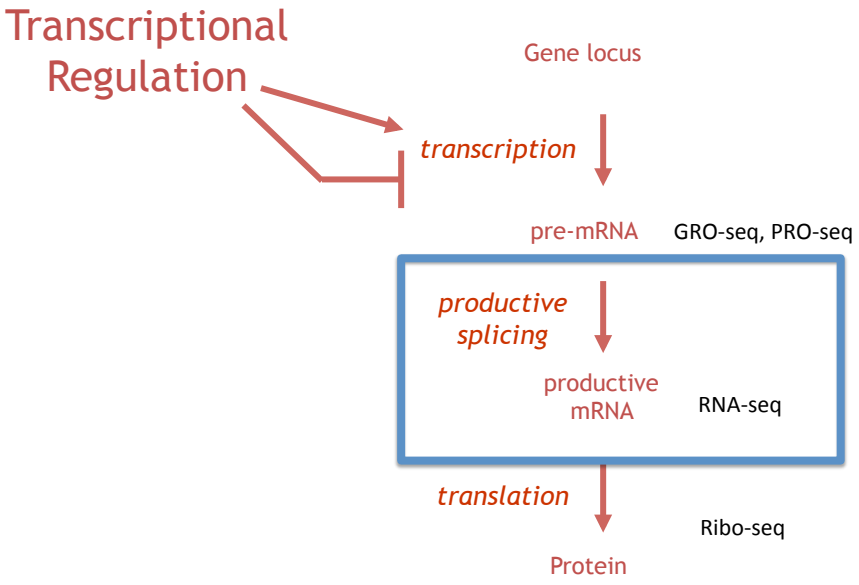
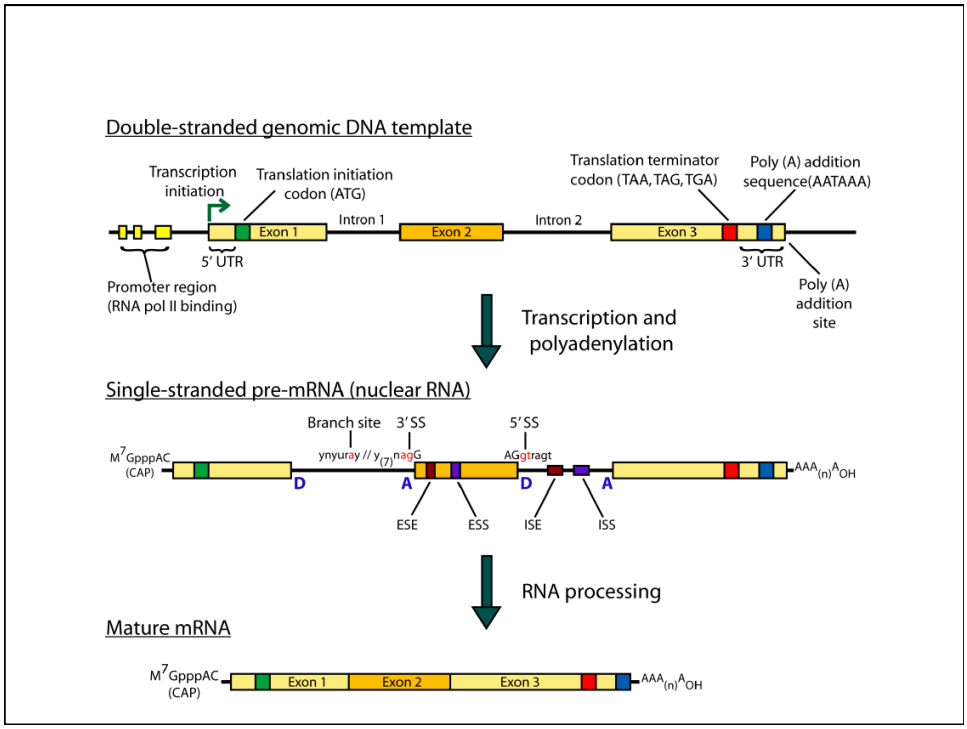
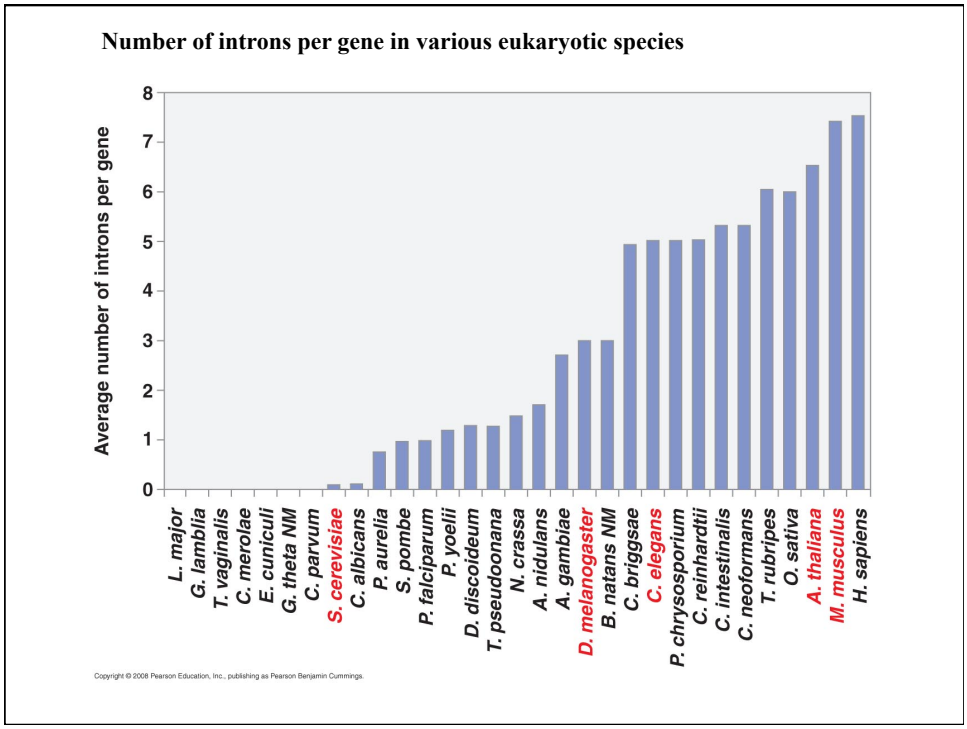
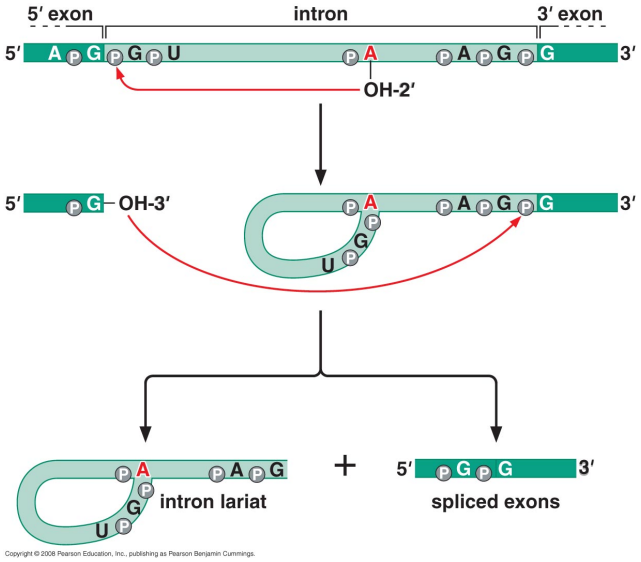
"Everybody is talented because everybody who is human has something to express."
— Brenda Ueland




Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
M7GpppAC (CAP)
ynyuray // y(7) nagG
AGgtragt
AAA(n)AOH
D
A
D
A
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
AAA(n)AOH


Number of introns per gene in various eukaryotic species
Average number of introns per gene
0
1
2
3
4
5
6
7
8
L. major
G. lamblia
T. vaginalis
C. merolae
E. cuniculi
G. theta NM
C. parvum
S. cerevisiae
C. albicans
P. aurelia
S. pombe
P. falciparum
P. yoelii
D. discoideum
T. pseudoonana
N. crassa
A. nidulans
A. gambiae
D. melanogaster
B. natans NM
C. briggsae
C. elegans
P. chrysosporium
C. reinhardtii
C. intestinalis
C. neoformans
T. rubripes
O. sativa
A. thaliana
M. musculus
H. sapiens
Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

The intron is removed in a form called lariat and the flanking exons are joined



The Nobel Prize in Chemistry 1989
"for their discovery of catalytic properties of RNA"

Sidney Altman

1/2 of the prize

Canada and USA

Yale University
New Haven, CT, USA

b. 1939

Thomas R. Cech

1/2 of the prize

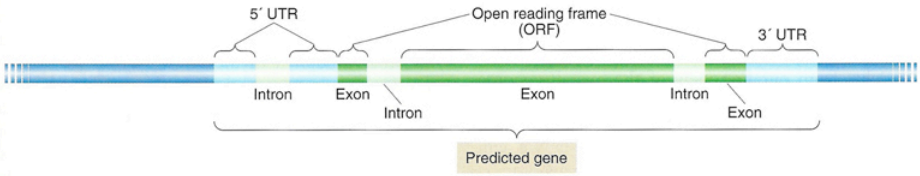
USA

University of Colorado
Boulder, CO, USA

b. 1947

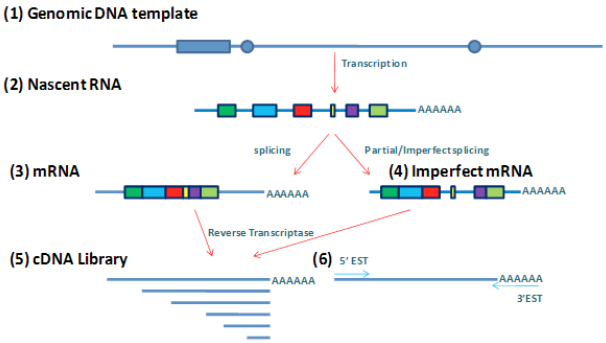
Applications of RNA-seq

- Differential expression
- Gene fusion
- Alternative splicing
- Novel transcribed regions
- Allele-specific expression
- RNA editing
- Transcriptome assembly for non-model organisms



The earliest approach to expression

Frederick "Fred" Sanger

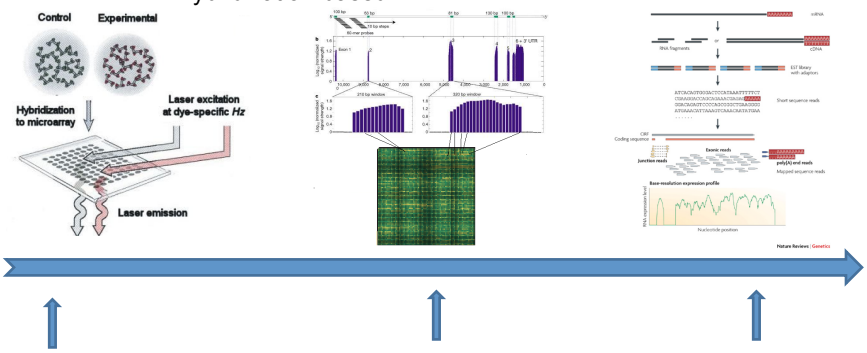


<u>Tag Based Sequencing Approaches</u>
Serial Analysis of Gene Expression (SAGE)
Cap Analysis of Gene Expression (CAGE)
} Expressed Sequence Tags (ESTs)

<u>Limitations of Sanger Sequencing</u>
Low throughput
Expensive
Not quantitative (no enough depth!)

The evolution of expression assays

Hybridization-based



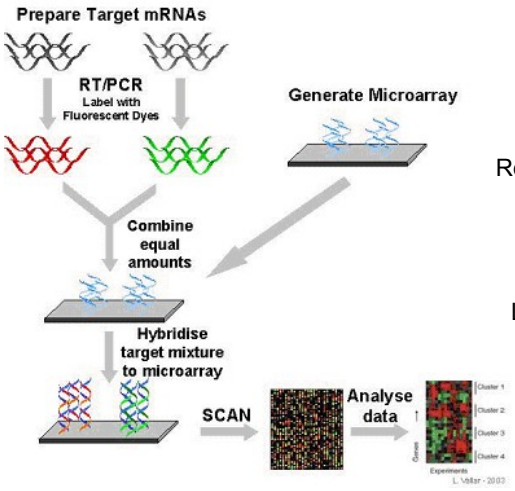
1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

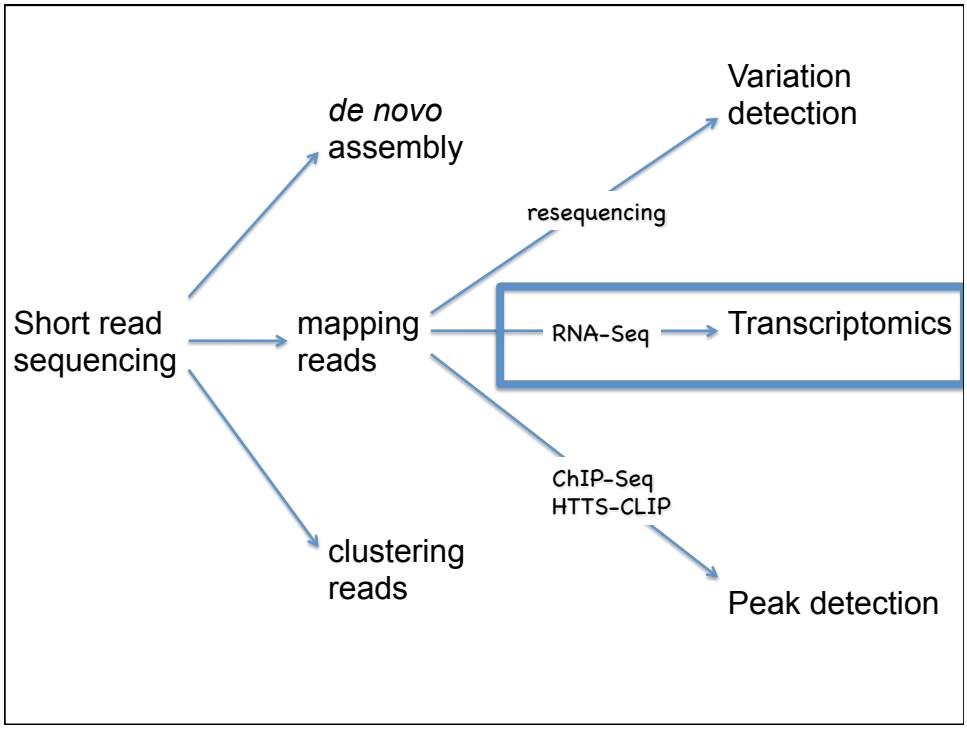
Microarrays

Hybridization Based Gene Expression Quantification

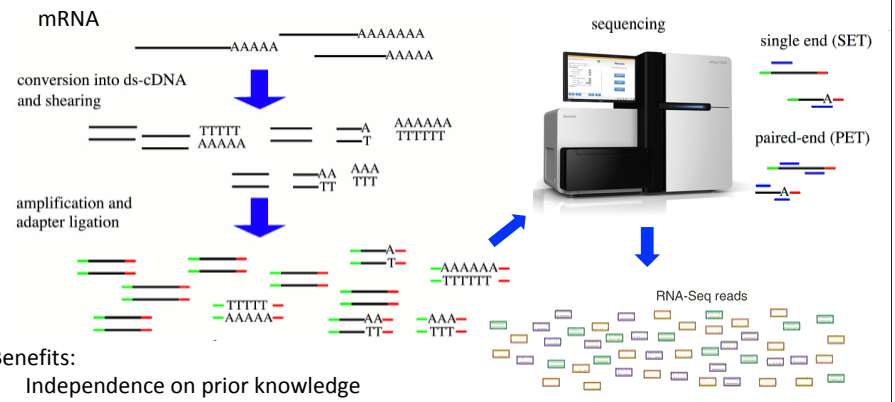


LIMITATIONS

Reliance on existing knowledge about genome sequence
High background due to cross-hybridization
Requires lots of starting material
Limited dynamic range of detection



Overview of RNA-Seq



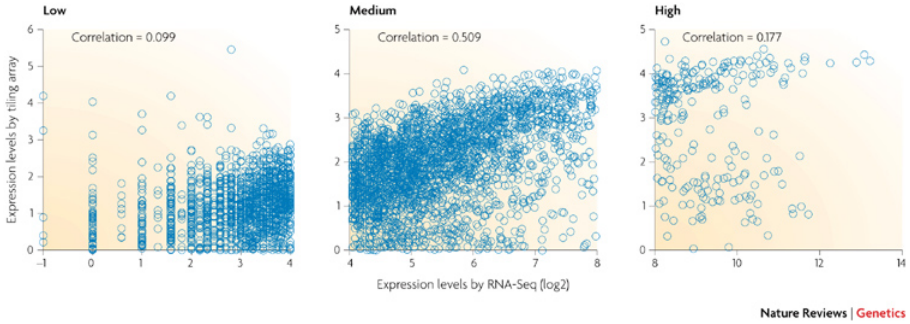
Benefits:

- Independence on prior knowledge
- High resolution, sensitivity and large dynamic range
- Unravel previously inaccessible complexities

Challenge:

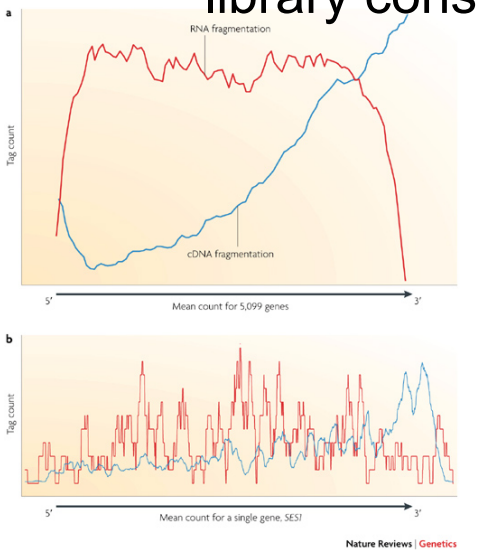
- Interpretation is not straightforward
- Procedures continue to evolve

RNA-seq and microarray agree fairly well only for genes with medium levels of expression



Saccharomyces cerevisiae cells grown in nutrient-rich media. Correlation is very low for genes with either low or high expression levels because microarray saturates at high levels and both techniques struggle with very lowly expressed.

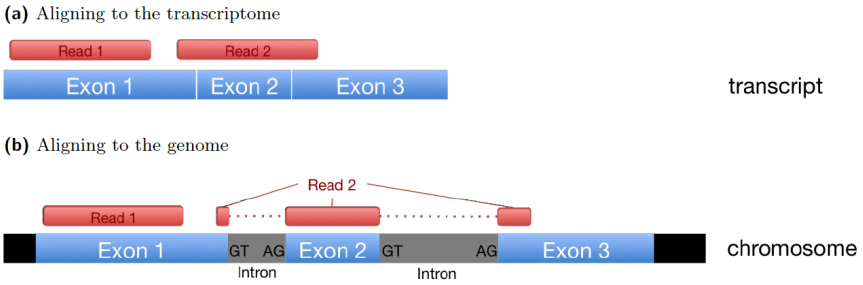
Challenges for RNA-Seq: library construction



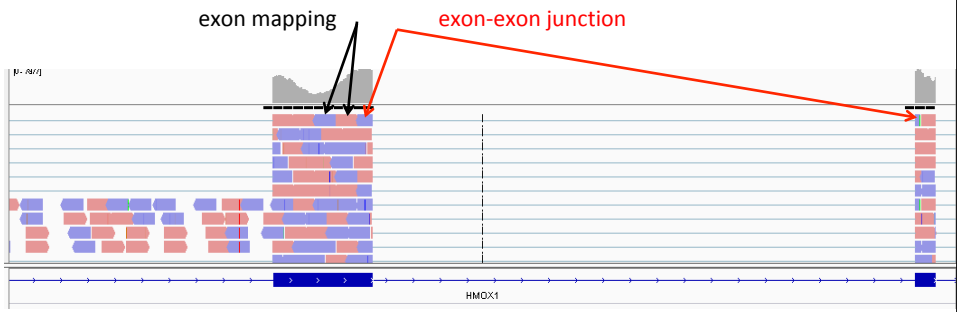
Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

A specific yeast gene, SES1 (seryl-tRNA synthetase)

Mature RNA presents a unique problem for read mapping …

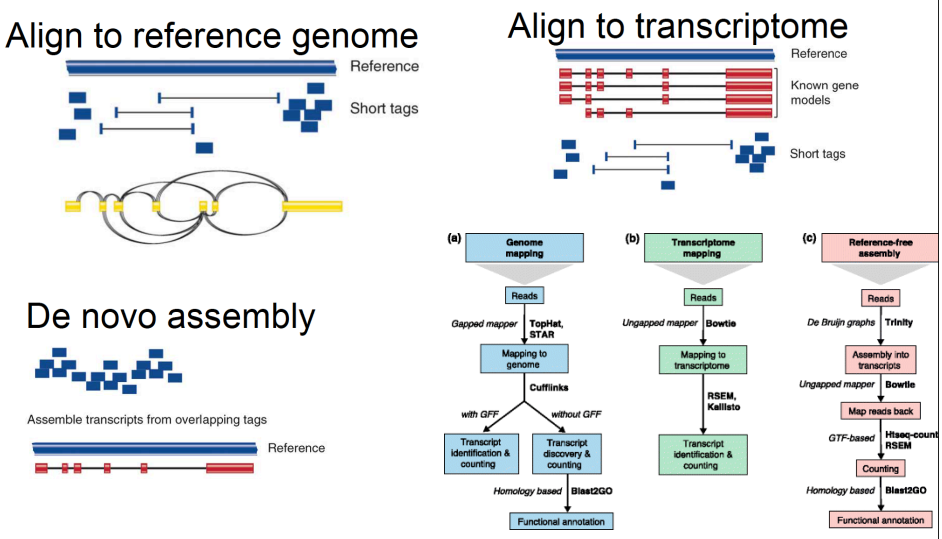


Read mapping



Unlike DNA-Seq, when mapping RNA-Seq reads back to reference genome, we need to pay attention to **exon-exon junction reads**

Three mapping strategies.

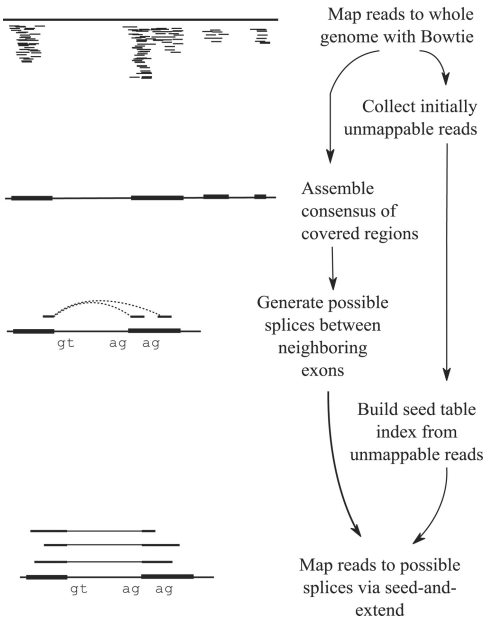


Diagrams from: Cloonan & Grimmond, Nature Methods 2010

Align to reference genome: TopHat

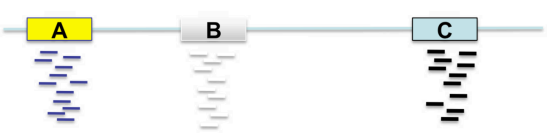
Do not assume annotation.

Putative exon defined by clustering reads (islands of coverage)



Trapnell et. al. (2009)

Consider all possible pairings of exons.



identify candidate exons via genomic mapping

Generate possible pairings of exons

Require the splice junction bases: gt … ag

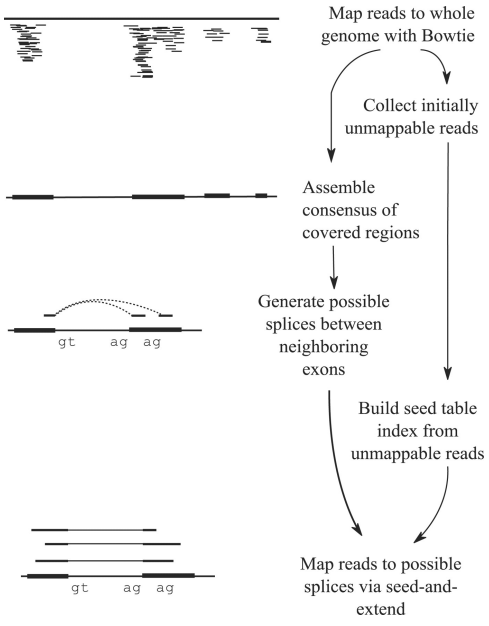
Align reads to possible junctions

TopHat: splice junctions

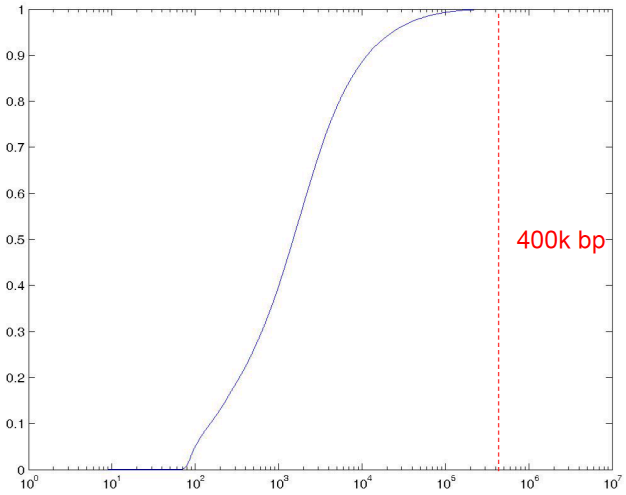
Find GT-AG pairing sites between neighboring islands

The distance between two sites should > 70bp and <20k bp, as most intron length lies within this range

Build exon based "genome" for mapping split reads.



The cumulative distribution function of the intron sizes

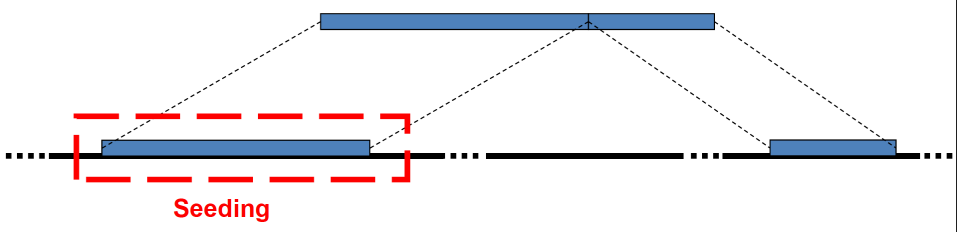


* Based on hg19 human Refseq annotation.

Alternative: Finding split reads

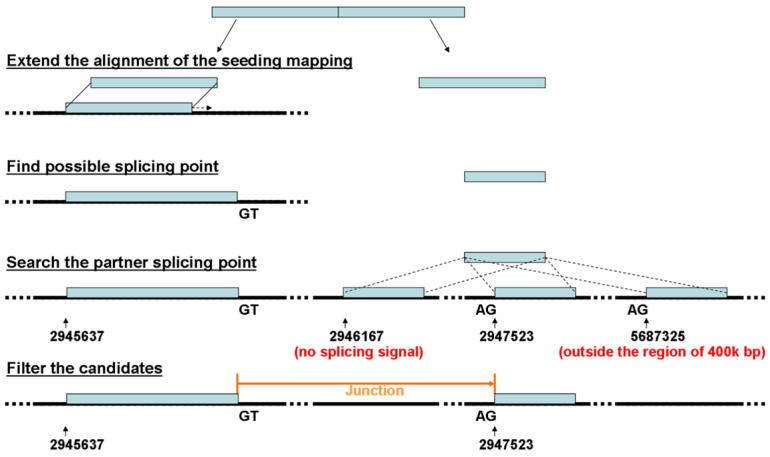
Basic concept

- **Split map**



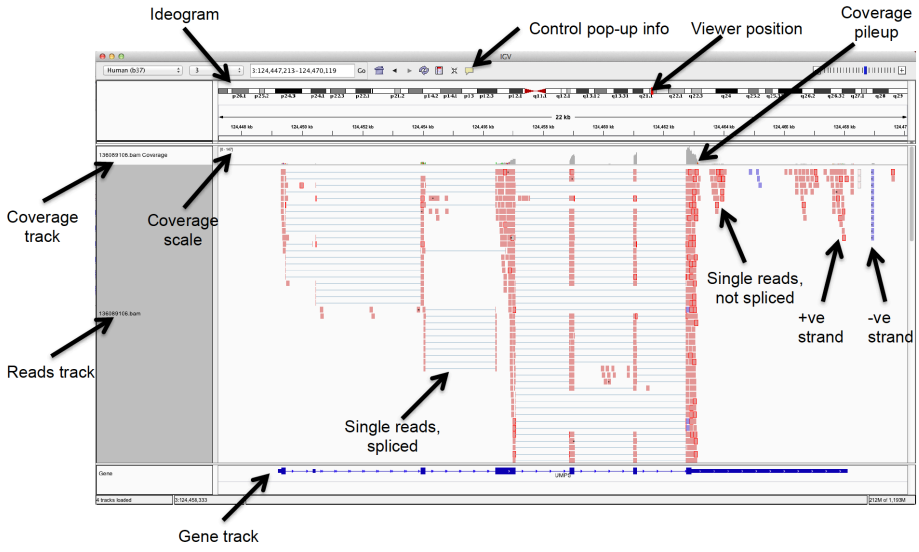
- **if read length >=50 bp**
- **at least one of the halves will have non-split map**

Junction search:

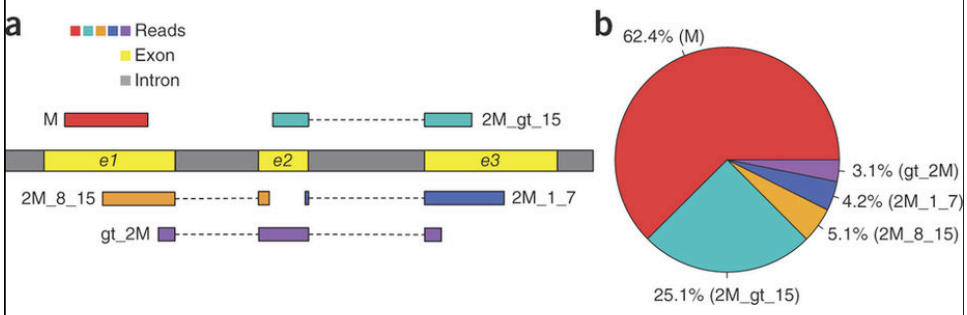


- **Residual length >= 10 bp**
- **Canonical splicing signal GT-AG (appears in 98% splices)**
- **Distance < 400k bp (existence of intron between two splicing exons)**

Many splice junctions per gene

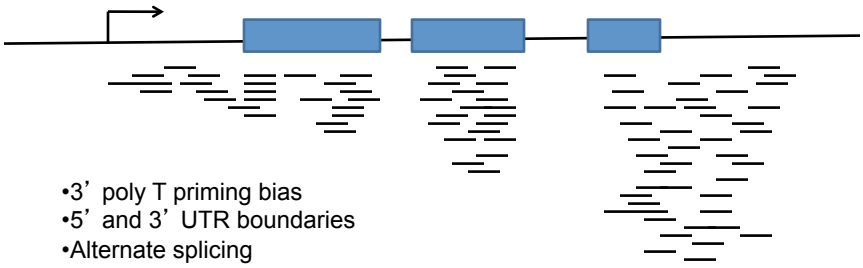


But many types of reads possible



Reads with poor anchoring (orange, dark blue) or spanning multiple exons (purple) were largely missed by first generation splice junction aligners (TopHat) but are now being discovered by second generation aligners (HISAT).

Expression quantification can be non-trivial



- 3' poly T priming bias
- 5' and 3' UTR boundaries
- Alternate splicing
- Cryptic exons
- Cryptic start sites
- Paired-end reads can help

Expression quantification

Gene A: 200

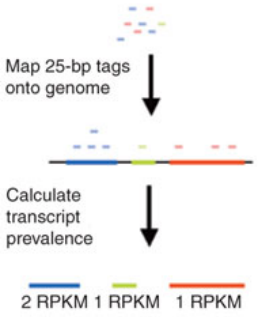
Gene B: 300

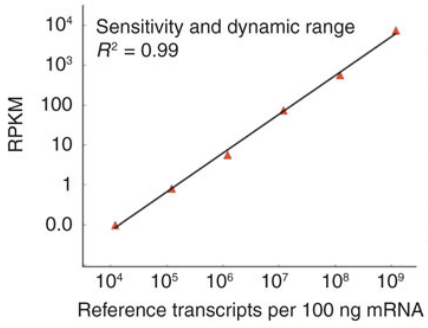
Is expression of Gene A < Expression of Gene B?

The number of reads is roughly proportional to

- the length of the gene
- the total number of reads in the library

Quantification of known transcripts





- The expression levels of known transcripts (*exon model*) are measured by the number of reads per kilobase of transcript per million mapped reads (RPKM)

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 621–628 (2008).

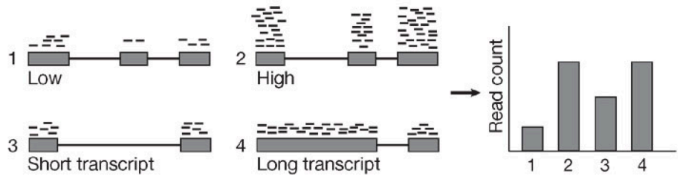
Gene and isoform abundance

Gene abundance
--RPKM: reads per kilobase of the transcript per million mapped reads to the transcriptome (Mortazavi et.al., 2008)

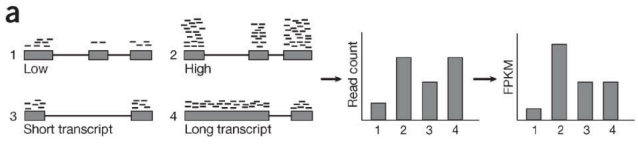
1 RKPM ~ 0.3 to 1 transcript per cell

Cons:
Could not be used to measure isoform abundance, because most reads are *shared by multiple isoforms*.

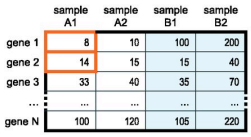
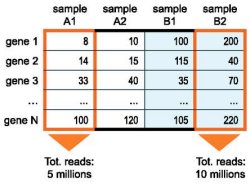
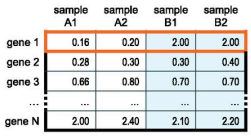
- Question: What are the RPKM-corrected expression values and why?



- Note especially normalization for fragment length (transcripts 3 and 4)



Graphic credit: Garber et al. (2011) *Nature Methods* **8**:469–477. Note that the authors here use the related term FPKM, **Fragments per KB per million reads**, which is suitable for paired-end reads (we will not cover the details here).

	ANALYSIS STEPS	CHALLENGES
	**READ MAPPING** Find original read source within the reference genome or transcriptome.	Computationally intensive. Different trade-offs between speed and accuracy produce different sub-optimal solutions.
	**COUNTS COMPUTATION** Estimate gene expression with "counts", i.e. with the number of reads mapped on each gene.	Counts depend on the strategy adopted to define gene boundaries and to handle reads mapping on multiple genes.
	**COUNTS NORMALIZATION** Eliminate biases to make expression levels comparable between samples (e.g. different sequencing depths of samples A1 and B2) and within samples (e.g. different lengths of gene 1 and gene 2).	Gene-specific biases are difficult to remove and can impact on downstream analyses.
	**DIFFERENTIAL EXPRESSION ANALYSIS** Identify genes with statistically different expression levels in the compared conditions (e.g. A and B).	FRIDAY True differential expression can be detected only with precise and powerful tests and when technical and biological variability is correctly modeled.