
RNA-seq Workflows and Tools. Stephen Turner. Figshare. http://dx.doi.org/10.6084/m9.figshare.662782

The “tuxedo” protocol for RNA-seq

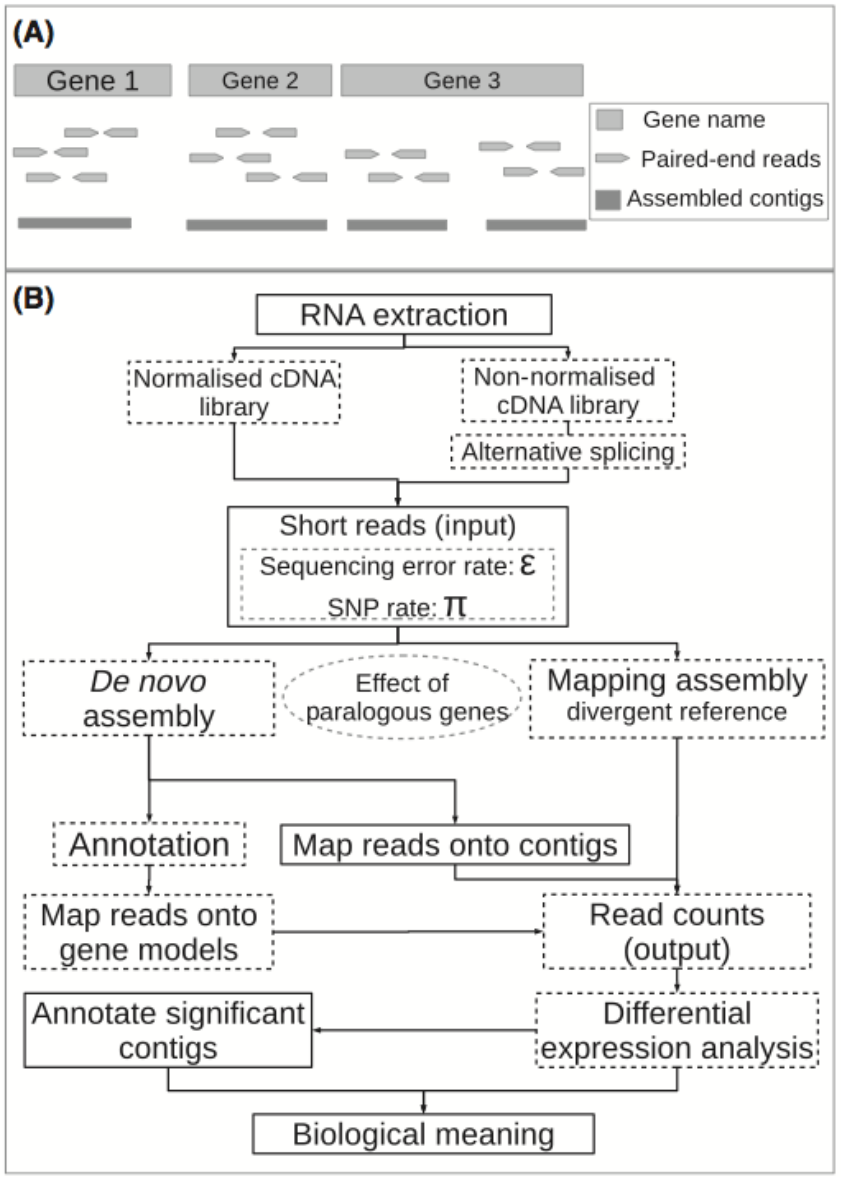


Trapnell et al 2012

Pipelines for RNA-seq (geared towards splicing)



Methods to Study Splicing from RNA-Seq. Eduardo Eyras, Gael P. Alamancos, Eneritz Agirre. Figshare. http://dx.doi.org/10.6084/m9.figshare.679993 also see http://arxiv.org/abs/1304.5952



Vijay et al 2012

The point…

- There is no single “best” way forward yet. It is probably best to try several pipelines and think carefully about each of the steps.

What we do

- Im my lab, we find that light quality trimming, and adapter pruning is very helpful.
- For species with well annotated genomes (Drosophila) we map using splice aware tools (STAR or tophat), and count using HTSeq. We have had bad luck with RSEM and cufflinks.
- For species without a reference genome (and too divergent from other species). We map to the de novo transcriptome and count using eXpress*.
- For both of these we then tend to use DESeq2 and limma.

Counting fragments to genes, transcripts, exons, kmers..

- I have been discussing this from a perspective of counting reads to features such as genes or transcripts.
- Cogent arguments are made to instead use using exons, or even individual nucleotides within exons (Lauren McIntyre).
- Recently a number of approaches of pseudo-mapped kmers have come along (sailfish, salmon, RNA-skim, kallisto). These are incredibly fast.

What features should we count

- An important (and impactful) issue that you need to consider in your analysis is what features you are trying to count:
  - Genes as features
  - Transcripts as features
  - Exons as features
  - Kmers as features

How should we map reads

- Do we want to map to a reference genome (with a “splice aware” aligner)?
- Or do we want to map to a transcriptome directly.

- What is preferable, to generate a *de novo transcriptome* or map to a “closely” related species?

And before we map reads…

- How should we filter (based on quality) reads (if at all)?

- What are some of the considerations?

Mapping to a transcriptome

- What are the downsides to mapping to a transcriptome?

Mapping to a transcriptome

- unspliced read aligners are useful against a transcript (or cDNA) database, such as that generated for a de novo transcriptome.

- For this BW is faster than seed based approaches (shrimb & stampy), but the latter may be preferred if mapping to "distant" transcriptomes.

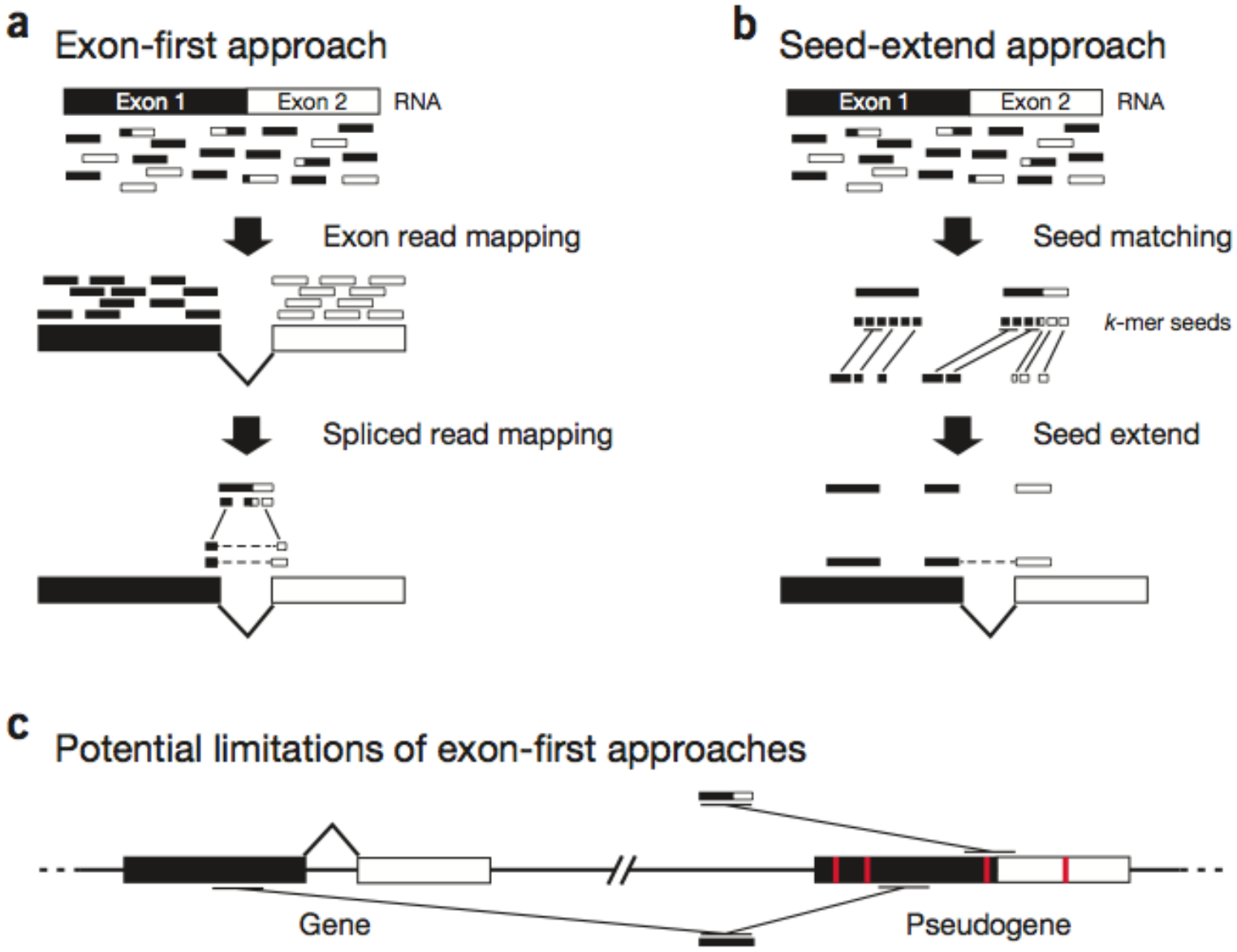
Mapping to the genome

- How do we deal with alternative transcripts or paralogs during mapping?

"splicing aware" aligners:

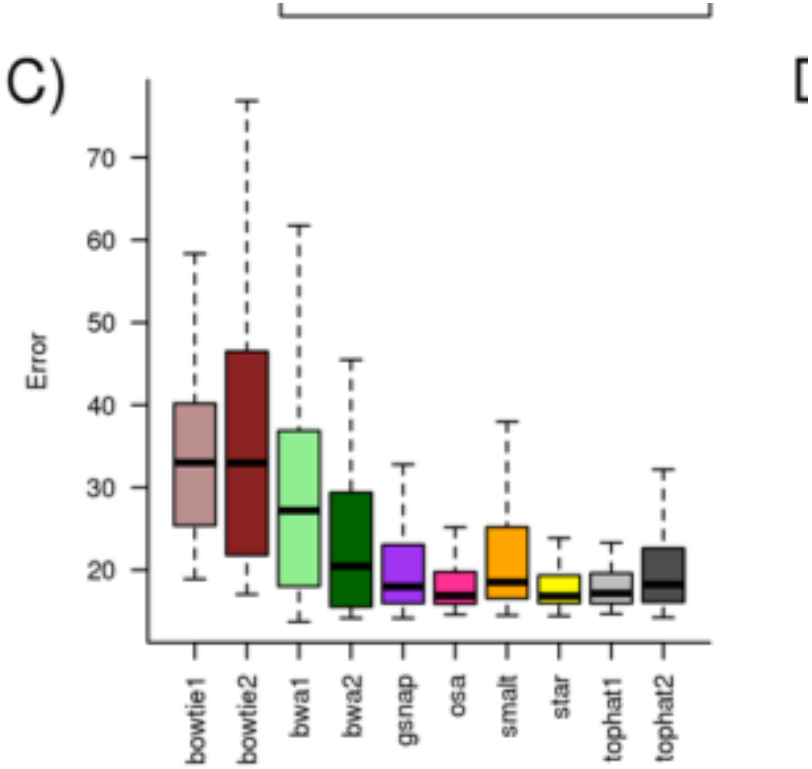
  - Exon First: (tophat, MapSplice, SpliceMap) Fig1A Garber
  - Step 1 - map reads to genome
  - Step 2 -unmapped reads are split, and aligned.

- Seed & extend (Fig1B Garber) (GSNAP, QPALMA)
  - kmers from reads are mapped (the seeds), and then extended



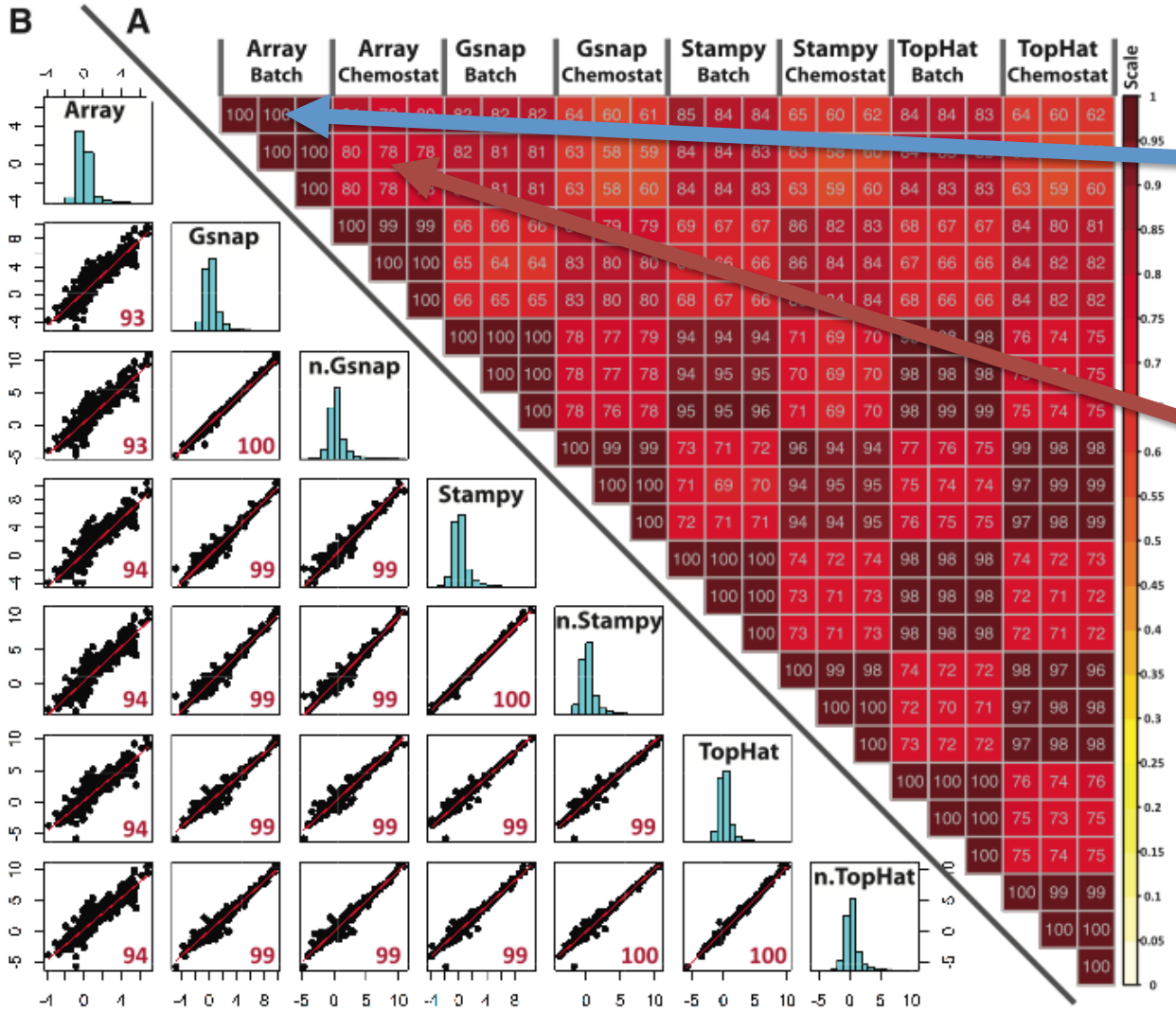
Garber et al. 2011

The variation in the mapping step (at least with a reference genome) seems to have modest effects.



RNA-seq gene profiling - a systematic empirical comparison
Fonseca et al (2014). http://dx.doi.org/10.1101/005207

What does this tell us?



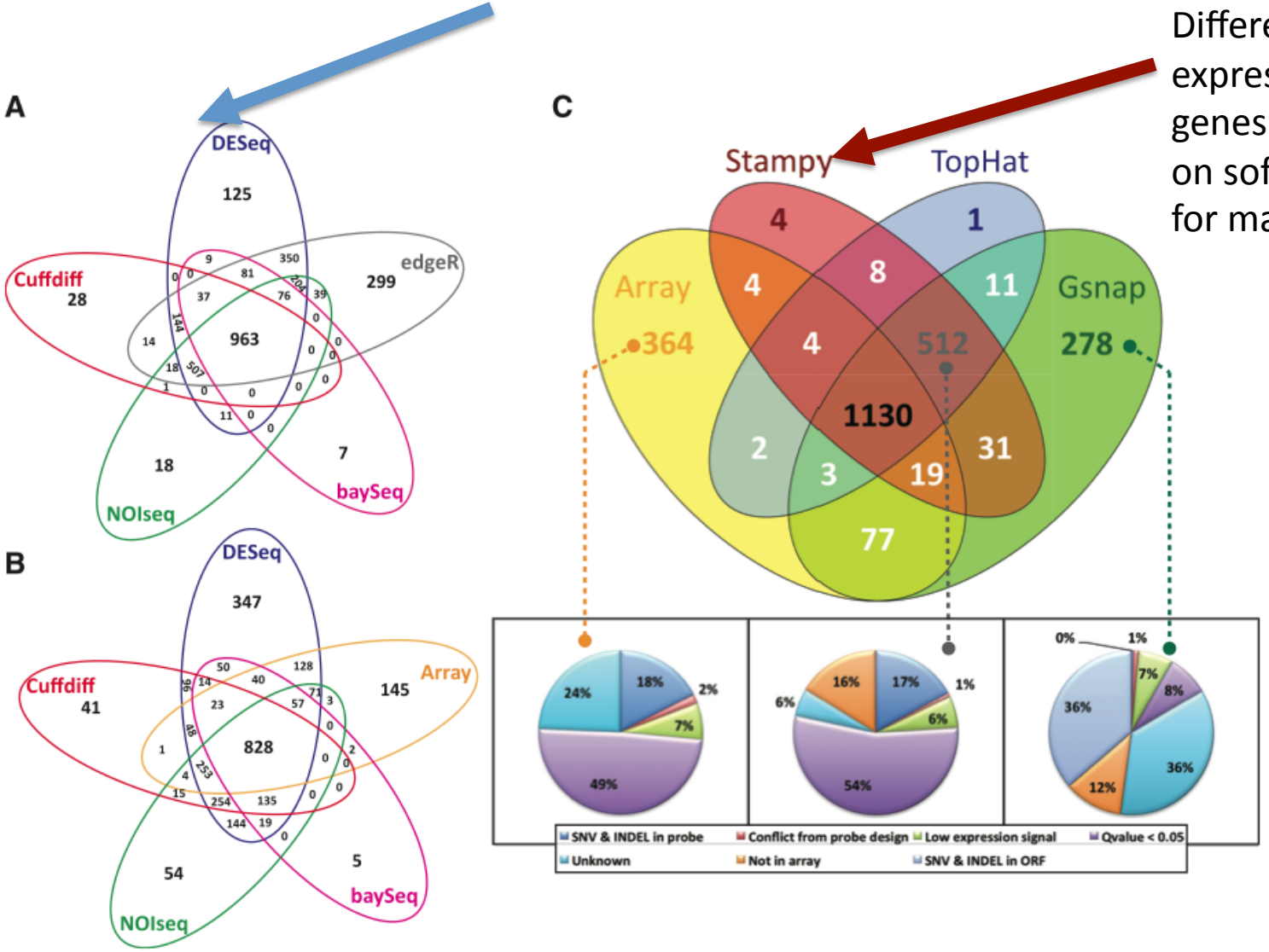
These cells are biological replicates (diagonals)

These cells are for different software

Nookaew et al 2102 NAR

Differentially expressed genes based on software for quantification

Differentially expressed genes based on software for mapping



Nookaew et al 2102 NAR

Which to use

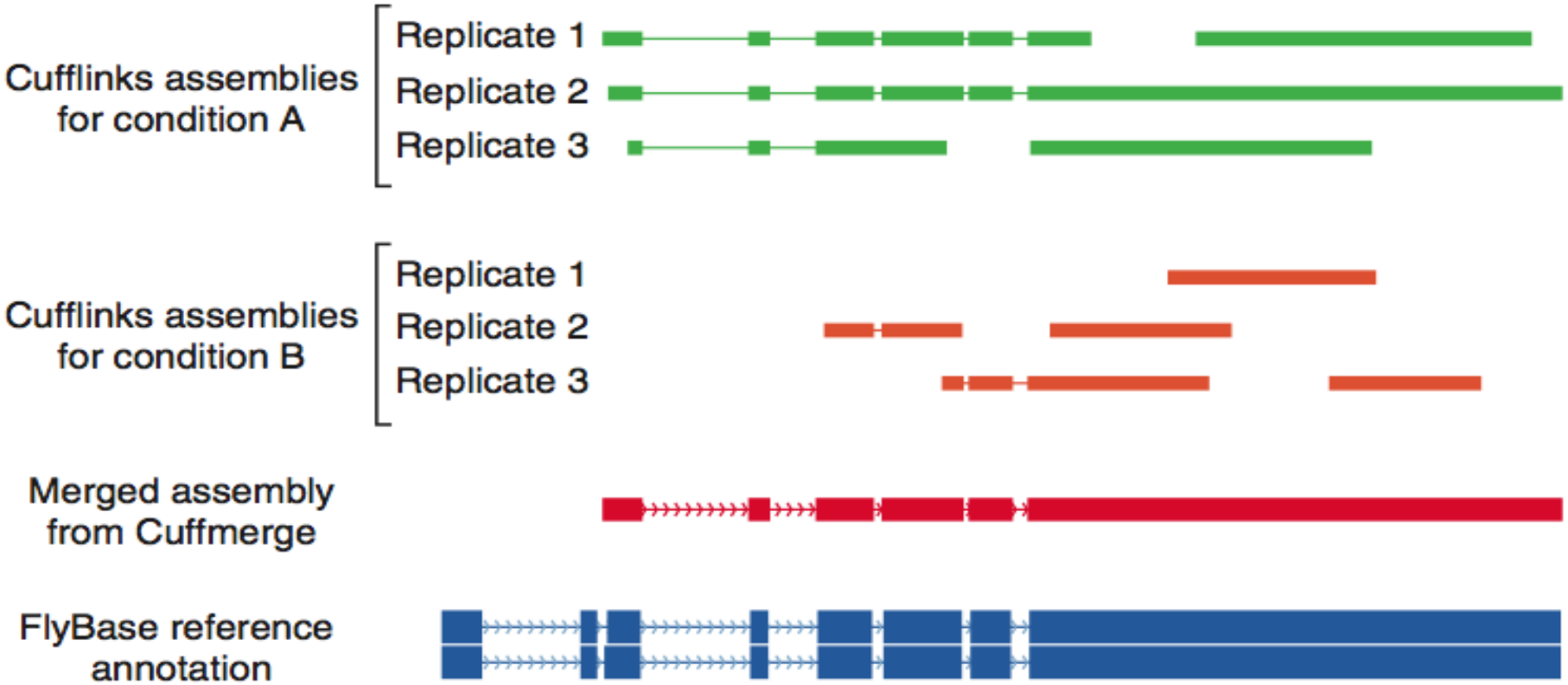
- If a (close to?) perfect match transcriptome assembly is available for mapping. Burrows-wheeler based aligners can be much faster than seed based methods (upto 15x faster)

- BW based aligners have reduced performance once mismatches are considered.
  - Exponential decrease in performance with each additional mismatch (iteratively performs perfect searches).
  - Seed methods may be more sensitive when mapping to transcriptomes of distantly related species (or high polymorphism rates).

From Garber et al. 2011

How could mapping reads (whether to a reference genome or transcriptome) influence our downstream counts?

Merging all transcripts?



Trapnell et al 2012.

Counting

- One of the most difficult issues has been how to count reads.

- What are some of the issues that we need to account for during counting of reads?
  - Transcript length (my be a minor concern depending on application).
  - Ambiguously (multi-)mapped reads. How should you count those.

Dealing with multi-mapped reads

- Several options:
  - Only use reads that map uniquely (exclude all multi-mapped reads).
  - What might be the problem with such an approach?

Dealing with multi-mapped reads

- Several options:
  - Only use reads that map uniquely (exclude all multi-mapped reads).
  - What might be the problem with such an approach?
  - What happens if your transcriptome assembly (because of polymorphism), has assembled two or more genes for a single true gene?

Dealing with multi-mapped reads

- Several options:
  - Only use reads that map uniquely.
  - Use all reads (unique + multi-mapped).
    - What are the problems here?

Dealing with multi-mapped reads

- Several options:
  - Only use reads that map uniquely.
  - Use all reads (unique + multi-mapped).
    - What are the problems here?
    - Pseudo-replication(ish)

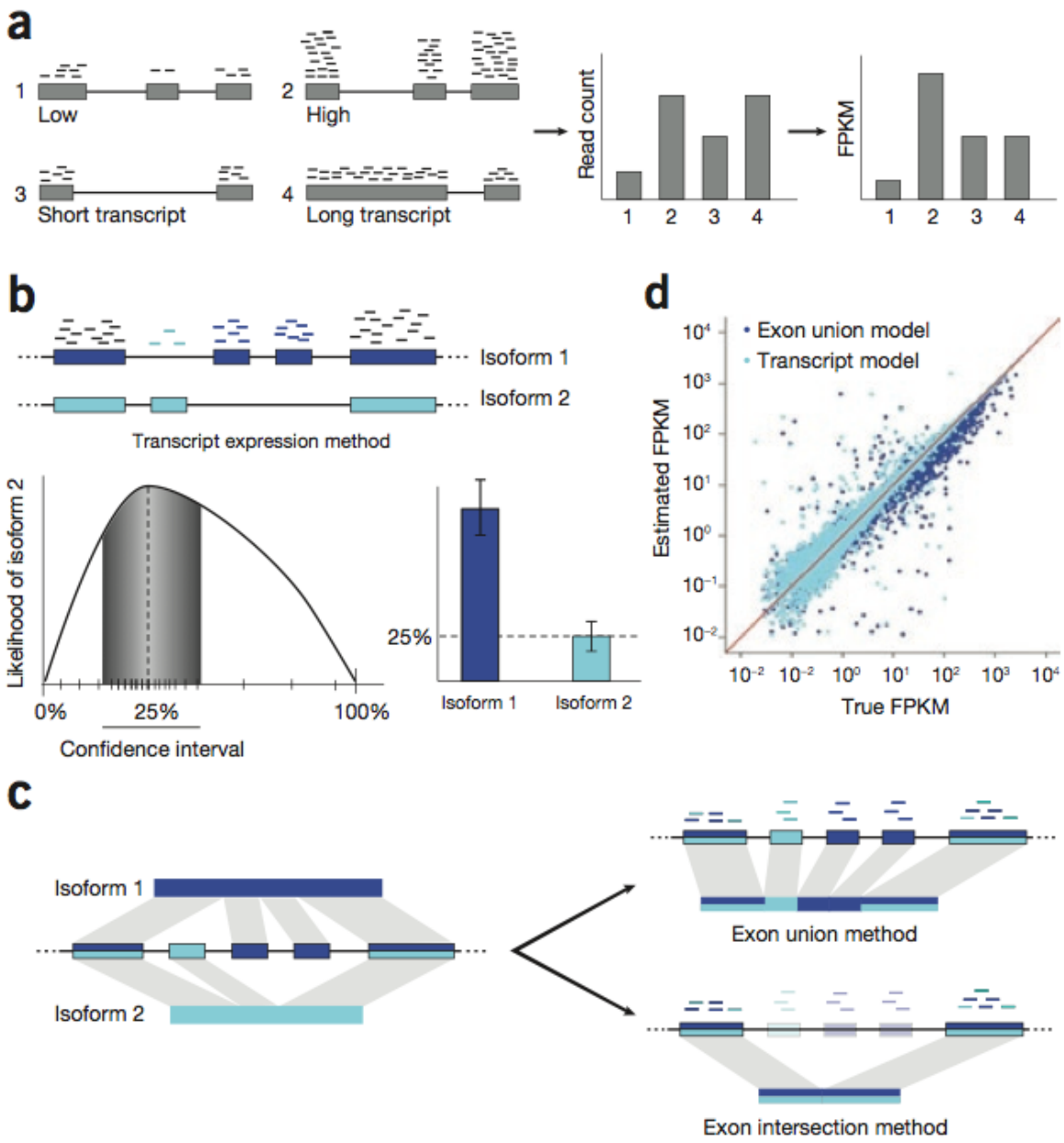
Dealing with multi-mapped reads

- Several options:
  - Only use reads that map uniquely (HTSeq, eXpress).
  - Use all reads (HTSeq, eXpress).
  - Unique reads + assigning multi-mapped reads "randomly"
  - Unique reads + model based inference for where reads belong (eXpress).

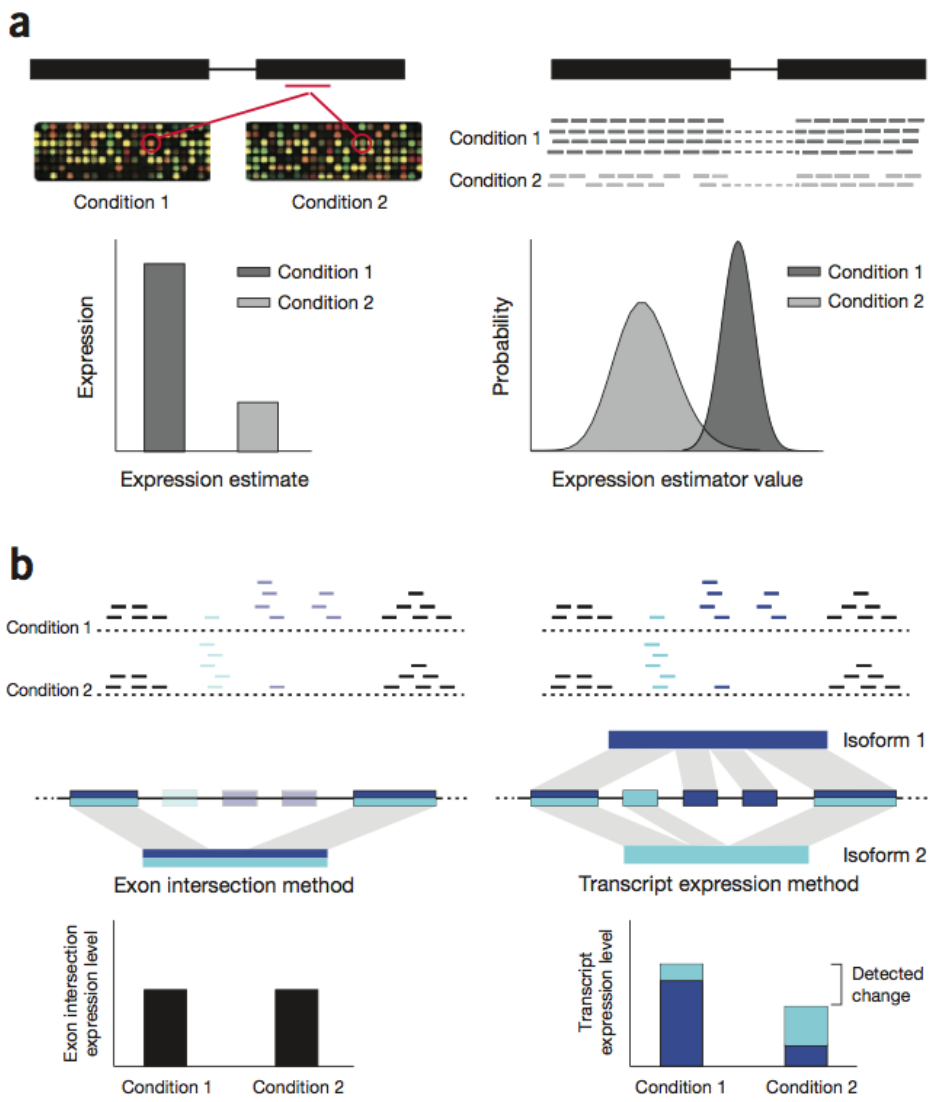
Accounting for multiple isoforms (when counting alternative transcripts).

- - Only count reads that map uniquely to an isoform (Alexa-Seq, HTSeq). Can be very problematic, when isoforms do not have unique exons.

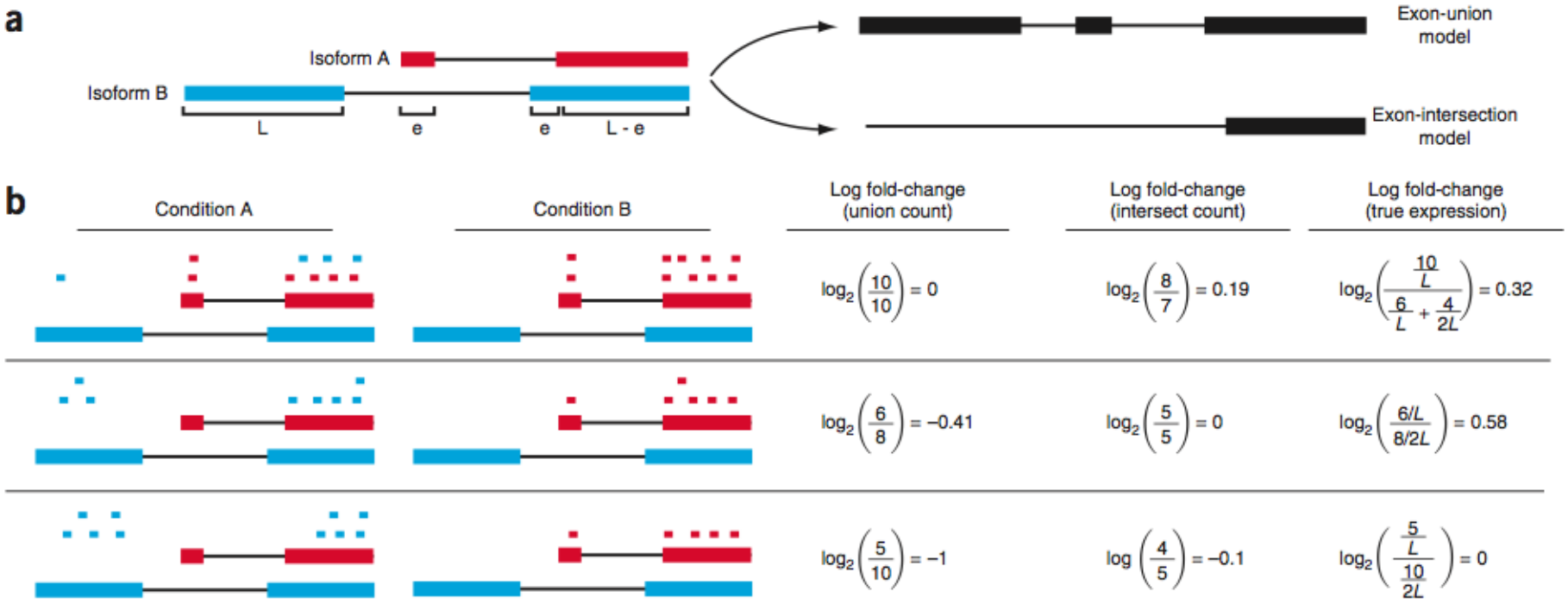
- - so called "isoform-expression" methods (cufflinks, MISO) model the uncertainty parametrically (often using MLE). The model with the best mix of isoforms that models the data (highest joint probability) is the best estimate. How this is handled differs a great deal by the different.



Garber et al. 2011



Garber et al. 2011



Trapnell et al 2013

Counting

- What are we trying to count?
- Gene level measure (eXpress, corset, RSEM, HTSeq, featureCounts, summarizeOverlaps).
- Exon level (HTSeq, ???)
- Transcript level (HTSeq, Cufflinks, ….)
- Clustering (corset)
- Kmer (sailfish, RNA-skim)

However…

- There has been a great deal of discussion and disagreement about this (see seqanswer forums, and “discussions” between Simon Anders and Lior Patcher).
- Fundamentally there are numerous cases where both methods fail. So take care.

Seqanswer or blog postings of use

- http://seqanswers.com/forums/showpost.php?p=102911&postcount=60
- http://gettinggeneticsdone.blogspot.com/2012/11/star-ultrafast-universal-rna-seq-aligner.html
- http://gettinggeneticsdone.blogspot.com/2012/12/differential-isoform-expression-cuffdiff2.html
- http://gettinggeneticsdone.blogspot.com/2012/09/deseq-vs-edger-comparison.html

Problems with cufflink and cuffdiff? Reproducibility…

- http://seqanswers.com/forums/showthread.php?t=20702
- http://seqanswers.com/forums/showthread.php?t=17662
- http://seqanswers.com/forums/showthread.php?t=23962
- http://seqanswers.com/forums/showthread.php?t=21020
- http://seqanswers.com/forums/showthread.php?t=21708
- http://www.biostars.org/p/6317/

Take home message

- For Differential expression analysis, by and large counts are best, not an adjusted count.
- For gene-by-gene analyses, accounting for transcript length is not essential.

- However, there are several situations where variation in counts due to the influence of transcript length is important.
  - Multivariate analyses (clustering, PCA, MDS).
  - Collapsing multiple transcripts (of potentially different length) into a "gene" level measure of counts.
  - You can also include transcript length (or effective length) as a covariate in the statistical analysis itself (either as an offset or a covariate).

Counting at the “gene” or exon level may be simpler (at least initially).

- i.e. all mapped reads for transcripts associated with a particular “gene” get counted (HTSeq, corset, eXpress, RSEM (?)).

Counting reads

- Htseq (python library) works with Deseq.

- In our experience this is both easy (ish) to use and counting in a sensible manner.
- I remain very confused about getting “counts” out of both RSEM and Cufflinks…
- eXpress has some nice features, and is fast.
- featureCounts in R
- Corset uses a clustering approach useful for mapping against de novo transcriptomes where there may be false contigs.
- Sailfish, salmon, RNA-skim and kallisto are all part of a new breed of pseudo-mapping and counting tools.

Differential expression

- Deseq/DESeq2 (http://www.ncbi.nlm.nih.gov/pubmed/20979621)
- EDGE-R
- EBseq (RSEM/EBseq)
- RSEM (http://deweylab.biostat.wisc.edu/rsem/)
- eXpress (http://bio.math.berkeley.edu/eXpress/overview.html)
- Beers simulation pipeline(http://www.cbil.upenn.edu/BEERS/)
- DEXseq (http://bioconductor.org/packages/release/bioc/html/DEXSeq.html)
- Limma (voom)
- Sleuth (works with kallisto)

Example workflows (also see papers I sent)

- http://f1000research.com/articles/4-1070/v1
- http://www.bioconductor.org/help/workflows/rnaseqGene/
- http://jura.wi.mit.edu/bio/education/hot_topics/QC_HTP/QC_HTP.pdf
- http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNAseqDE_Dec2011.pdf

Why do we care about multiple comparisons?

- Let's go to the R tutorial at
- https://github.com/DworkinLab/Bio720/blob/master/DealingWithMultipleComparisons/SimulationsMultipleComparisons.R

How can we deal with multiple comparisons