General considerations for RNA-seq quantification for differential expression and or splice variant assessment

What are the goals of your research? Why did you generate all of the RNAseq data in the first place.

- RNA-seq is generated for a number of reasons
- Transcriptome assembly (& SNP discovery)
- Transcript discovery (variants for Transcription start site, alternative splicing, etc..)
- Quantification of (alternative transcripts)
- Differential expression analysis across treatments.

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

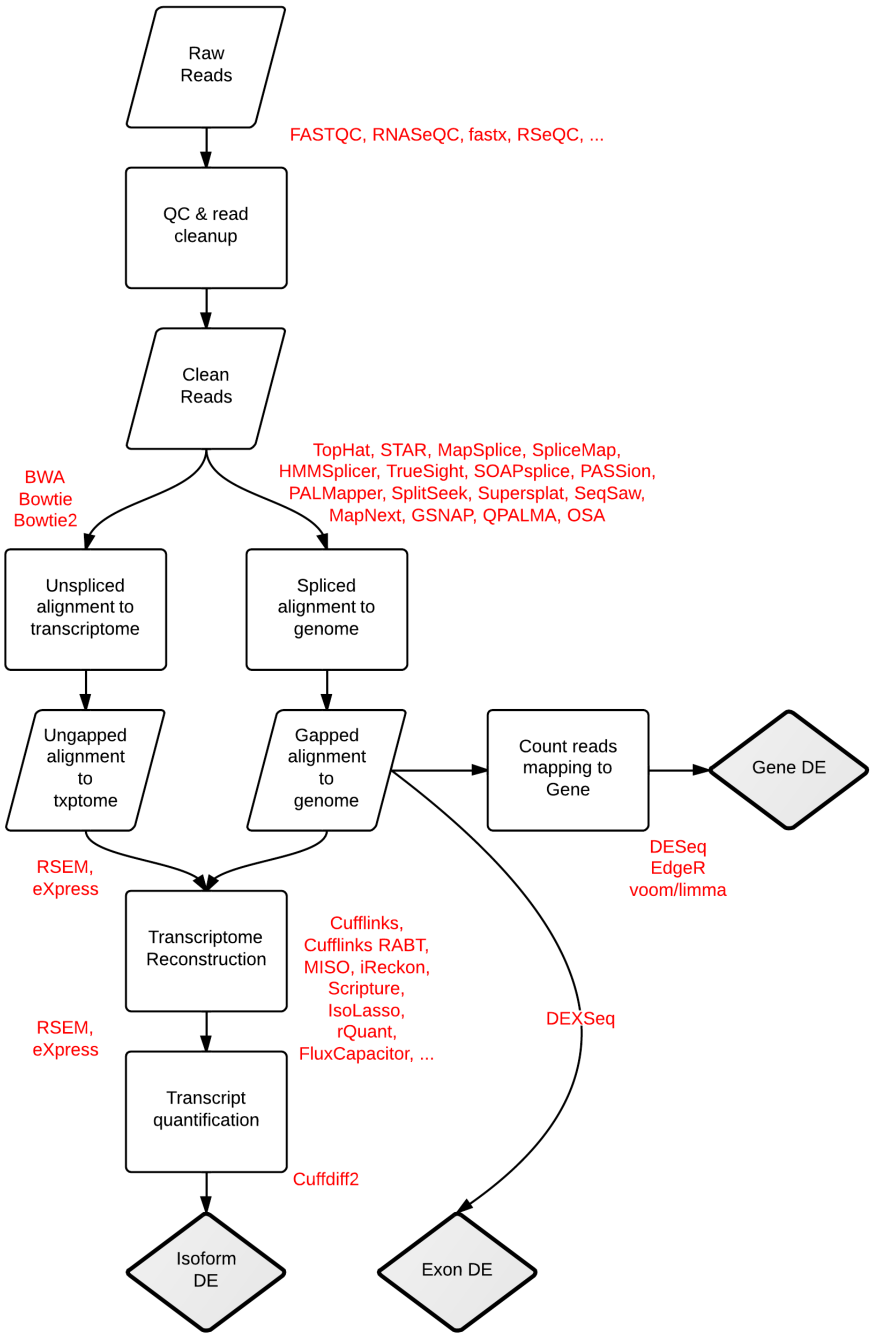
- Early work for RNA-seq tried to “mirror” the type of gene level analysis used in microarrays.
- However, RNA-seq has demonstrated how important it is to take into account alternative transcripts, even when attempting to get “gene level” measures.

How do we put together a useful pipeline for RNAseq

- What are the steps we need to consider?

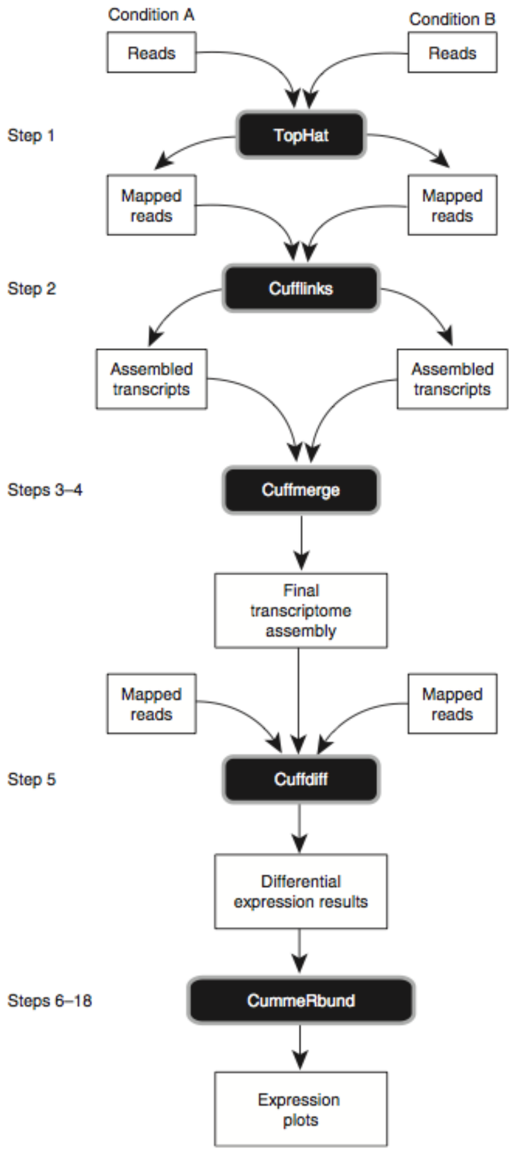
How do we put together a useful pipeline for RNAseq

- 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



Condition A

Condition B

Reads

Reads

Step 1

TopHat

Mapped reads

Mapped reads

Step 2

Cufflinks

Assembled transcripts

Assembled transcripts

Steps 3–4

Cuffmerge

Final transcriptome assembly

Mapped reads

Mapped reads

Step 5

Cuffdiff

Differential expression results

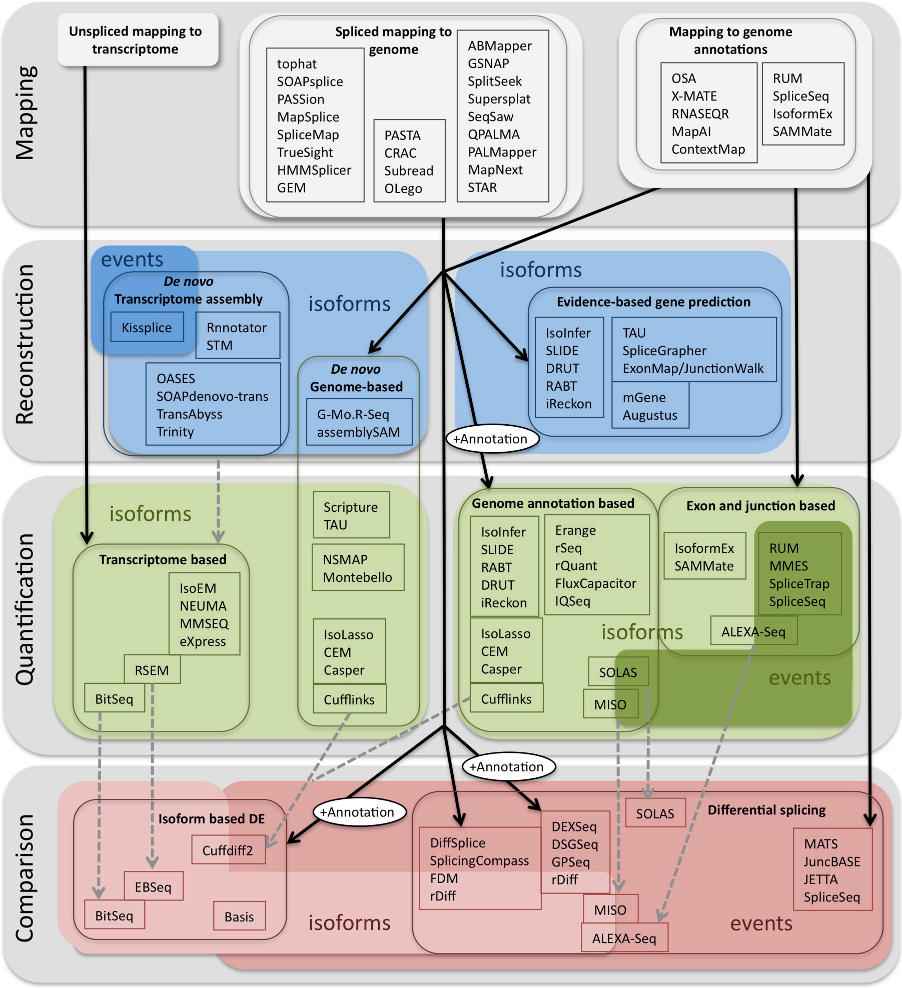
Steps 6–18

CummeRbund

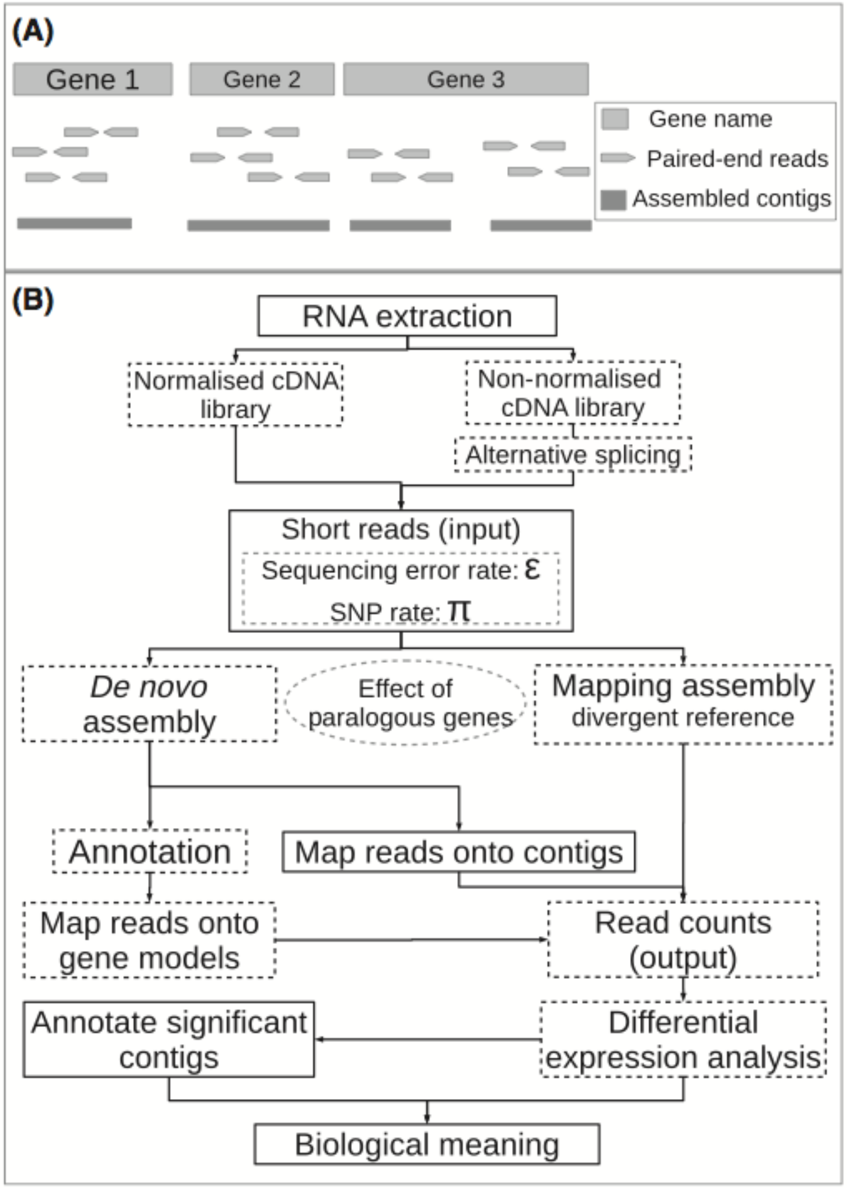
Expression plots

Trapnell et al 2012

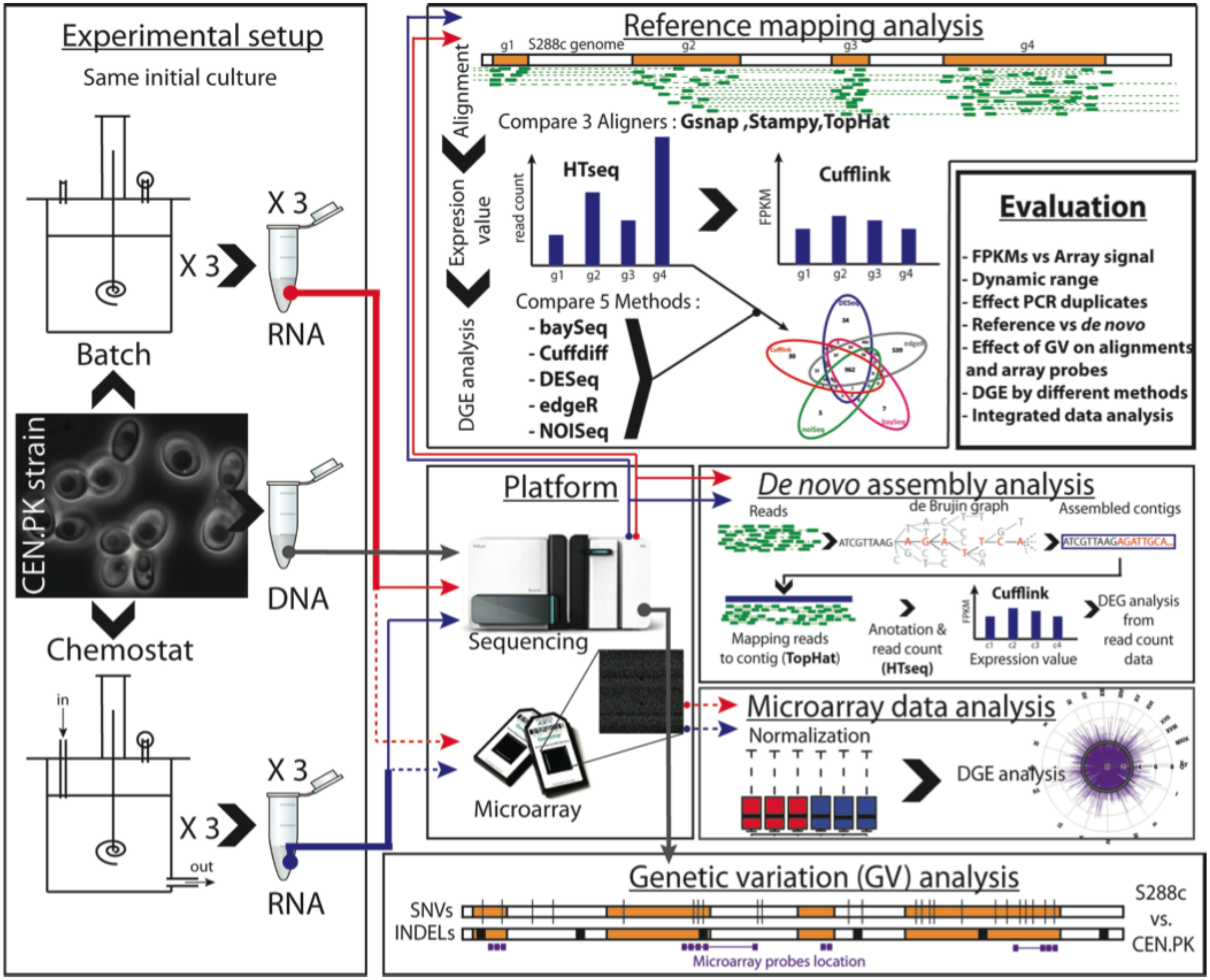
Pipelines for RNA-seq (geared towards splicing)



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



Nookaew et al 2102 NAR

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.

Mapping to a transcriptome

- What might be the downside to mapping to the 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.

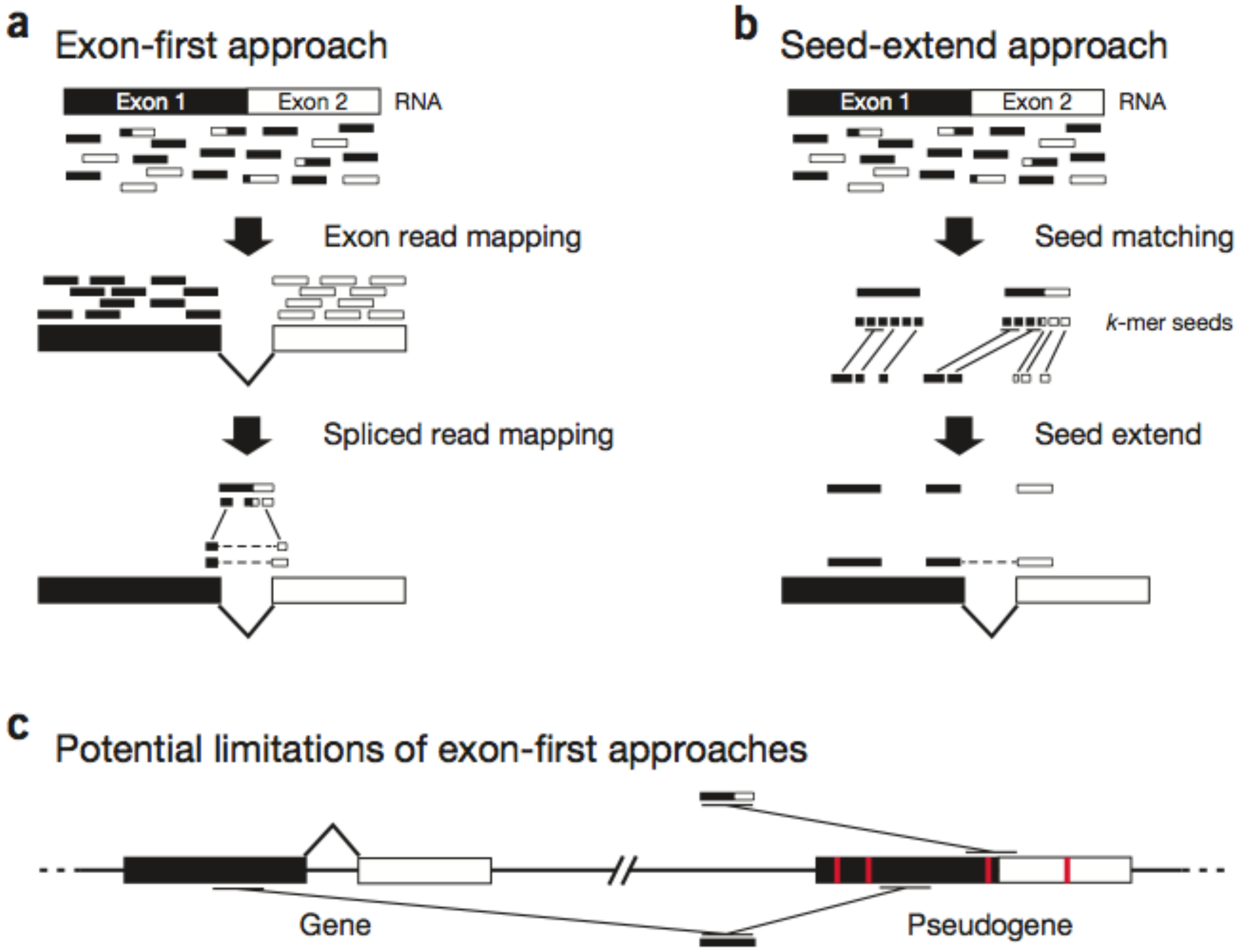
Mapping to the genome

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

"splicing aware" aligners:

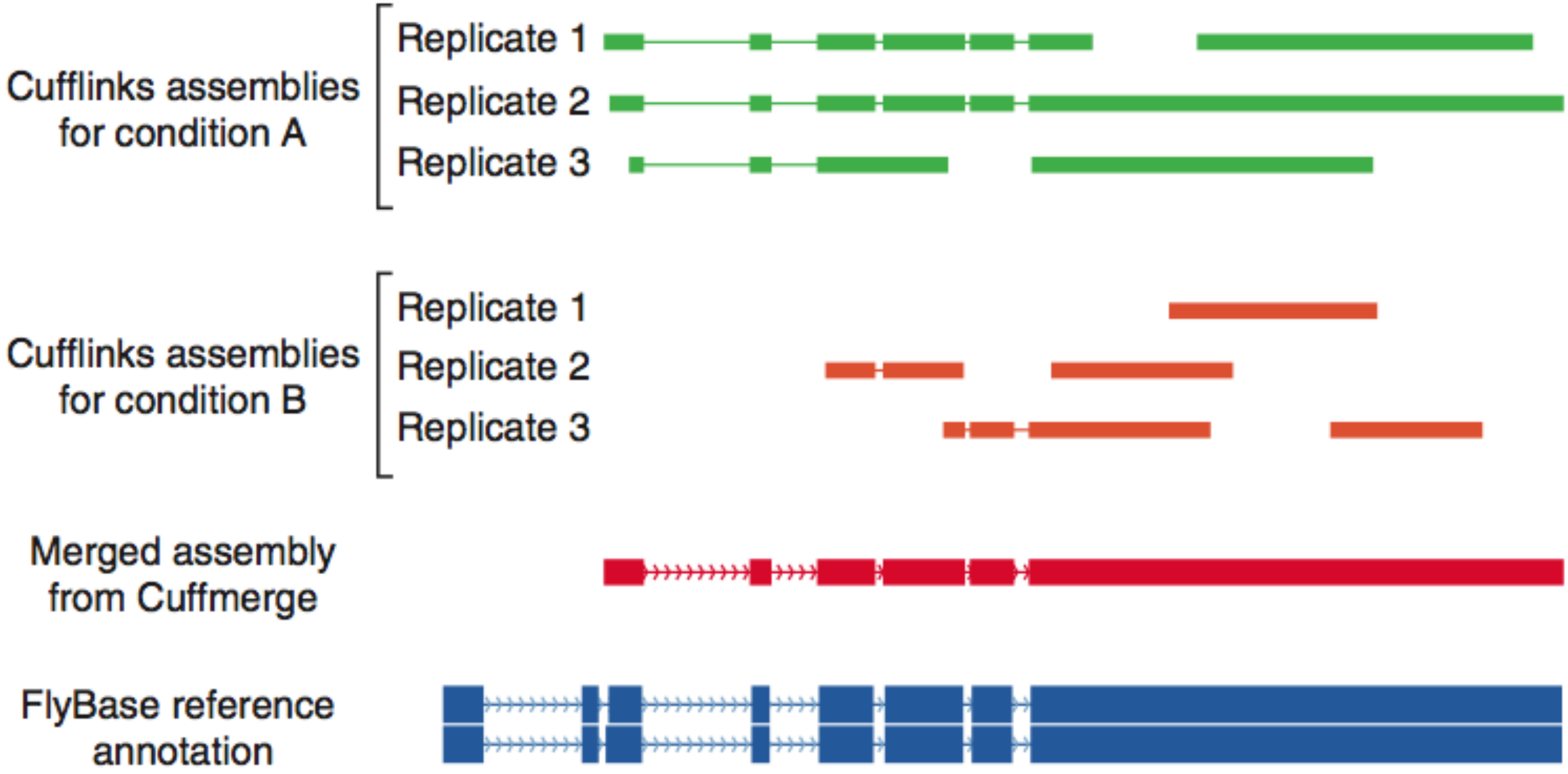
  - Exon First: (tophat, MapSplice, SpliceMap) Fig1A Garber
  - Step 1 - map reads to genome
  - Step 2 -unmapped reads are split, and aligned.

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



Garber et al. 2011

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

- 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

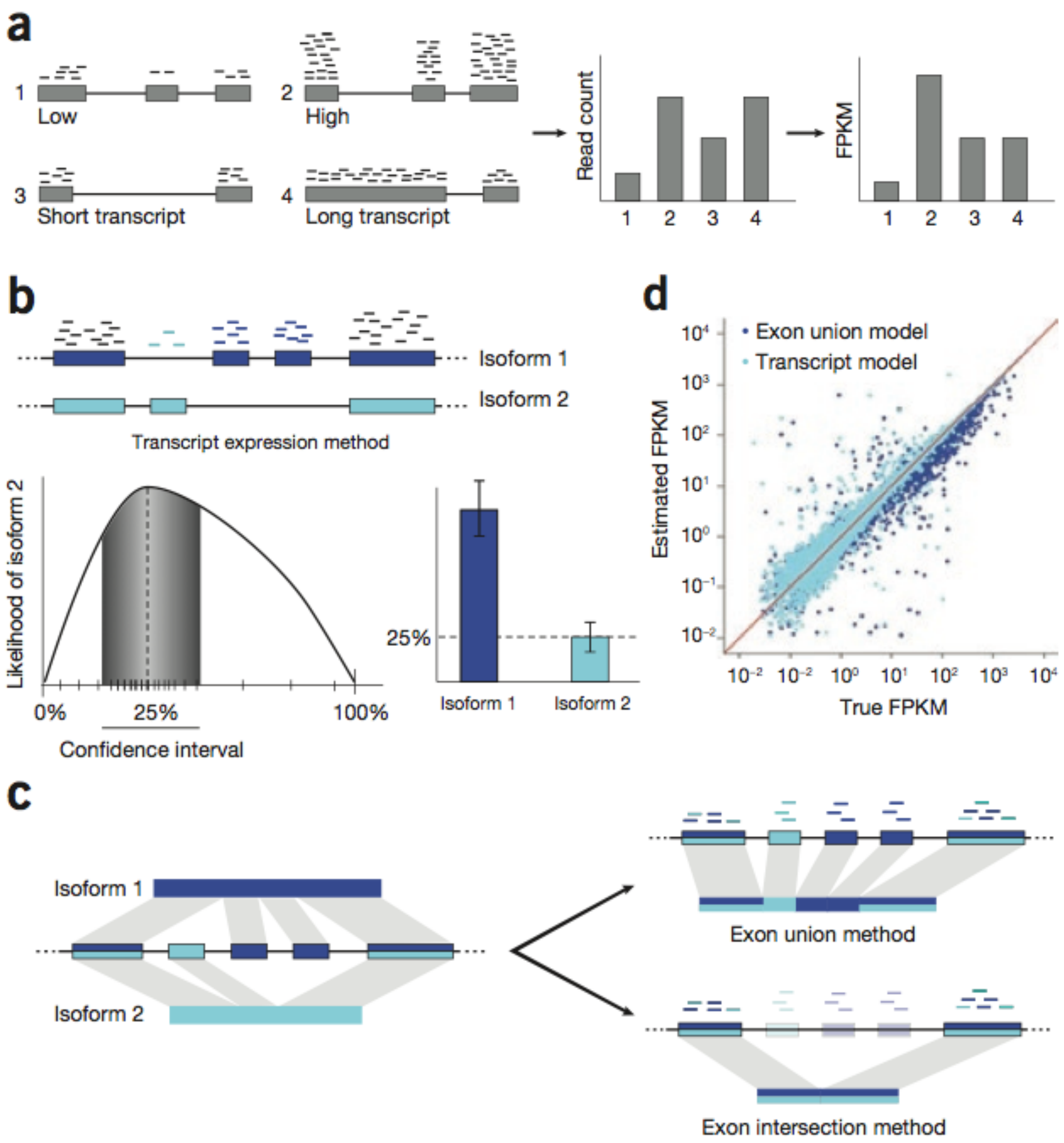
Counting

- 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.

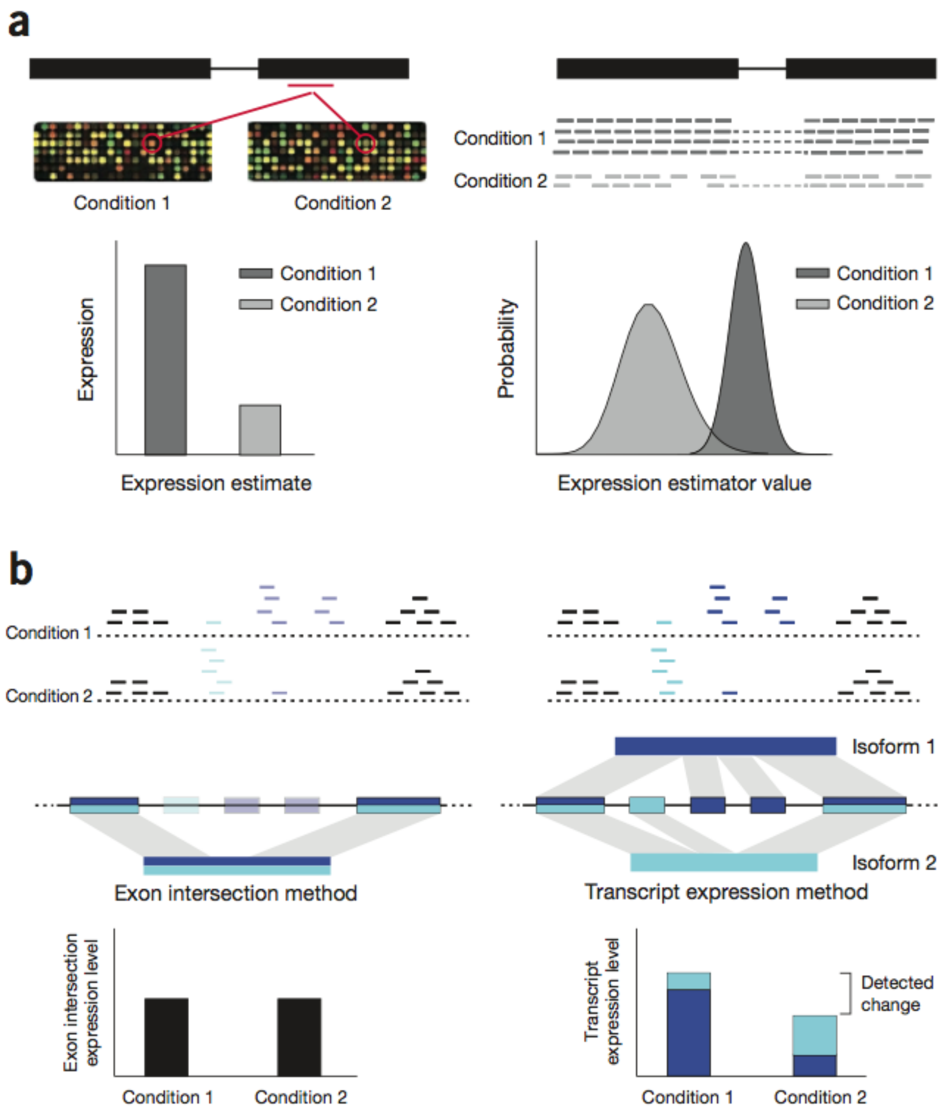
Accounting for multiple isoforms.

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

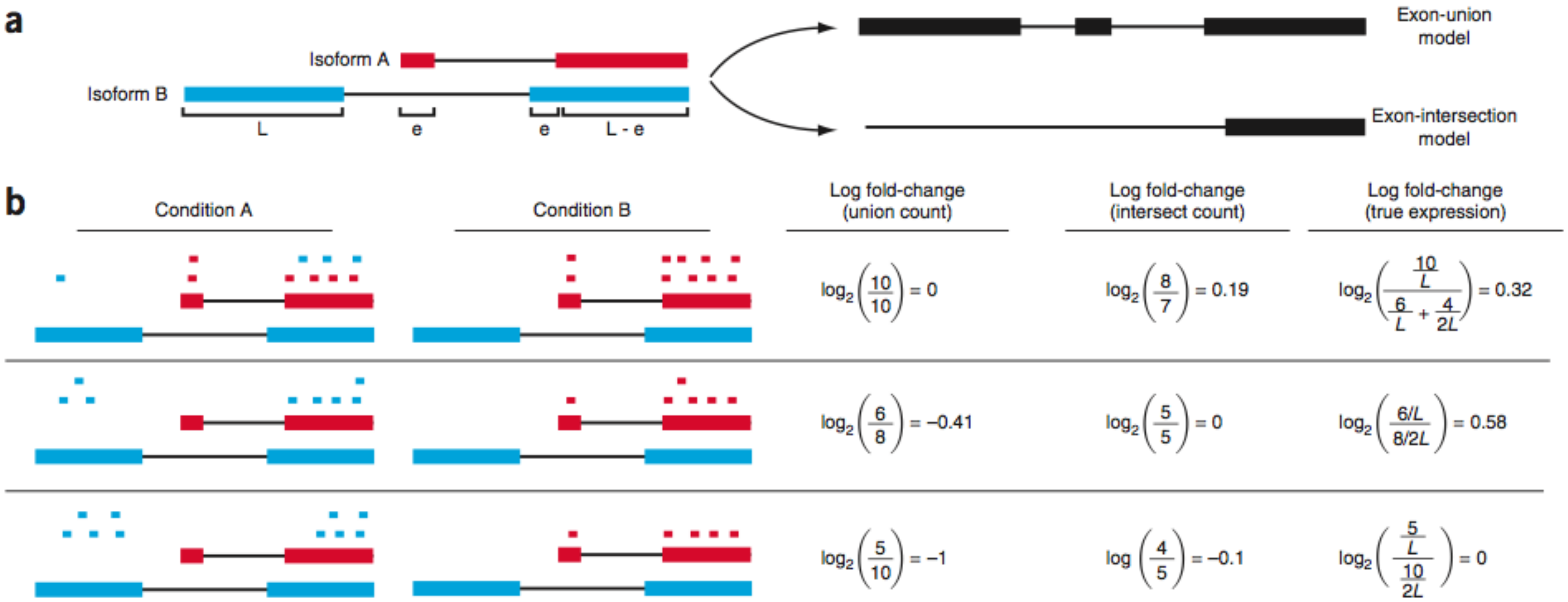
- - so called "isoform-expression" methods (cufflinks, MISO) model the uncertainty parametrically (often using MLE). The model with the best mix of isoforms that models the data (highest joint probability) is the best estimate. How this is handled differs a great deal by the different



Garber et al. 2011

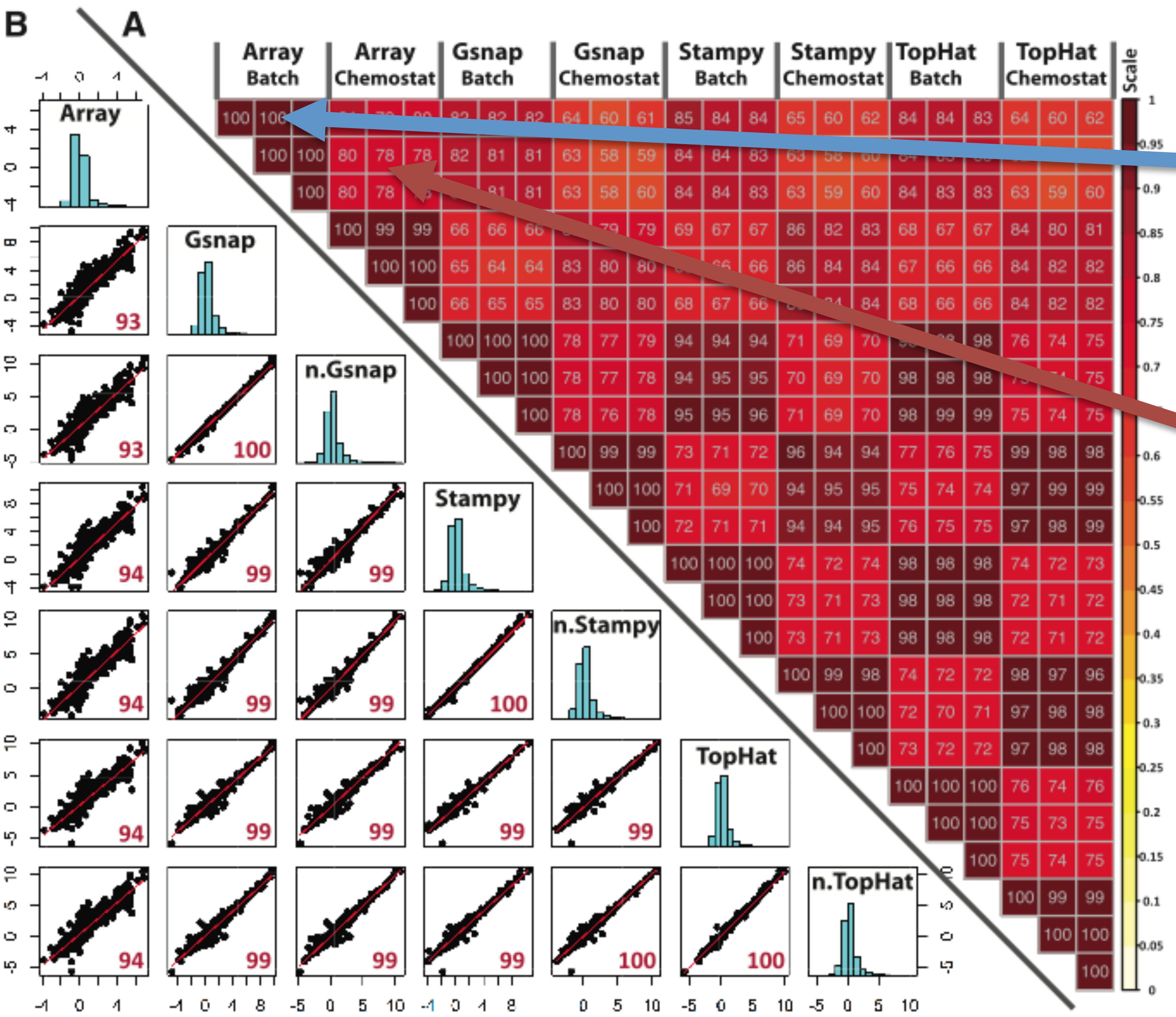


Garber et al. 2011


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

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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		100	80	78	[illegible]	[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	[illegible]	74	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

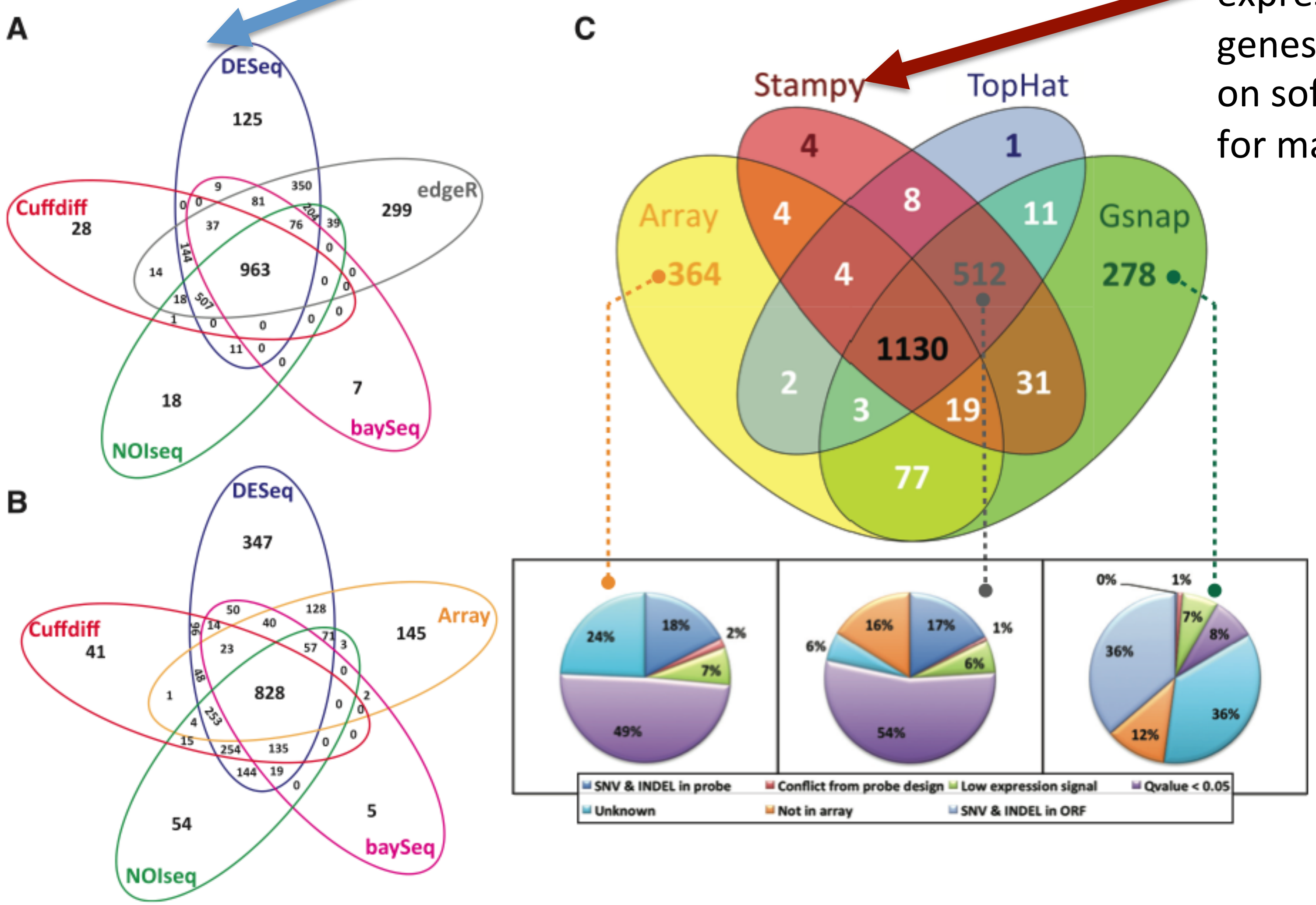
These cells are biological replicates (diagonals)

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

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 reads

- Htseq (python library) works with DEseq

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)

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