Transcriptomics

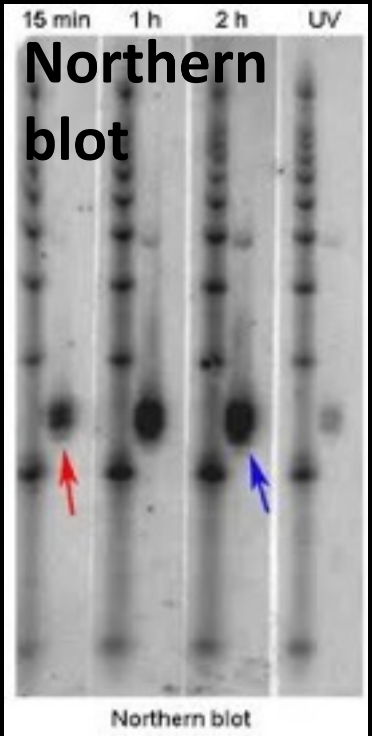
Katie Pollard
Alisha Holloway

BMI 206
docpollard.org/bmi206
October 10, 2016

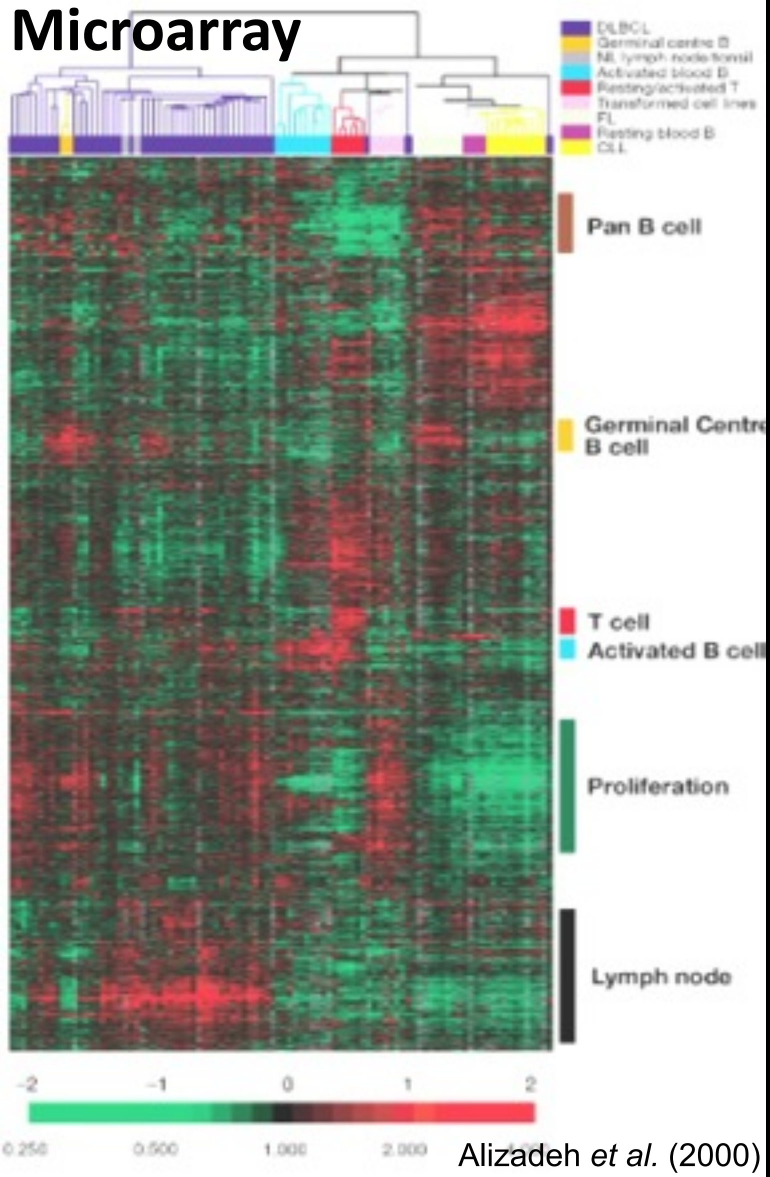
Measuring gene expression

PCR

Northern blot

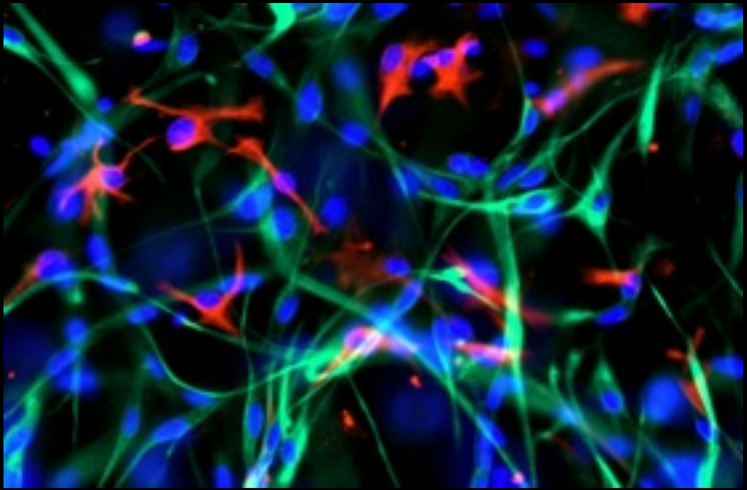


Microarray



Alizadeh *et al.* (2000)

Imaging

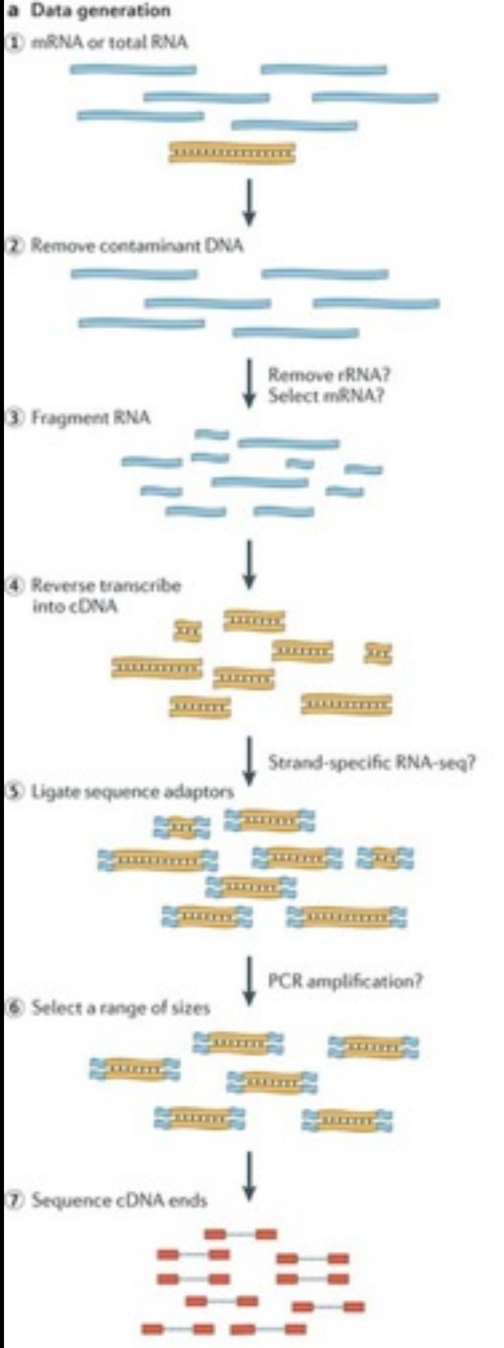


RNA-seq

Use of high-throughput sequencing technologies to assess the RNA content of a sample.

Number of sequence reads that map to a transcript is a measure of expression level.

Count data!



Martin & Wang 2011

Applications

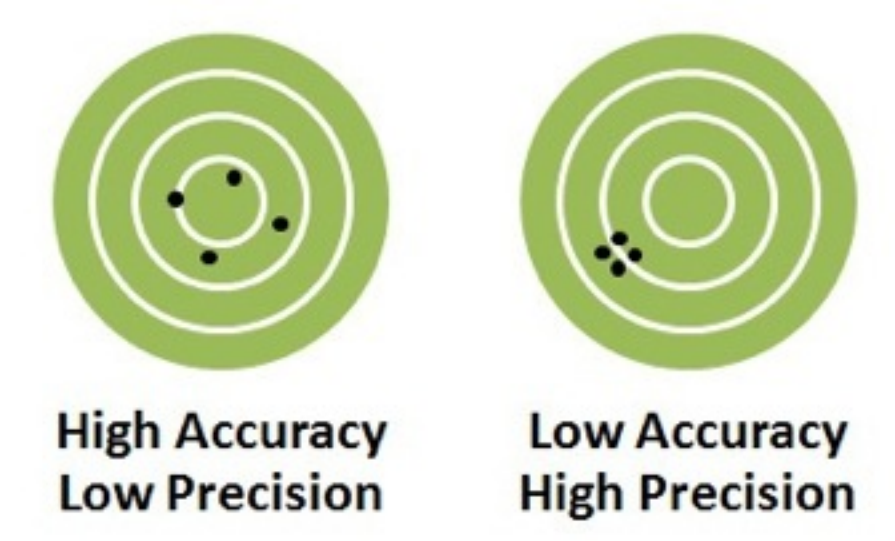
- Profile transcriptome-wide expression patterns
- Assess allele-specific expression
- Quantify alternative transcript usage
- Discover novel genes/transcripts, gene fusions
- Identify RNA-editing events
- Ribosome profiling to measure translation
- Massively parallel reporter assays to measure transcription from candidate enhancers

Variants of RNA-seq

- Paired end sequencing
- GRO-seq (to measure rate of transcription)
- CAGE (5’ ends of transcripts)
- Small RNA sequencing (need to enrich to see them)
- Single cell RNA-seq

Experimental choices

- Study design
  - Biological replicates
  - Reference genome?
  - Good gene annotation?
- Read depth
- Barcoding
- Read length
- Paired vs. single-end



Need biological replicates to measure accuracy
Technical replicates measure precision

How many reads needed?

Human Transcriptomics:

- ~15-20K genes expressed in a tissue or cell line.
- Genes are on average 3KB
- For 1x coverage using 100 bp reads, would need 600K sequence reads (on average)
- In reality, we need MUCH higher coverage to accurately estimate gene expression levels.
- **30-50 million reads**

Analysis pipelines

- QC
- Alignment (or kallisto pseudo alignment)
- Statistical analysis
  - Quantification
  - Hypothesis tests
  - Clustering
  - Integrate with other data
- Visualization

QC

- FastQC
  - Before alignment
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- RNA-SeQC
  - After alignment
  - https://confluence.broadinstitute.org/display/CGATools/RNA-SeQC
- Proportion of reads that mapped uniquely
  - Mark duplicates to assess PCR over-amplification
- Assess ribosomal RNA and possible contaminant content
  - human RNA (if not human samples)
  - Mycoplasma (if cell lines)
- Quality of de novo assembled transcripts:
  - http://hibberdlab.com/transrate/

Quantification

1. Assign reads to transcripts
    - pre-defined versus de novo transcriptome
    - multi-mapping reads
2. Normalization of transcript read counts
    - Library size
    - Gene length
    - Base composition biases (hexamers, GC%)

Example summary statistics: RPKM, FPKM, CPM, TMM

Differential Expression

- Goal: determine whether observed difference in read counts is greater than would be expected due to random variation.

- If reads independently sampled from population, they would follow multinomial distribution approximated by Poisson

Differential Expression

- BUT! We know that the count data show more variance than expected under Poisson
- Over-dispersion problem mitigated by using the **negative binomial distribution**, which is determined by mean and dispersion parameters
- Dispersion is hard to estimate
  - High false positive rate: http://biorxiv.org/content/early/2015/06/11/020784
  - Estimates based on different methods vary

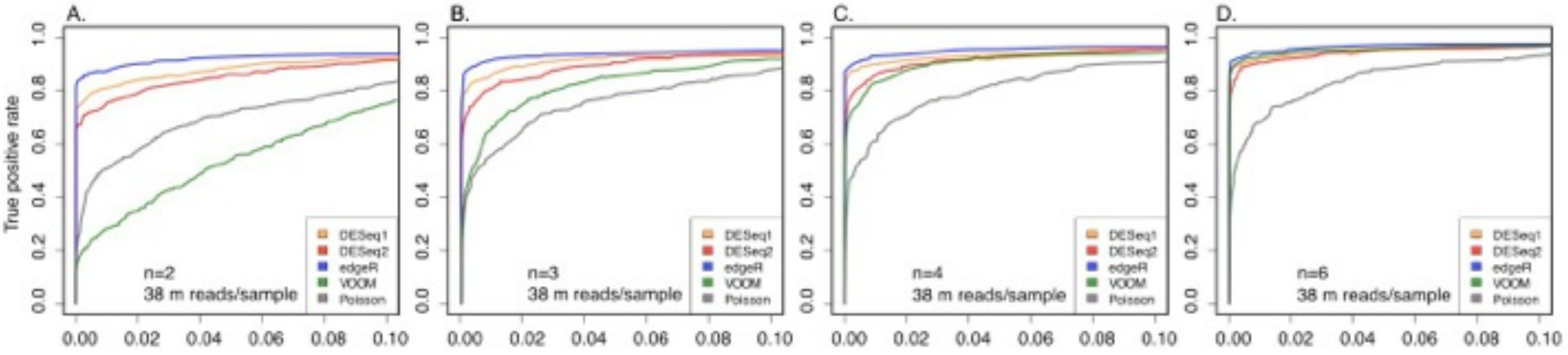
Many software packages

Quantification and statistical analysis:

- edgeR
- DESeq / DESeq2
- VOOM (+ limma)
- Others…

Differential Expression - sample size

n=2 n=3 n=4 n=6



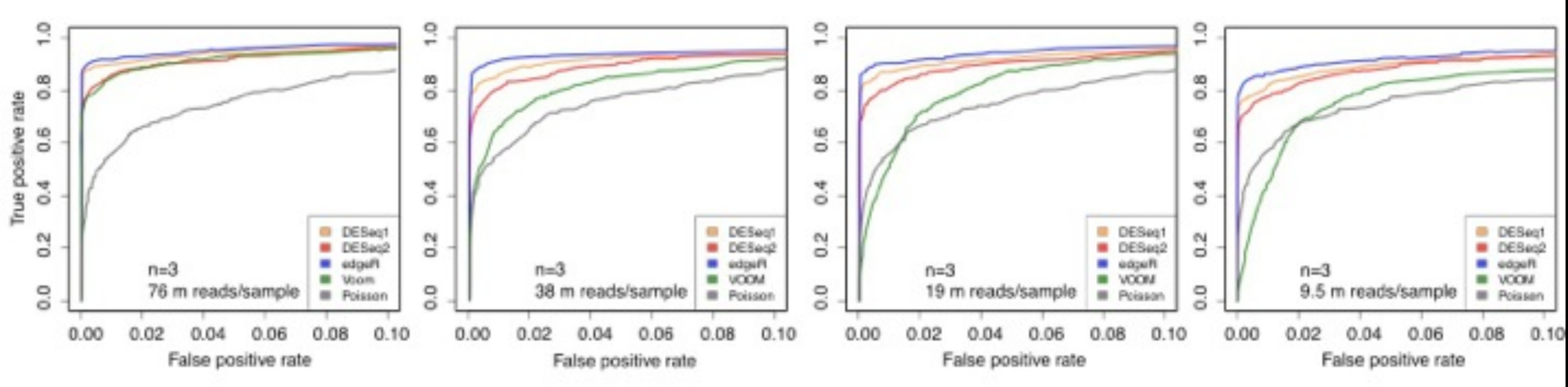
False Positive Rate

- Sensitivity increases with samples size
- EdgeR wins

Williams et al. 2014

Differential Expression - mapped reads

76m 38m 19m 9.5m



False Positive Rate

- Sensitivity increases with number mapped reads
- EdgeR wins

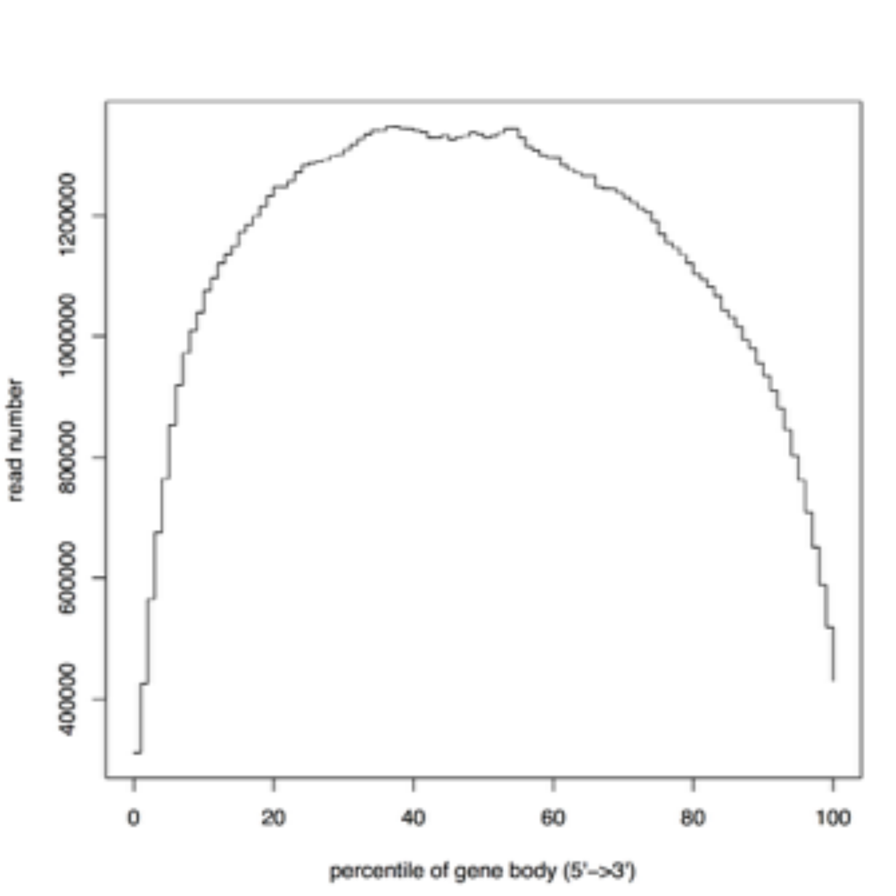
Williams et al. 2014

Additional Details

Power of paired-end reads

- Impact on read mapping
  - Pairs give two locations to determine whether fragment is unique (assess PCR over-amplification)
- Useful for estimating transcript-level abundance
  - Increases number of splice junction spanning reads and fragments
    - Either the read maps over a splice junction or each end of a pair maps to different exons
- Single end is often good enough

Distribution of reads over gene body



TSS END

Normalized by gene length

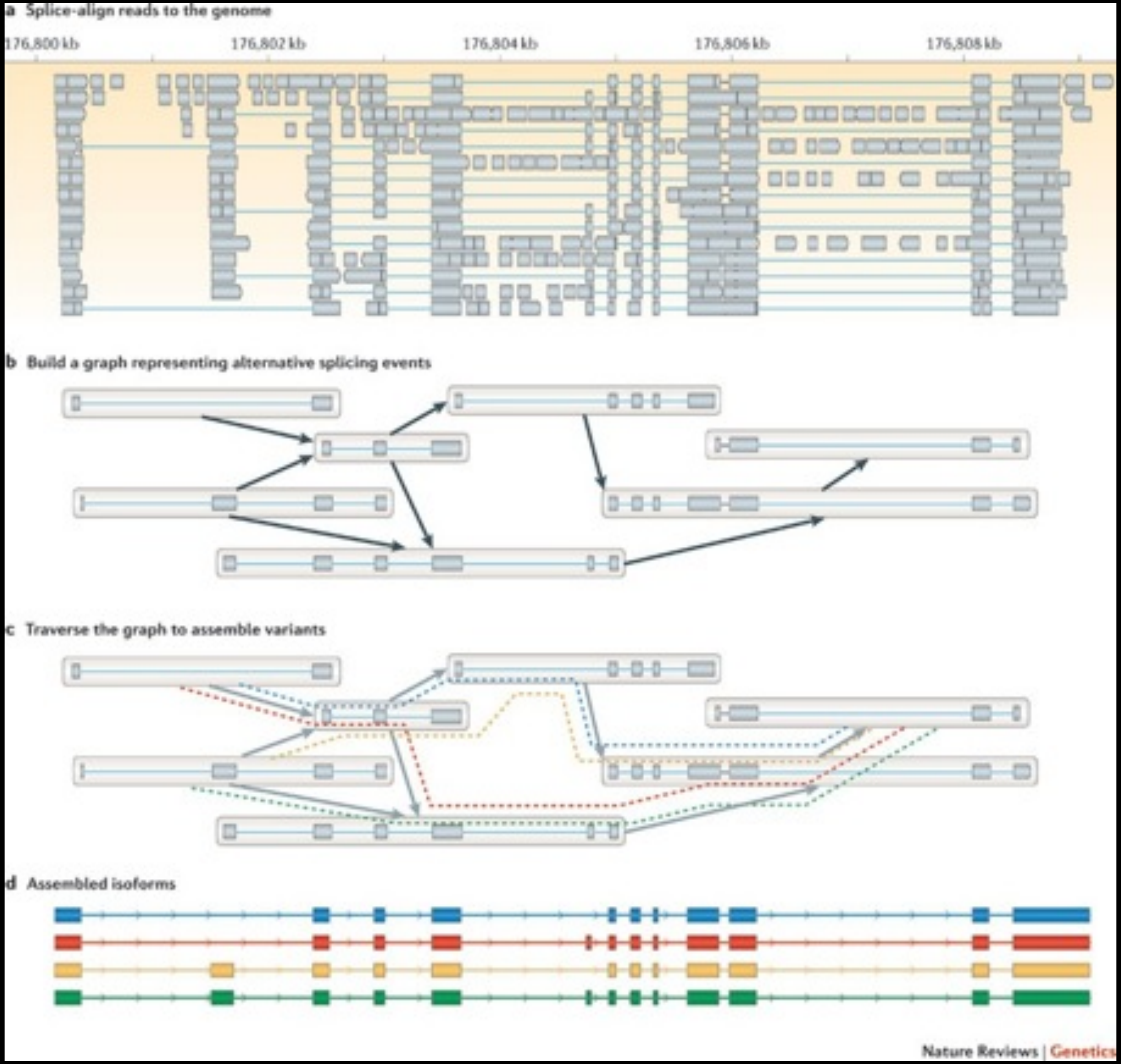
Transcript Assignment

Aligned contiguous and spliced reads

Build graph to connect neighboring concordant alignments

Traverse graph to assemble variants

Assemble possible isoforms


a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms
Nature Reviews | Genetics

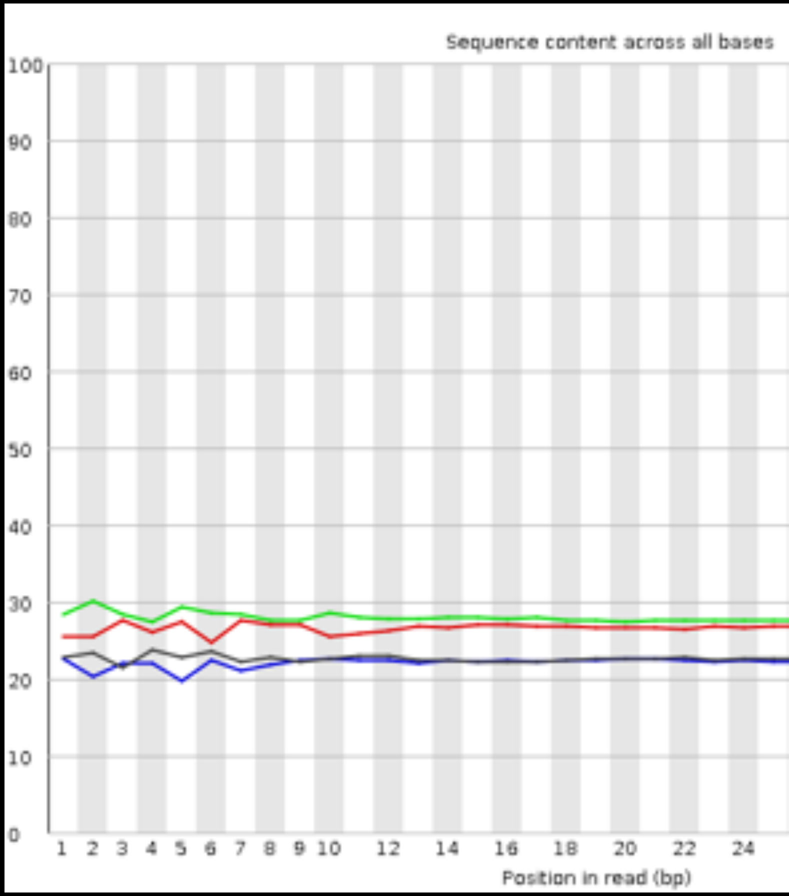
Martin & Wang, Nature Reviews Genetics 2011

Transcript Assignment Tools

- Annotated transcript assembly
  - Cufflinks
  - RSEM
  - TIGAR
  - MISO
- *De novo* transcript assembly
  - Cufflinks
  - Trans-ABySS
  - Trinity
  - RSEM

Bias Correction and Normalization

- Random hexamer bias (Hansen et al. 2010)
  - From PCR or RT primers
  - Re-estimate read counts to account for bias
- Resources for normalization
  - Bullard et al. 2010
  - Williams et al. 2014
  - http://www.rna-seqblog.com/data-analysis/which-method-should-you-use-for-normalization-of-rna-seq-data/

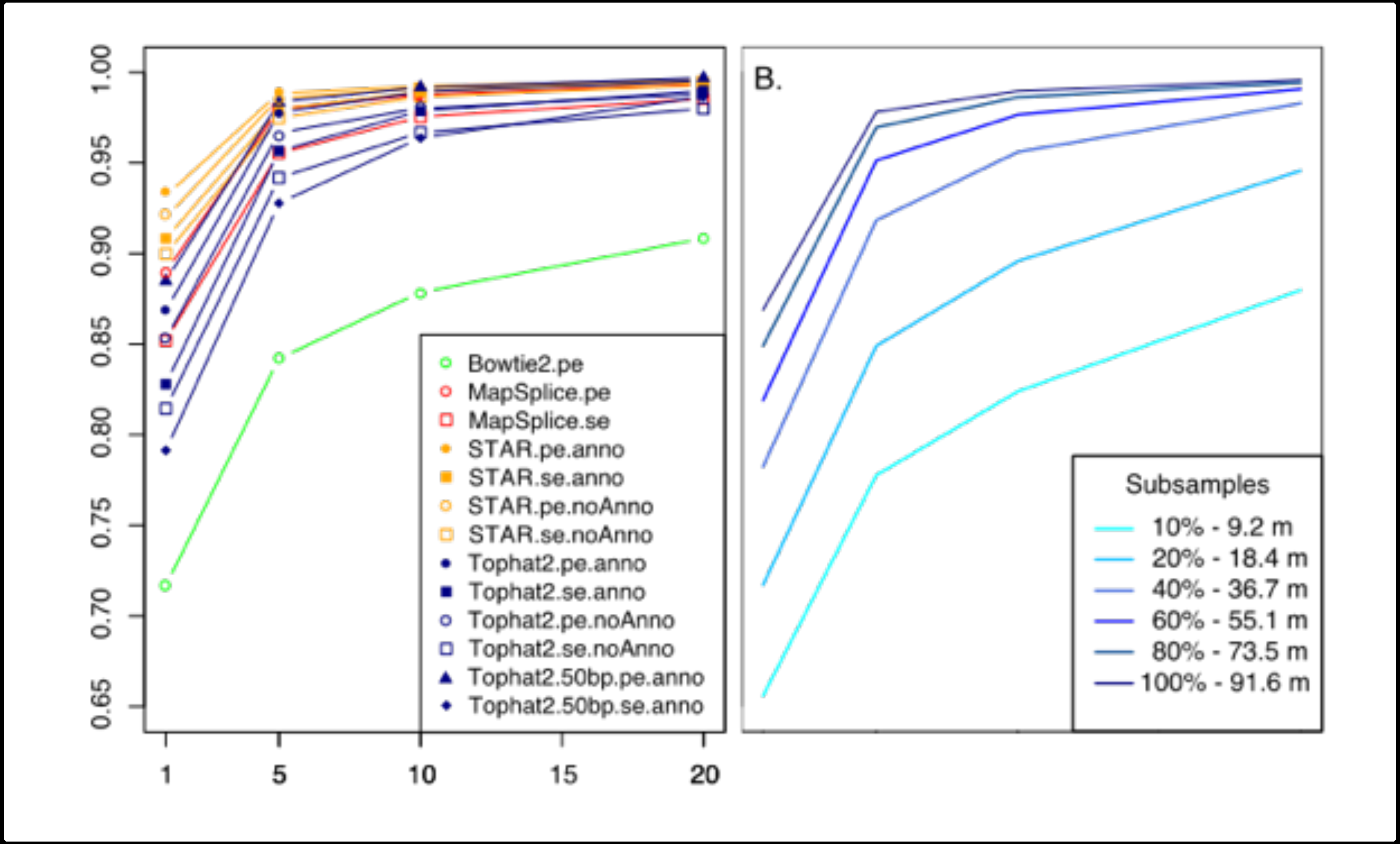


Compare Splice Junction Mappers

SJM	Length	SE | PE	Annotation	Prop. SJs Relative to BWA*	% Splice Junctions Recovered
Mapsplice	100	PE	No	0.89	89.0%
Mapsplice	100	SE	No	0.43	85.2%
STAR	100	PE	Yes	0.94	93.2%
STAR	100	SE	Yes	0.44	90.8%
STAR	100	PE	No	0.83	92.0%
STAR	100	SE	No	0.35	90.0%
Tophat2	100	PE	Yes	0.73	86.9%
Tophat2	100	SE	Yes	0.41	82.8%
Tophat2	100	PE	No	0.64	85.3%
Tophat2	100	SE	No	0.37	81.5%
Tophat2	50	PE	Yes	0.82	88.5%
Tophat2	50	SE	Yes	0.22	79.1%

*Mapped reads to transcriptome using BWA to establish ground truth.

Compare tools for splice junction mapping


Bowtie2.pe
MapSplice.pe
MapSplice.se
STAR.pe.anno
STAR.se.anno
STAR.pe.noAnno
STAR.se.noAnno
Tophat2.pe.anno
Tophat2.se.anno
Tophat2.pe.noAnno
Tophat2.se.noAnno
Tophat2.50bp.pe.anno
Tophat2.50bp.se.anno
B.
Subsamples
10% - 9.2 m
20% - 18.4 m
40% - 36.7 m
60% - 55.1 m
80% - 73.5 m
100% - 91.6 m