NGS2014 Monday & Tuesday

- Today, Tomorrow and maybe part of Wed we are going to get more in depth on aspects of RNAseq.
- I will be giving a lecture on some additional aspects of RNAseq (M&T).
- Matt will go over in more detail Transcriptome assembly (and will discuss SOAP-deNovo Trans as well as Trinity).
- Chris, Matt and Meg will help lead you through a "choose your own adventure" of RNAseq analysis.

NGS2014 Monday & Tuesday

- Tomorrow I will give a short lecture on some of the basics of performing differential expression analysis with counts from RNAseq data.
- Meg and Matt may discuss or do a short tutorial on aspects of annotating and "cleaning" your transcriptomes.
- We will also do a tutorial comparing some of the results of your "counting" from today.

General considerations for RNA-seq quantification for differential expression

or how to count.

Ian Dworkin

First some apologies (I am moving back to Canada, and I need to practice this).

- From all I have seen and heard, Meg has already covered a number of important aspects of this, so there will be some repetition.
- I am purposefully repeating some of the same highlights in a few places, because it truly bears repeating!

Biological replicates Not technical ones.

- Meg already went over this, but there is little purpose in using technical replication 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.

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

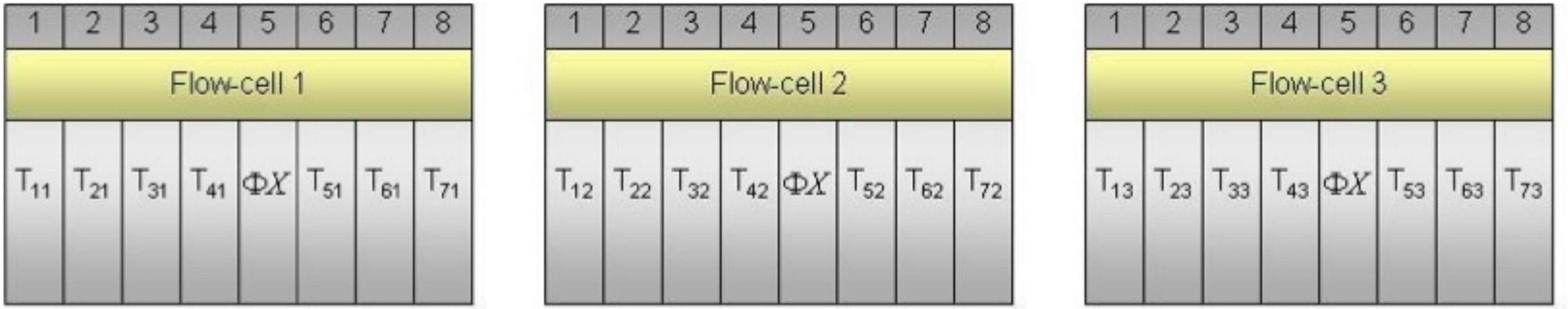
Replication

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, or lanes within the flow cell.



1	2	3	4	5	6	7	8
Flow-cell 1							
T_{11}	T_{21}	T_{31}	T_{41}	ΦX	T_{51}	T_{61}	T_{71}

1	2	3	4	5	6	7	8
Flow-cell 2							
T_{12}	T_{22}	T_{32}	T_{42}	ΦX	T_{52}	T_{62}	T_{72}

1	2	3	4	5	6	7	8
Flow-cell 3							
T_{13}	T_{23}	T_{33}	T_{43}	ΦX	T_{53}	T_{63}	T_{73}

Auer and Doerge 2010

Blocking

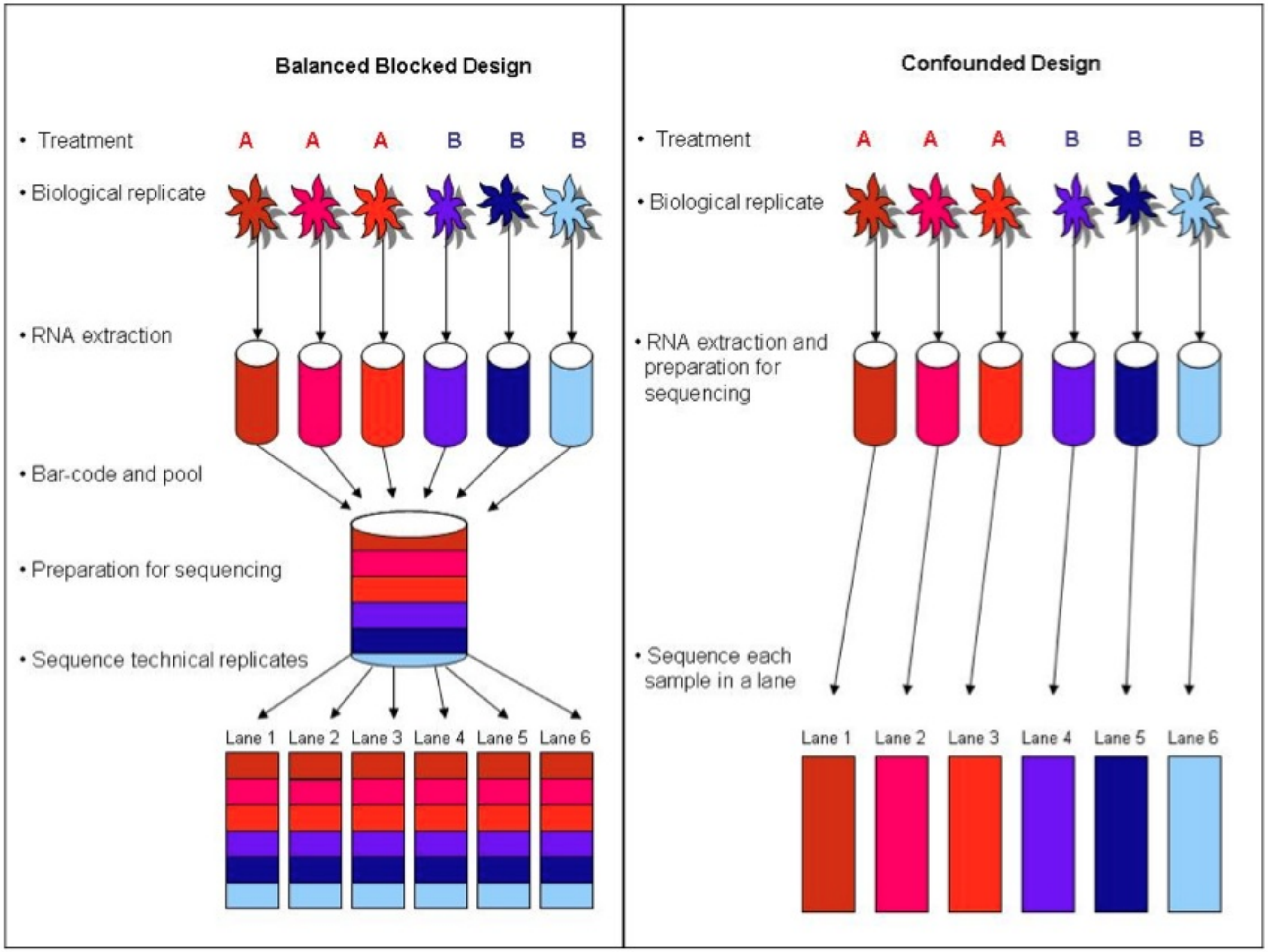
Incorporating lanes as a blocking effect

Sampling

Replication

Blocking

Randomization



Auer and Doerge 2010

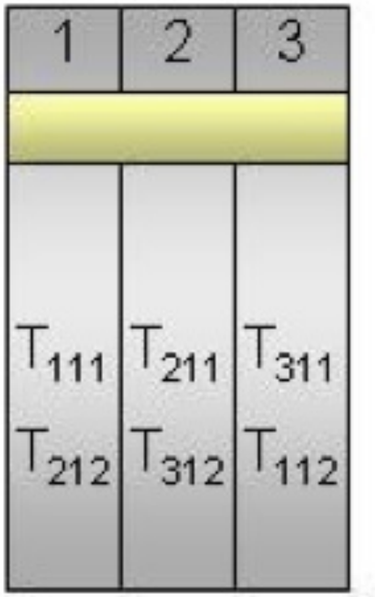
Blocking designs

Sampling

Replication

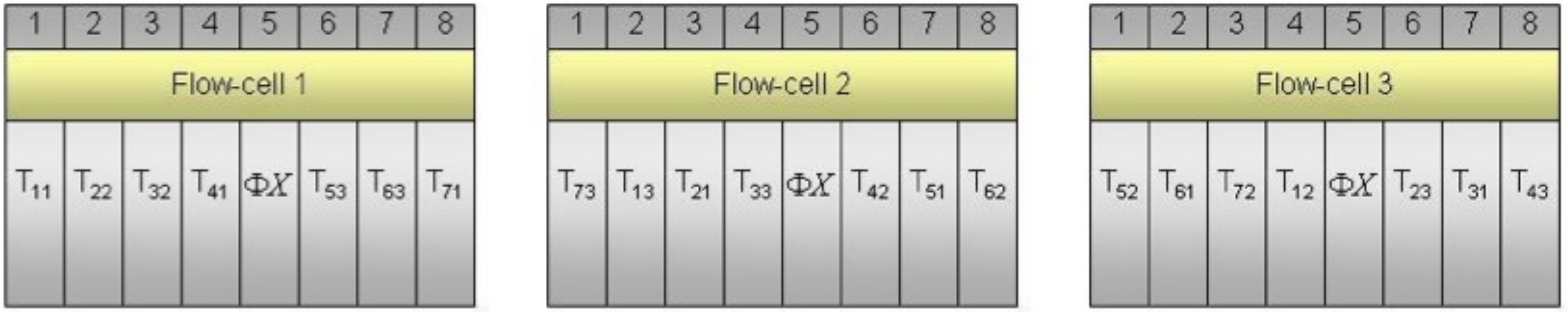
Blocking

Randomization



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

Let’s dissect these subscripts.


1 2 3 4 5 6 7 8
Flow-cell 1
T_{11} T_{22} T_{32} T_{41} ΦX T_{53} T_{63} T_{71}
1 2 3 4 5 6 7 8
Flow-cell 2
T_{73} T_{13} T_{21} T_{33} ΦX T_{42} T_{51} T_{62}
1 2 3 4 5 6 7 8
Flow-cell 3
T_{52} T_{61} T_{72} T_{12} ΦX T_{23} T_{31} T_{43}

Balanced for treatments across flow cells.. Randomized for location

Auer and Doerge 2010

What is your research question?:

What are the goals of your research?

Why did you generate all of the RNAseq data in the first place?

What can you use RNAseq data for?

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

- Transcriptome assembly.
- Improving genome assembly/annotation.
- SNP discovery (large genomes)
- Transcript discovery (variants for Transcription start site, alternative splicing, etc..)

- Differential expression analysis…

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 experiment should guide how you perform your experiment.

- 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 (see Meg's slides).

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.

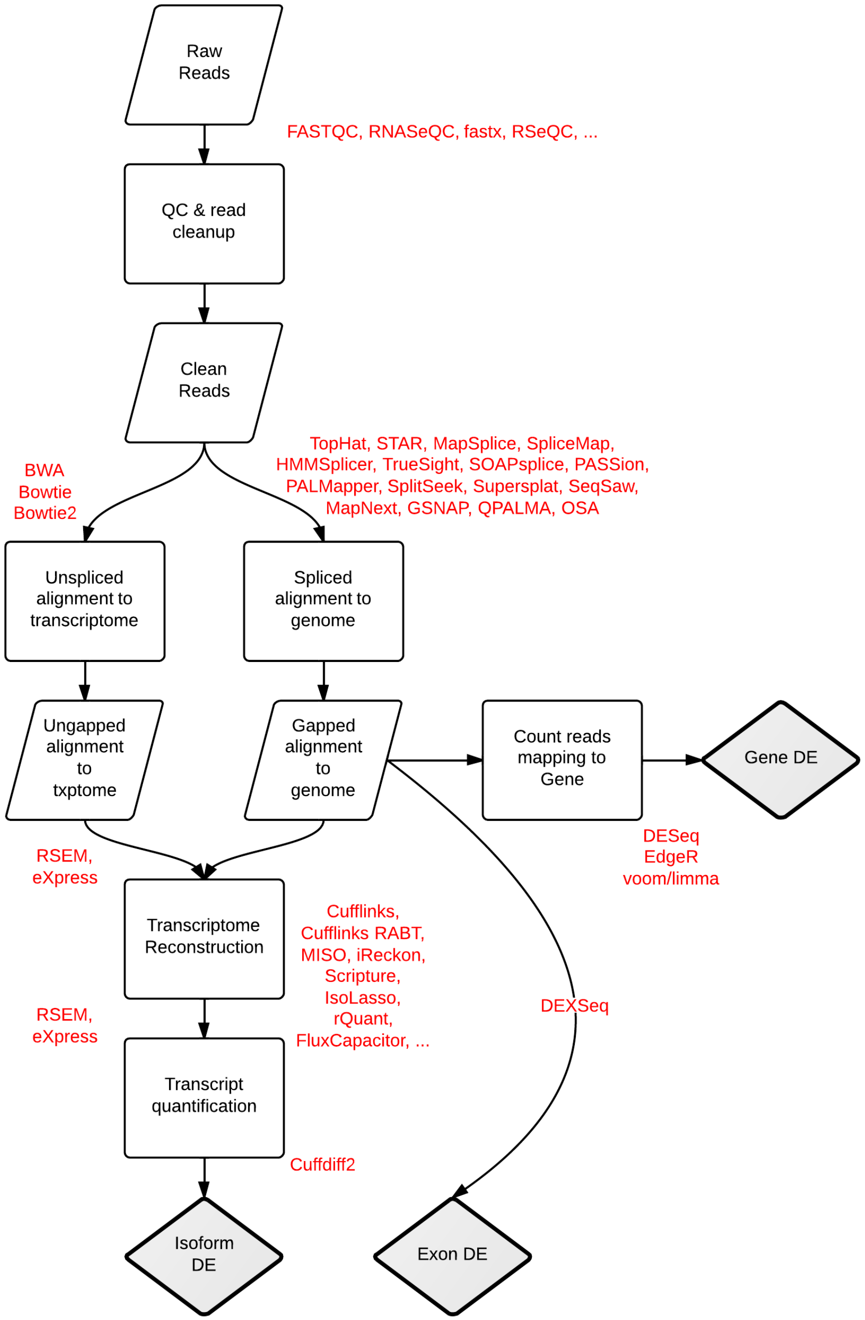
- Early work for RNA-seq tried to “mirror” the type of gene level analysis used in microarrays.
- However, RNA-seq has demonstrated how important it is to take into account alternative transcripts, even when attempting to get “gene level” measures.

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.
- Mapping reads to genome/transcriptome.
- Deal with alternative transcripts (new transcriptome)?
- Remap & count reads.
- Differential expression.



Raw Reads

FASTQC, RNASeQC, fastx, RSeQC, ...

QC & read cleanup

Clean Reads

BWA
Bowtie
Bowtie2

TopHat, STAR, MapSplice, SpliceMap, HMMSplicer, TrueSight, SOAPsplice, PASSion, PALMapper, SplitSeek, Supersplat, SeqSaw, MapNext, GSNAP, QPALMA, OSA

Unspliced alignment to transcriptome

Spliced alignment to genome

Ungapped alignment to txptome

Gapped alignment to genome

Count reads mapping to Gene

Gene DE

DESeq
EdgeR
voom/limma

RSEM, eXpress

Transcriptome Reconstruction

Cufflinks, Cufflinks RABT, MISO, iReckon, Scripture, IsoLasso, rQuant, FluxCapacitor, ...

DEXSeq

RSEM, eXpress

Transcript quantification

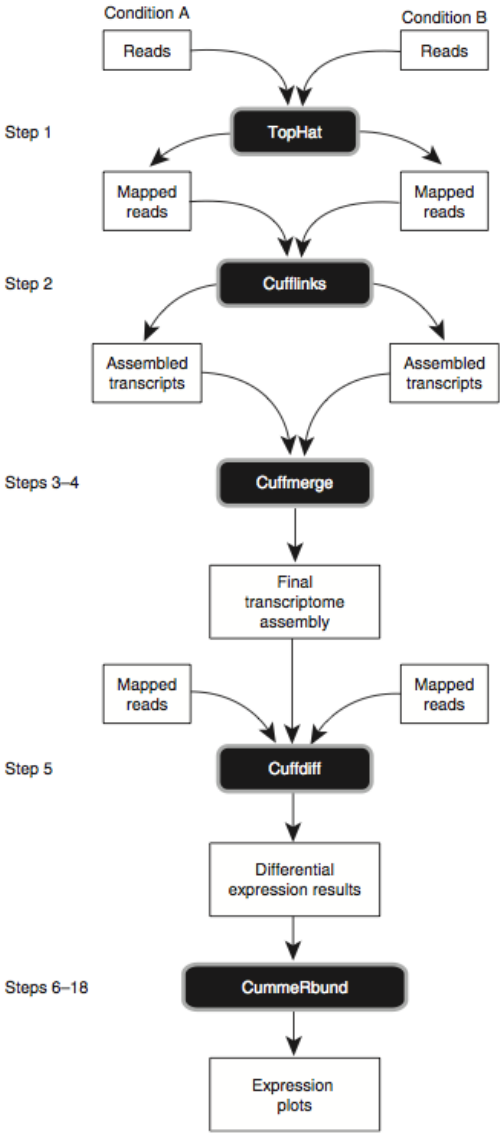
Cuffdiff2

Isoform DE

Exon DE

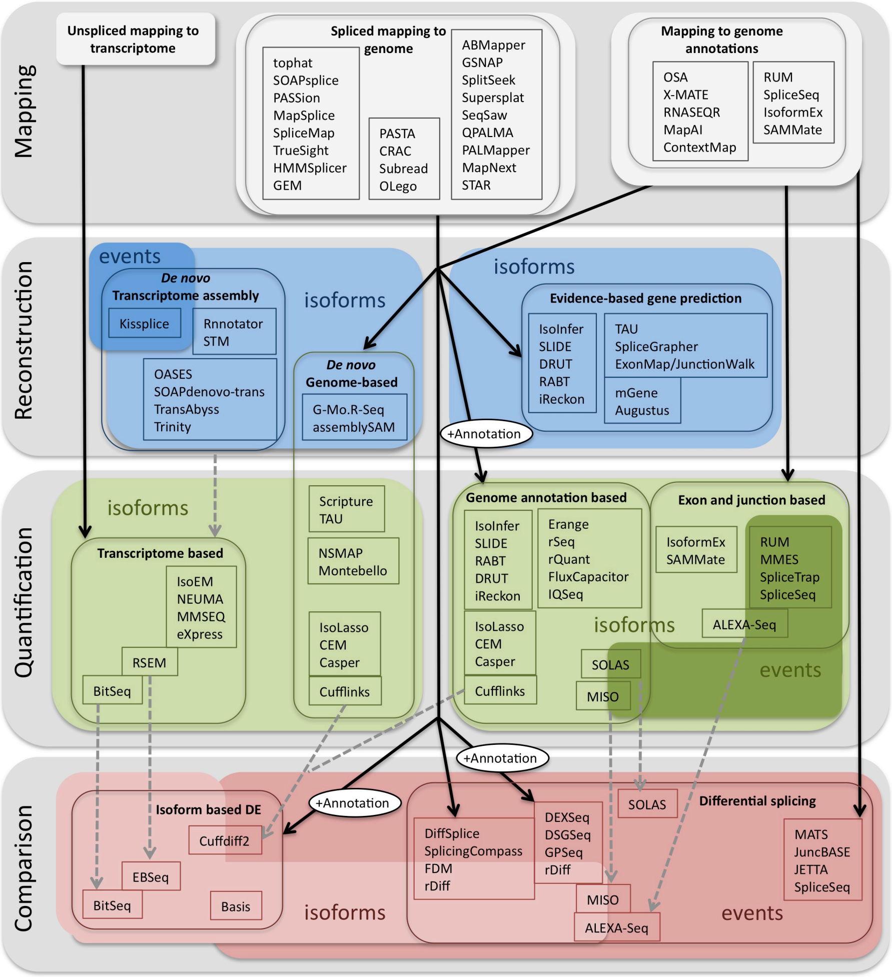
RNA-seq Workflows and Tools. Stephen Turner. Figshare. http://dx.doi.org/10.6084/m9.figshare.662782

The “tuxedo” protocol for RNA-seq



Trapnell et al 2012

Pipelines for RNA-seq (geared towards splicing)



Mapping

Unspliced mapping to transcriptome

Spliced mapping to genome

tophat
SOAPsplice
PASSion
MapSplice
SpliceMap
TrueSight
HMMSplicer
GEM

PASTA
CRAC
Subread
OLego

ABMapper
GSNAP
SplitSeek
Supersplat
SeqSaw
QPALMA
PALMapper
MapNext
STAR

Mapping to genome annotations

OSA
X-MATE
RNASEQR
MapAI
ContextMap

RUM
SpliceSeq
IsoformEx
SAMMate

Reconstruction

events

De novo Transcriptome assembly

Kissplice

Rnnotator
STM

OASES
SOAPdenovo-trans
TransAbyss
Trinity

isoforms

De novo Genome-based

G-Mo.R-Seq
assemblySAM

isoforms

Evidence-based gene prediction

IsoInfer
SLIDE
DRUT
RABT
iReckon

TAU
SpliceGrapher
ExonMap/JunctionWalk

mGene
Augustus

+Annotation

Quantification

isoforms

Transcriptome based

IsoEM
NEUMA
MMSEQ
eXpress

RSEM

BitSeq

Scripture
TAU

NSMAP
Montebello

IsoLasso
CEM
Casper

Cufflinks

Genome annotation based

IsoInfer
SLIDE
RABT
DRUT
iReckon

Erange
rSeq
rQuant
FluxCapacitor
IQSeq

IsoLasso
CEM
Casper

Cufflinks

isoforms

SOLAS

MISO

Exon and junction based

IsoformEx
SAMMate

RUM
MMES
SpliceTrap
SpliceSeq

ALEXA-Seq

events

+Annotation

Comparison

Isoform based DE

Cuffdiff2

EBSeq

BitSeq

Basis

+Annotation

isoforms

DiffSplice
SplicingCompass
FDM
rDiff

DEXSeq
DSGSeq
GPSeq
rDiff

SOLAS

MISO

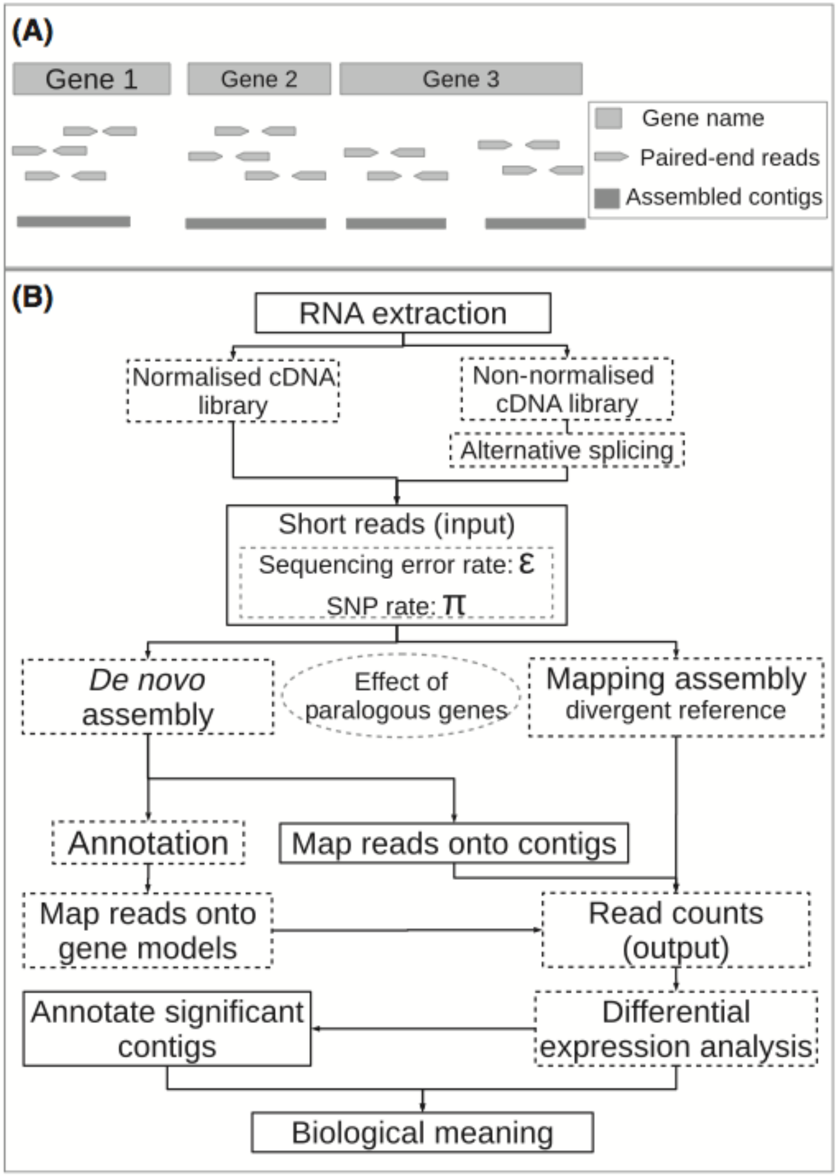
ALEXA-Seq

Differential splicing

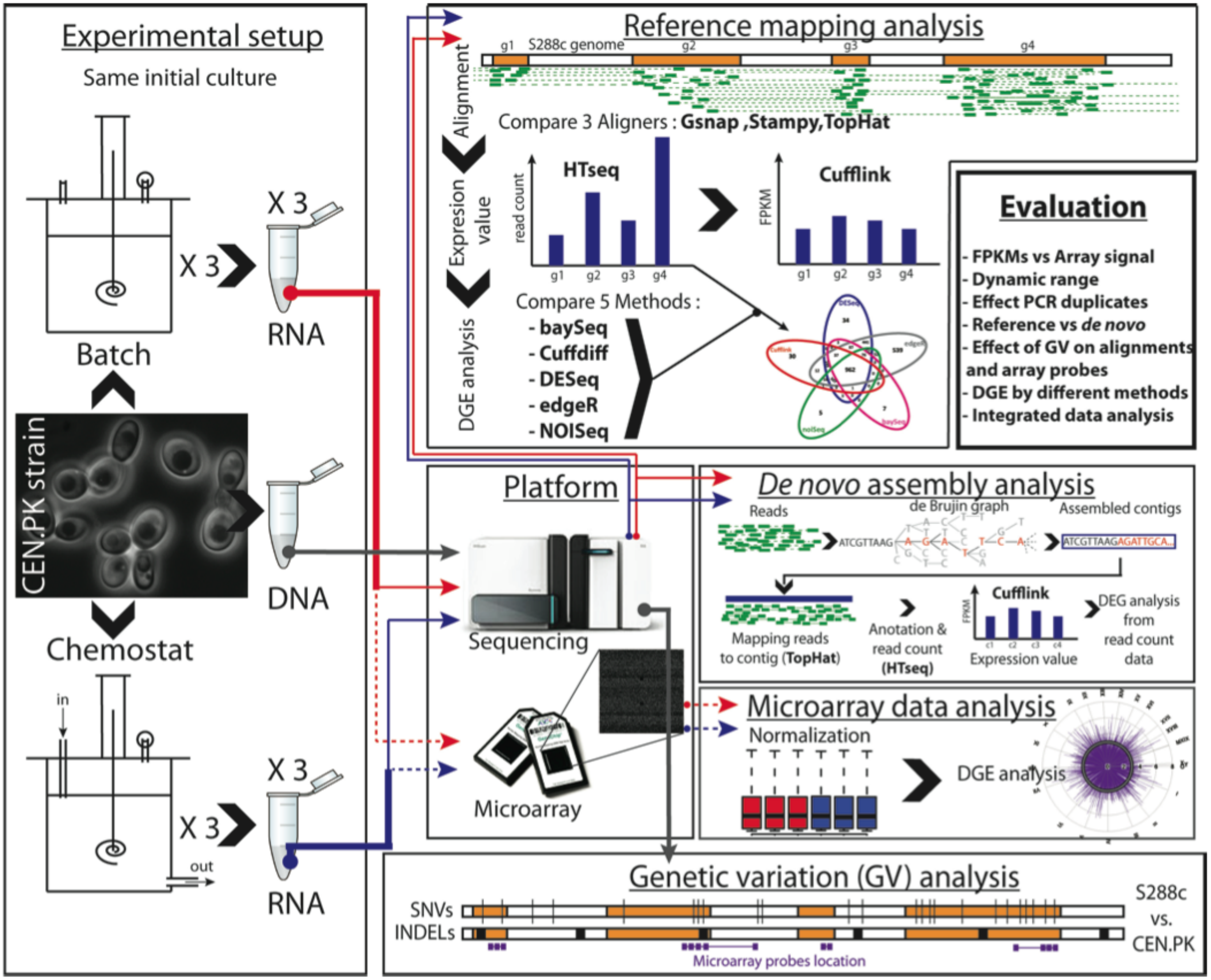
MATS
JuncBASE
JETTA
SpliceSeq

events

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



Nookaew et al 2102 NAR

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.

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 (Matt…)

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.

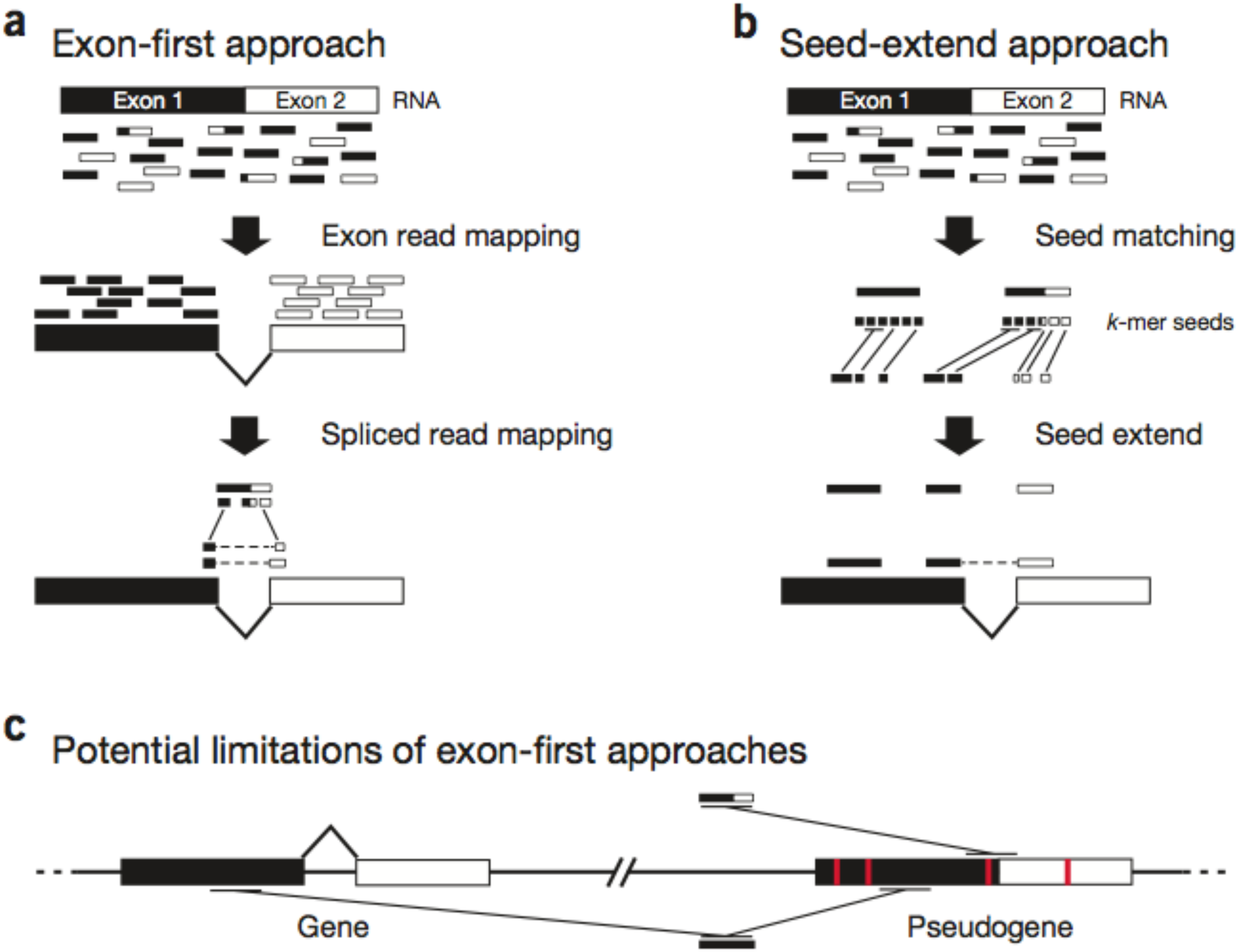
Mapping to the genome

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

"splicing aware" aligners:

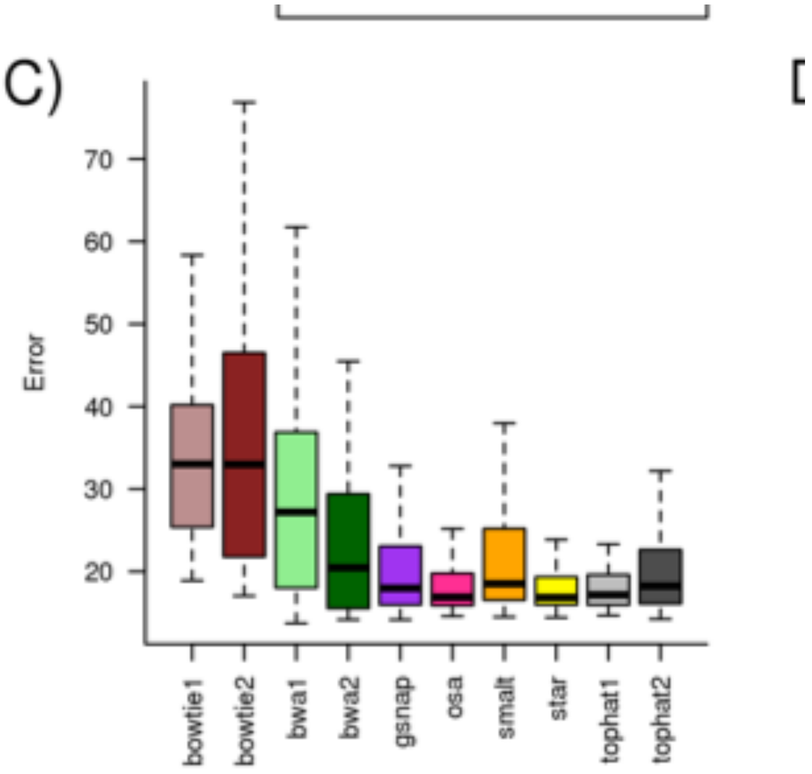
- Exon First: (tophat, MapSplice, SpliceMap) Fig1A Garber
- Step 1 - map reads to genome
- Step 2 -unmapped reads are split, and aligned.

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



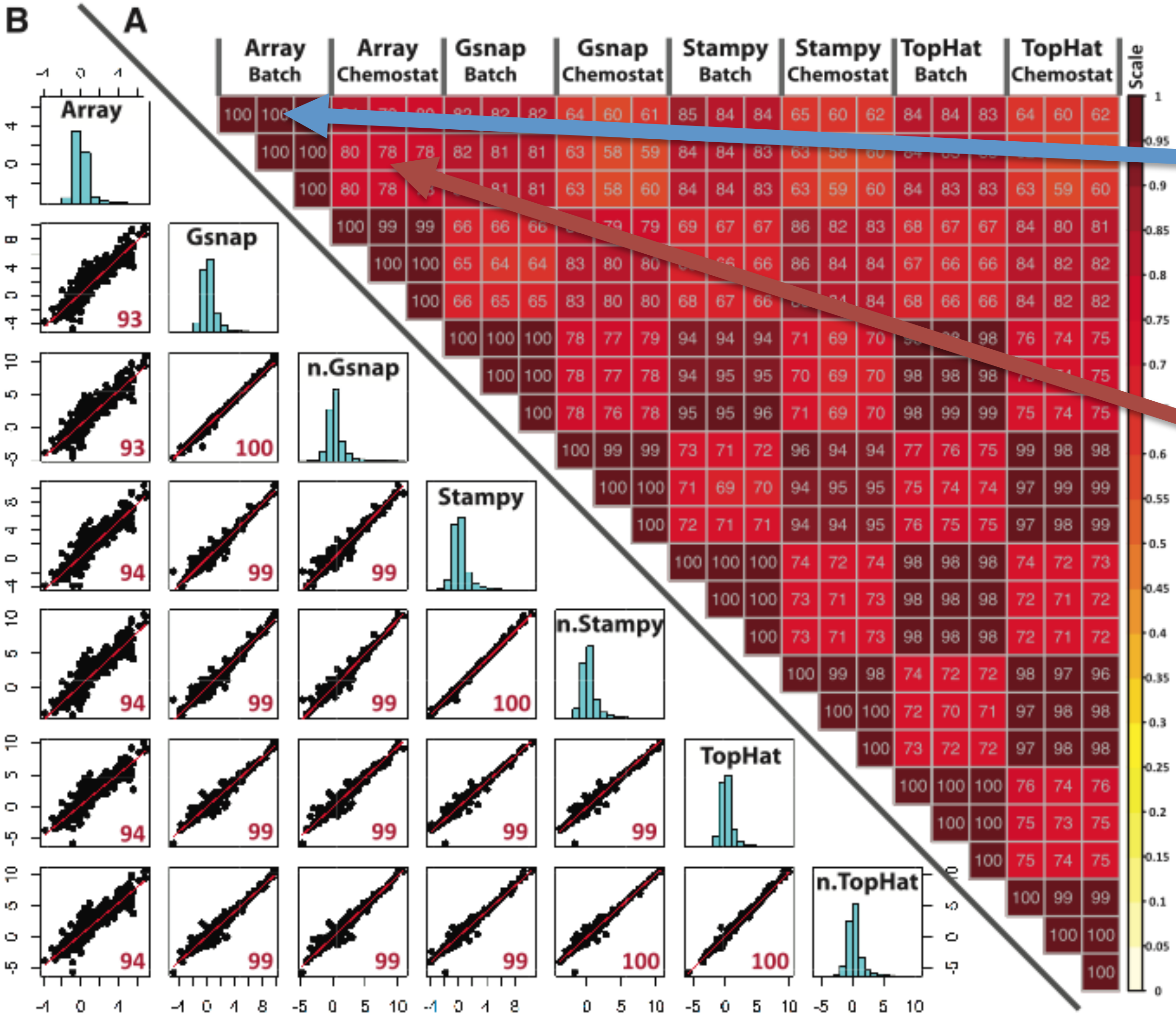
Garber et al. 2011

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?



Array Batch			Array Chemostat			Gsnap Batch			Gsnap Chemostat			Stampy Batch			Stampy Chemostat			TopHat Batch			TopHat Chemostat		
100	100	[illegible]	81	79	80	82	82	82	64	60	61	85	84	84	65	60	62	84	84	83	64	60	62
	100	100	80	78	78	82	81	81	63	58	59	84	84	83	63	58	60	84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		100	80	78	78	[illegible]	81	81	63	58	60	84	84	83	63	59	60	84	83	83	63	59	60
			100	99	99	66	66	66	[illegible]	79	79	69	67	67	86	82	83	68	67	67	84	80	81
				100	100	65	64	64	83	80	80	[illegible]	66	66	86	84	84	67	66	66	84	82	82
					100	66	65	65	83	80	80	68	67	66	[illegible]	84	84	68	66	66	84	82	82
						100	100	100	78	77	79	94	94	94	71	69	70	[illegible]	98	98	76	74	75
							100	100	78	77	78	94	95	95	70	69	70	98	98	98	75	[illegible]	75
								100	78	76	78	95	95	96	71	69	70	98	99	99	75	74	75
									100	99	99	73	71	72	96	94	94	77	76	75	99	98	98
										100	100	71	69	70	94	95	95	75	74	74	97	99	99
											100	72	71	71	94	94	95	76	75	75	97	98	99
												100	100	100	74	72	74	98	98	98	74	72	73
													100	100	73	71	73	98	98	98	72	71	72
														100	73	71	73	98	98	98	72	71	72
															100	99	98	74	72	72	98	97	96
																100	100	72	70	71	97	98	98
																	100	73	72	72	97	98	98
																		100	100	100	76	74	76
																			100	100	75	73	75
																				100	75	74	75
																					100	99	99
																						100	100
																							100

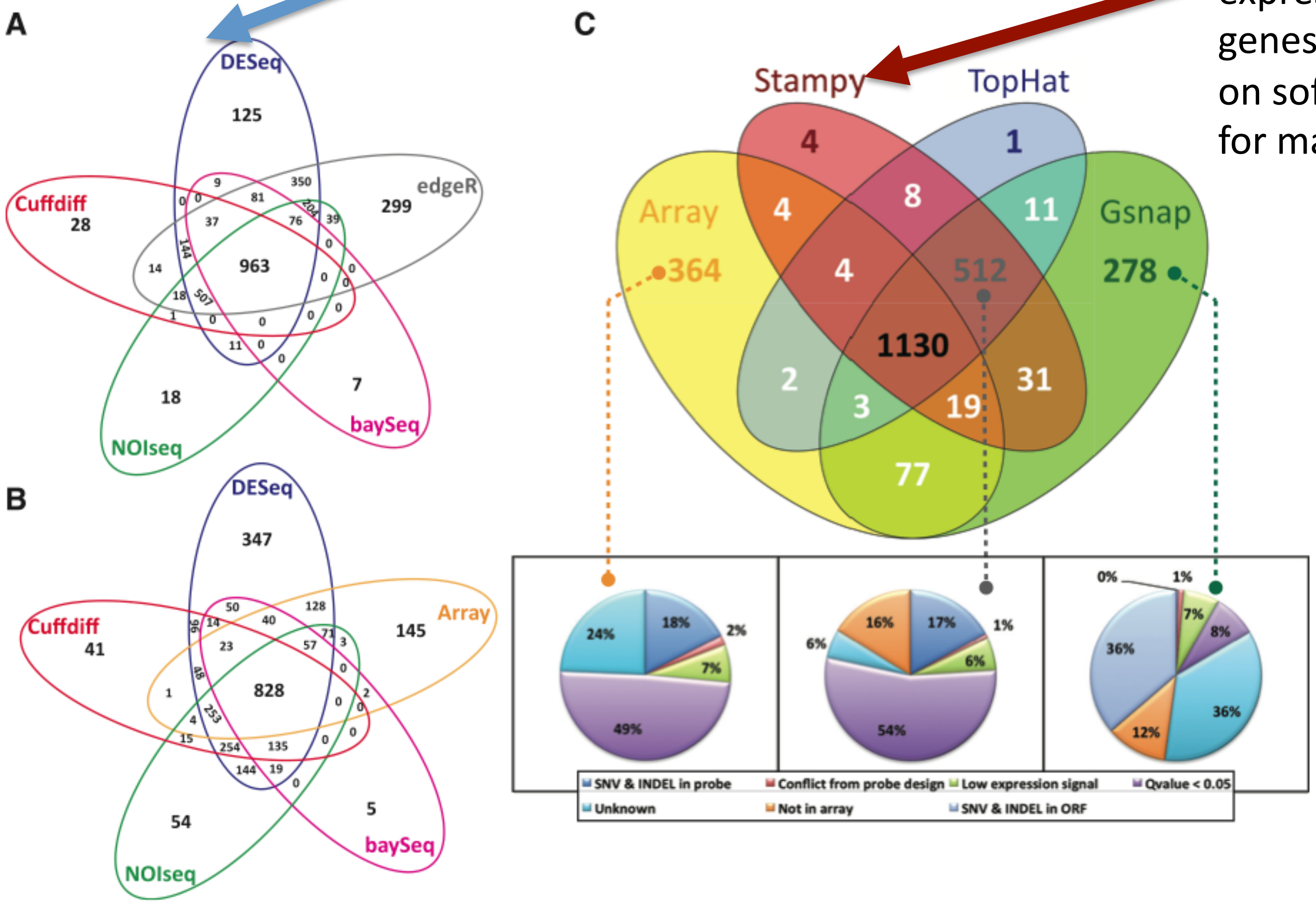
These cells are biological replicates (diagonals)

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)

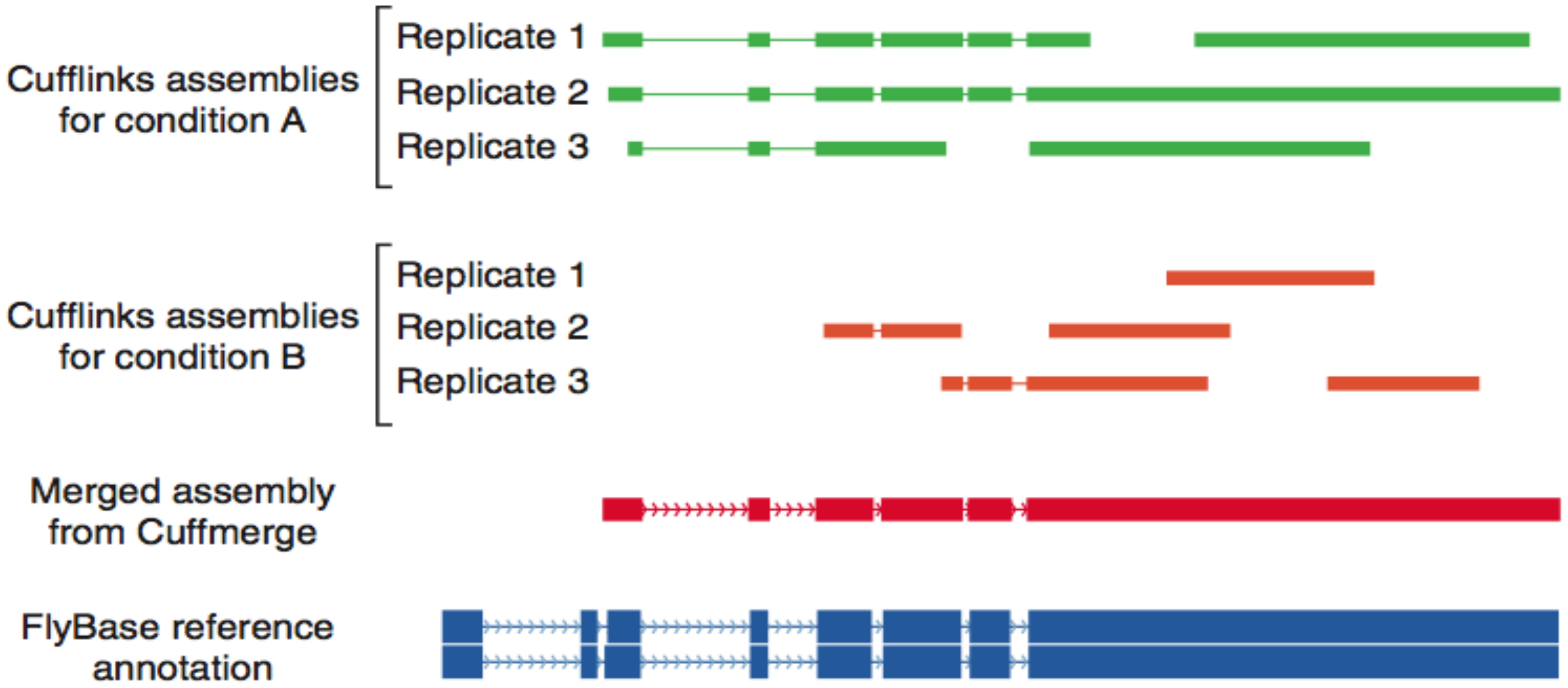
- BW based aligners have reduced performance once mismatches are considered.
  - Exponential decrease in performance with each additional mismatch (iteratively performs perfect searches).
  - Seed methods may be more sensitive when mapping to transcriptomes of distantly related species (or high polymorphism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

From Garber et al. 2011

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

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

Merging all transcripts?


Cufflinks assemblies for condition A
Replicate 1
Replicate 2
Replicate 3
Cufflinks assemblies for condition B
Replicate 1
Replicate 2
Replicate 3
Merged assembly from Cuffmerge
FlyBase reference annotation

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?

Counting

Counting

- What are we trying to count?
- Gene level measure (eXpress, corset, RSEM, HTSeq,…).
- Exon level (HTSeq, ???)
- Transcript level (HTSeq, Cufflinks, ….)

Counting

- We are interested in transcript abundance.
- But we need to take into account a number of things.
- How many reads in the sample.
- Length of transcripts
- GC content and sequencing bias
- (how many transcripts)

Old ways of counting Counting

- RPKM (reads aligned per kilobase of exon per million reads mapped) – Mortazavi et al 2008
- FPKM (fragments per kilobase of exon per million fragments mapped). Same idea for paired end sequencing.
- Transcripts per million (we will come to that).

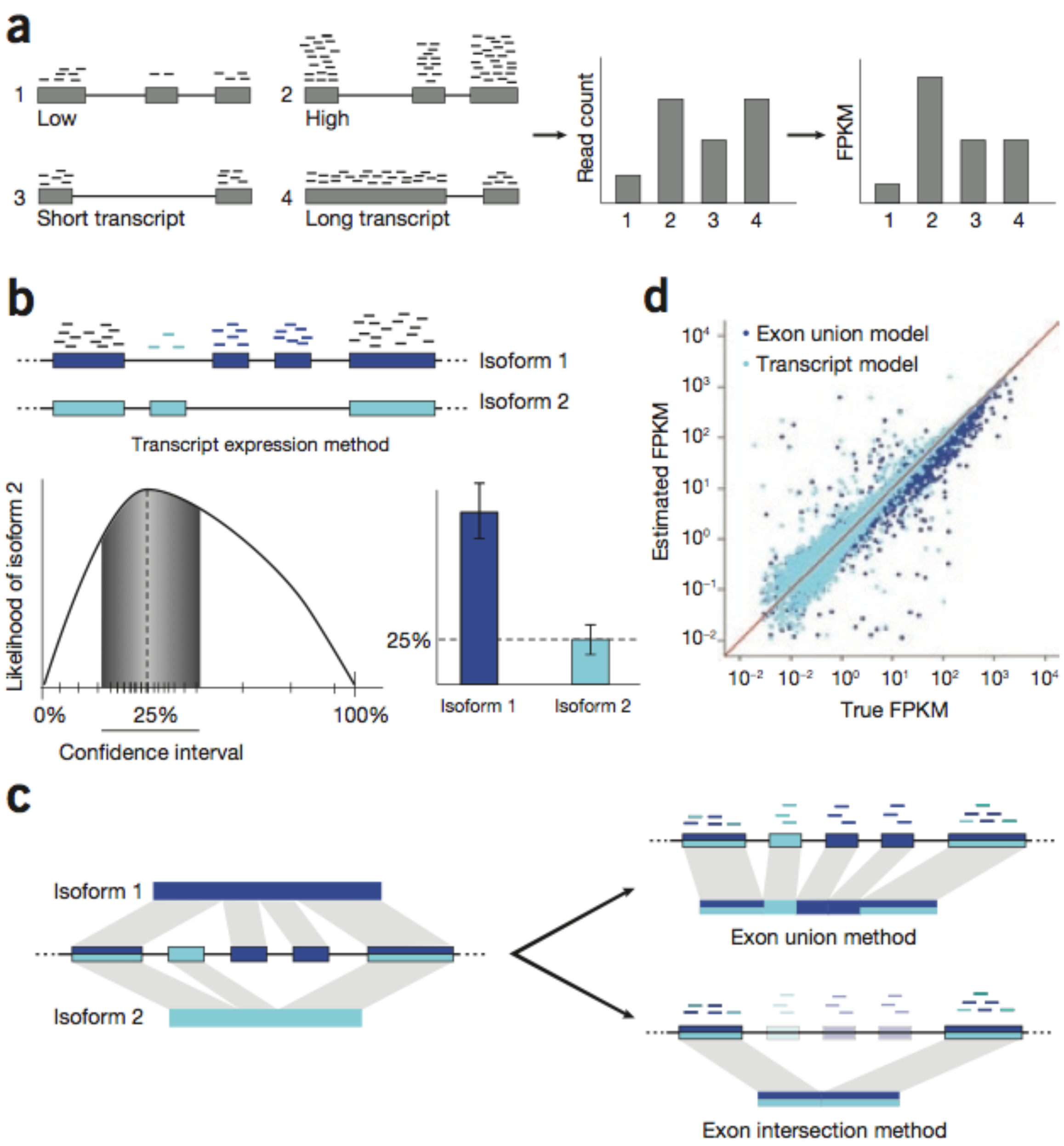
None of these measures are great for differential expression analysis.

- For appropriate differential expression analysis (as with all statistical modeling), keeping all of the data is better.
- So having counts of mapped reads, along with information like GC content, transcript length, total # reads is far more useful.
- We will discuss this tomorrow.

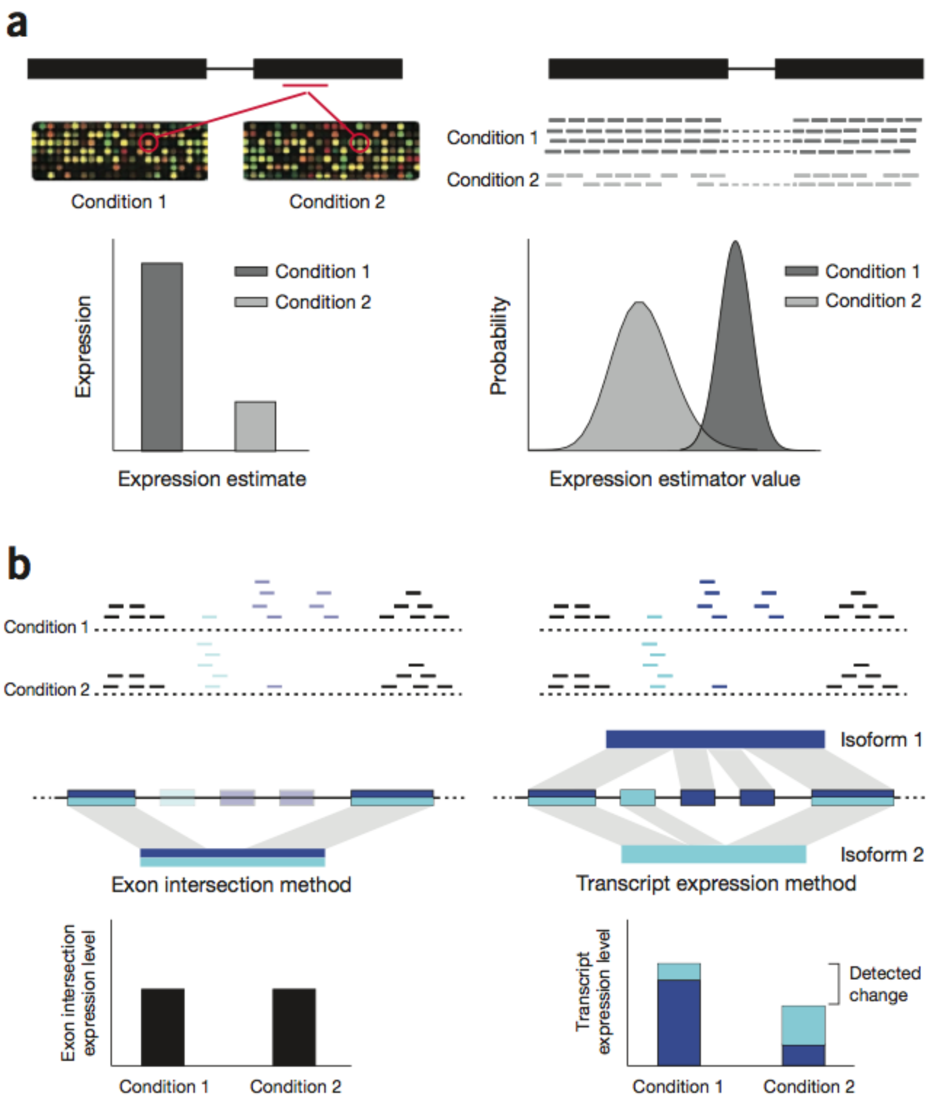
Accounting for multiple isoforms (when counting alternative transcripts).

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

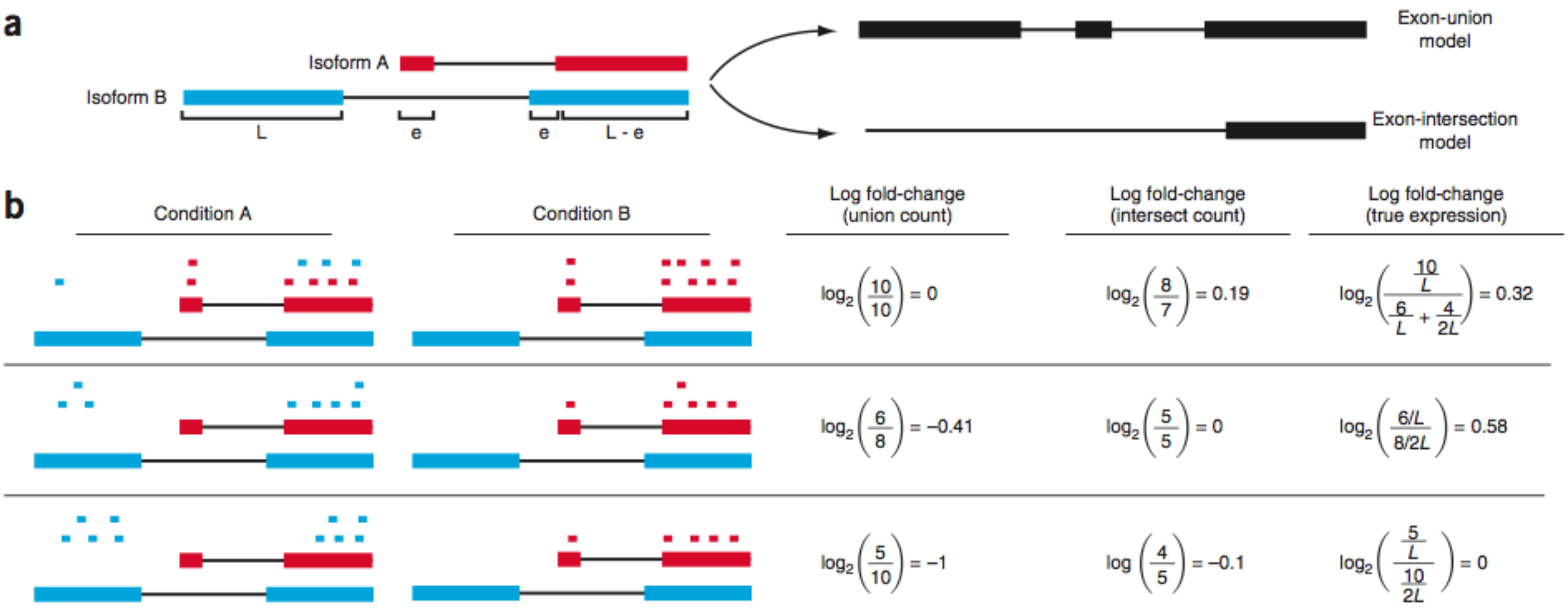
- - so called "isoform-expression" methods (cufflinks, MISO) model the uncertainty parametrically (often using MLE). The model with the best mix of isoforms that models the data (highest joint probability) is the best estimate. How this is handled differs a great deal by the different.



Garber et al. 2011



Garber et al. 2011


a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Trapnell et al 2013

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/

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…

Differential expression

- DEseq (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)

Example workflows

- 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