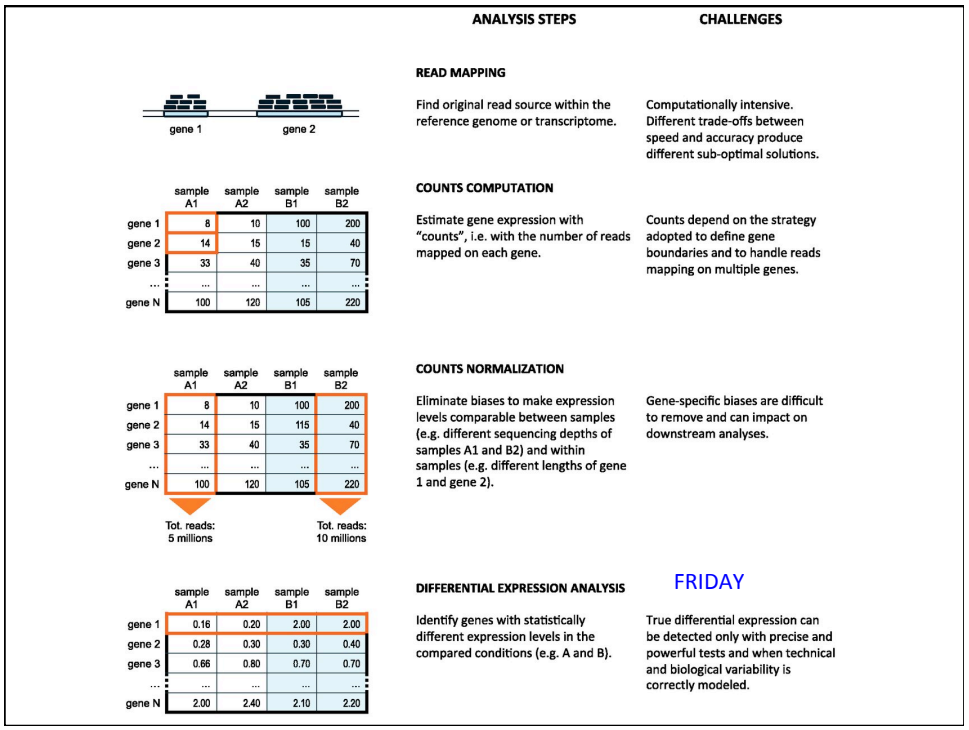
"Sometimes a writer wants to cut out unnecessary information from a *quotation* to place emphasis on the part of it that matters most or to eliminate unnecessary information."

https://quoting1.weebly.com/splicing-quotes.html



Handling multi-map reads in counts

- a non-negligible fraction of RNA-seq reads are 'multireads': reads that map with comparable fidelity on multiple positions of the reference.

  - Ignore them (earliest approaches)
  - Weight them ("fragments")
  - Sophisticated approaches ("nearby" signal)

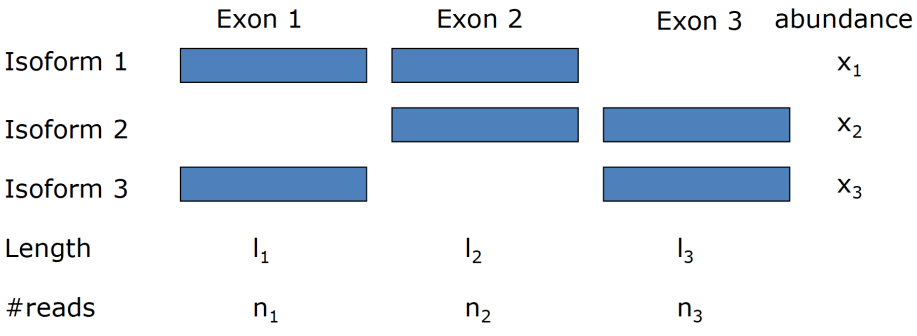
Recall: Normalization methods

- RPKM: Reads Per Kilobase Million
  - Count up the total reads in a sample and divide that number by 1,000,000 – this is our "per million" scaling factor.
  - Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
  - Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.
- FPKM: Fragments Per Kilobase Million
  - With paired-end RNA-seq, two reads can correspond to a single fragment, or, if one read in the pair did not map, one read can correspond to a single fragment. FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).
  - Also handles multi-read mapping better ("fragments")

One more simple normalization

- TPM: Transcripts Per Kilobase Million
  - Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
  - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your "per million" scaling factor.
  - Divide the RPK values by the "per million" scaling factor. This gives you TPM.
- The sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample.
- However, requires that your "feature set" be more or less complete.

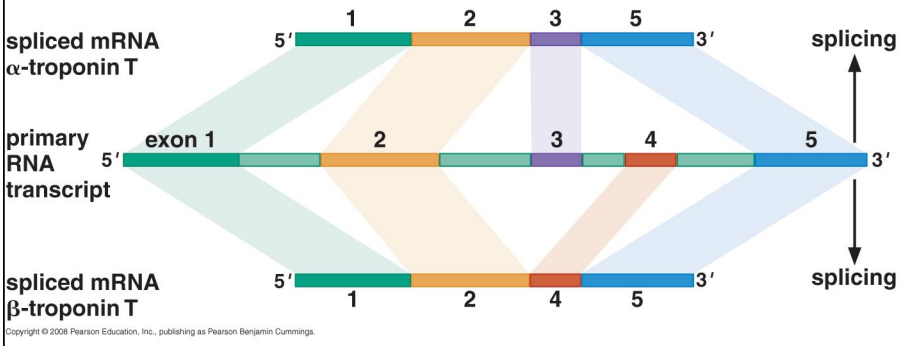
Simple abundance estimates are inappropriate for isoforms.



In other words, RPKM/FPKM is too simplistic for isoforms because a read's origin is unclear. So how to do we infer isoforms to properly identify a read's likely "origin"?

ALTERNATIVE SPLICING

Single gene can produce multiple products by alternative splicing



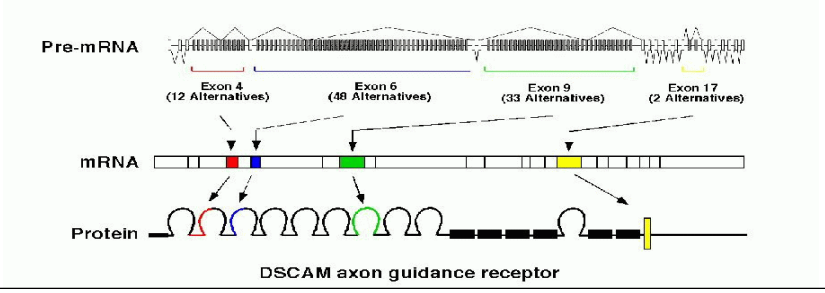
Alternative Splicing (AS)

MOST human genes show Alt Splicing

(92-94% of all genes; > 86% genes with alt isoform > 15% abundance)

some genes have a huge number of isoforms

(*slo* >500, *neurexin* >1000, *DSCAM* > 38000)



Large collections of splicing data

(pre-short read sequencing era)

Curated databases

SWISS-PROT and RefSeq both support annotation of experimentally supported alternative splicing

cDNA Sequencing Projects

RIKEN sequenced >21000 full length mouse cDNAs

Many other projects underway (human, fly, plants,...)

Shinagawa et al. (2001) *Nature* **409**:685-90

Microarray detection

Direct or indirect alternative splicing detection

Hu et al. (2001)*Genome Res* **11**:1237-45
Yeakley et al. (2002) *Nat Biotech* **20**:353-9

Public EST data sources (dbEST)

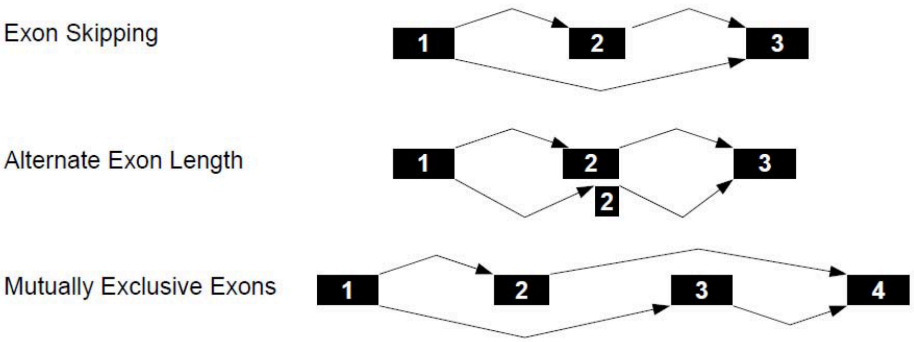
>4.5 million human EST sequences

>12 million total EST sequences

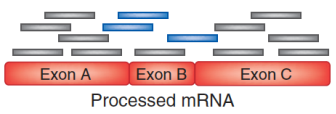
About 1000 new sequences per day

Boguski et al. (1993) *Nat Gen* **4**:332-3

First step: detect splice junctions

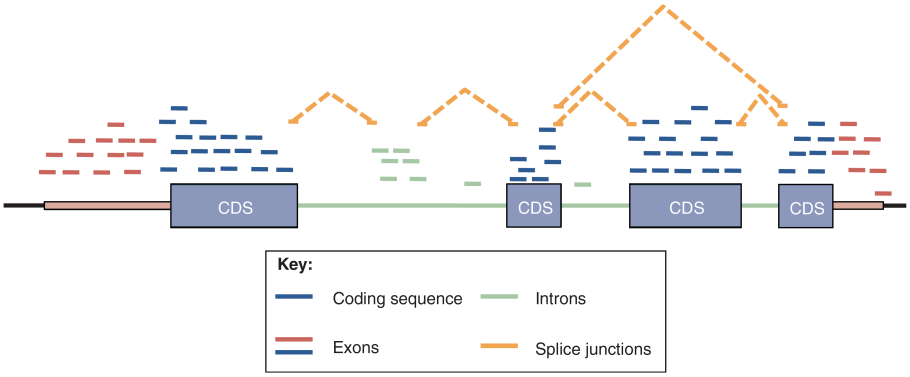


Mapping RNA-seq reads and quantifying transcripts

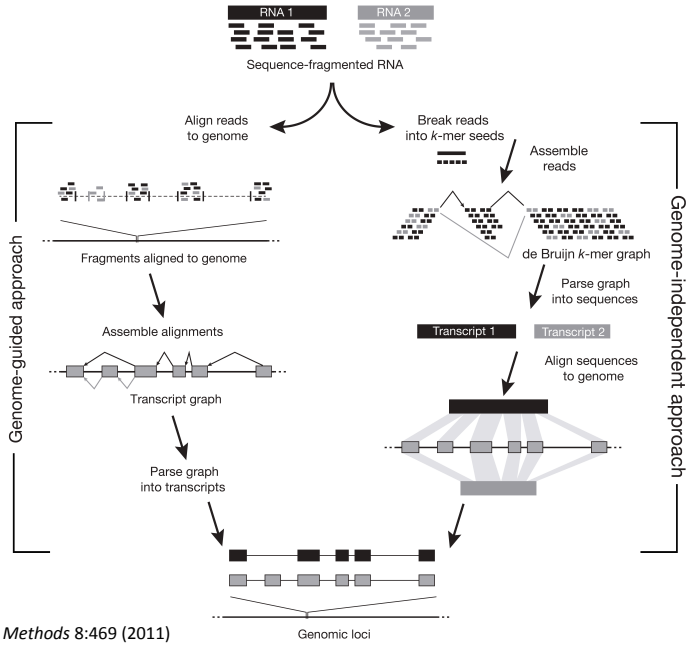


Two steps:

1. Infer Isoform (transcript) Use
2. Quantify Transcripts

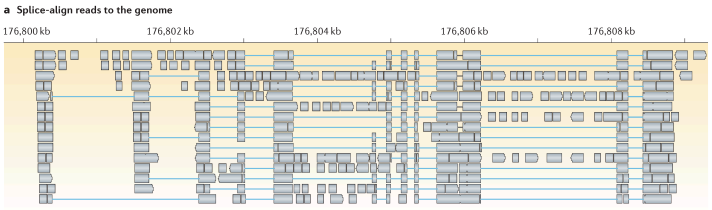


Strategies for Inferring Isoforms (Transcript Assembly)



Garber *et al. Nat Methods* 8:469 (2011)

Reference based transcript assembly

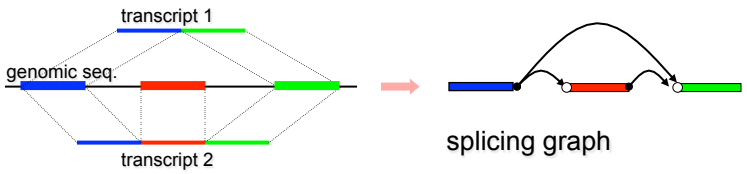


Martin and Wang *Nat Rev Genet* 12:671 (2011)

Splicing graph construction

If a reference genome is used:

- Map reads to the reference genome (short read aligner)
- Check alignment (splice sites, quality)
- Connect consecutive positions
- Build splicing graph

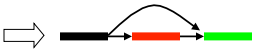


Splice graph approach

Replace the problem of finding a list of consensus sequences

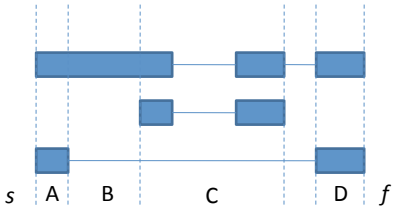
with ***Graph Reconstruction Problem***:

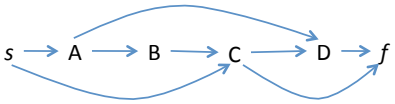
Given an set of expressed sequence, find a minimal graph (*splicing graph*) representing **all** transcripts as paths.



Heber, et. al. ISMB 2002

Splicing graph and splicing variants





An edge in the splicing graph, called a *block*, represents a maximal sequence of adjacent exons or exon fragments that always appear together in a given set of splicing variants. Therefore, variants can be represented by sequence of blocks, e.g. {ABCD, C, AD}.

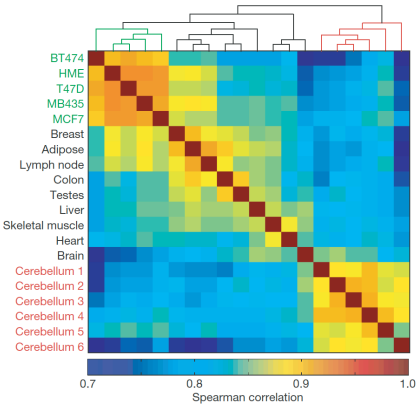
Vertices *s* and *f* are included into graph, and are linked to the 5' and 3' of each variant, respectively.

Larcroix, et. al. WABI, 2009

Diversity of alternative splicing events in human tissues

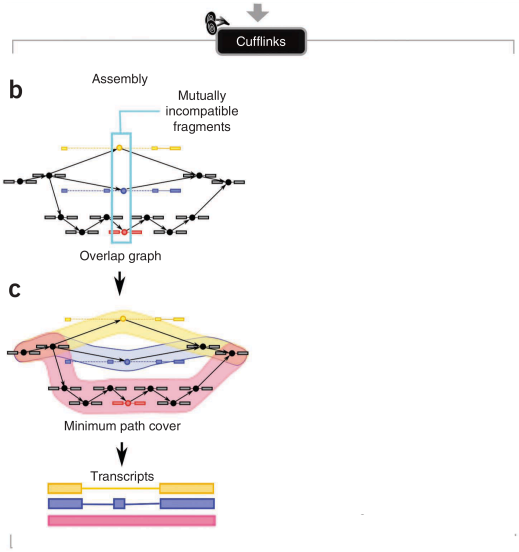
Alternative transcript events	Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

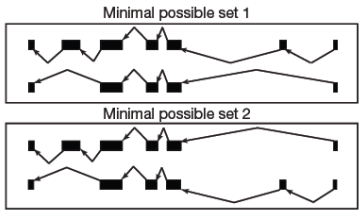
Constitutive exon or region — Body read — Junction read — pA Polyadenylation site
Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms



Wang et al. *Nature* 456:470 (2008)

Cufflinks: *ab initio* transcript assembly

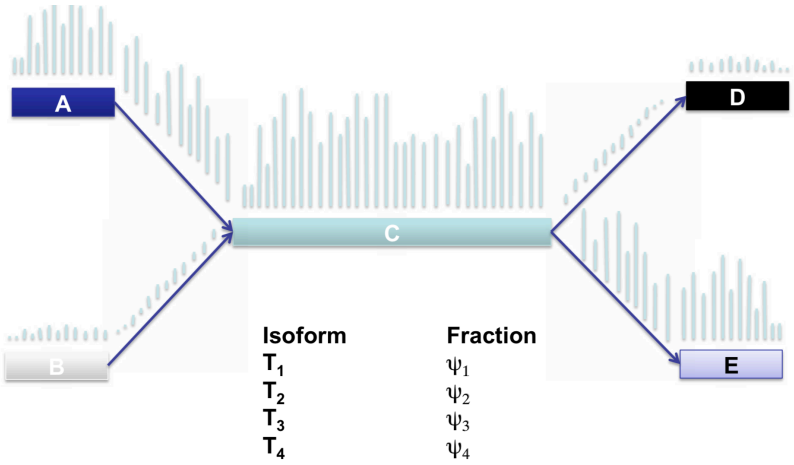




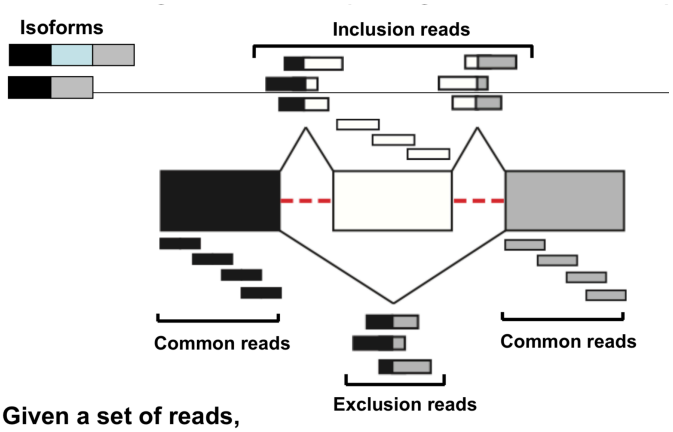
Isoform abundances estimated by maximum likelihood

Trapnell *et al. Nat. Biotechnology* 28:511 (2010)

We can use mapped reads to learn the mixture of isoforms present.



Our dataset contains a mixture of isoforms



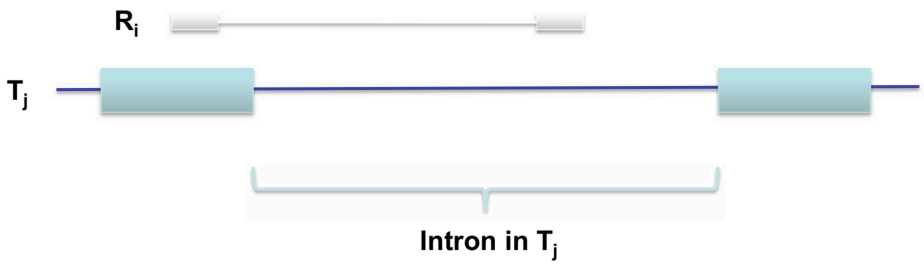
Given a set of reads, estimate:

Ψ = **Distribution of isoforms**

Some reads can be excluded from some isoform possibilities.

If a single ended read or read pair R_i is structurally incompatible with transcript T_j, then

$$P(R = R_i \mid T = T_j) = 0$$

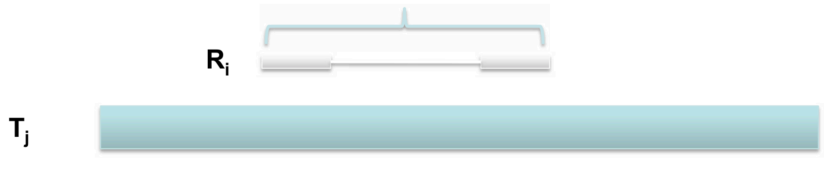


Some reads have a probability of originating with a particular isoform.

Assume our library fragments have a length distribution described by a probability density F. Thus, the probability of observing a particular paired alignment to a transcript:

$$P(R = R_i \mid T = T_j) = \frac{F(l_j(R_j))}{l_j}$$

Implied fragment length $l_j(R_i)$



Our probabilistic model

- Find expression abundances $\psi_1,\ldots,\psi_n$ for a set of isoforms $T_1,\ldots,T_n$
- Observations are the set of reads $R_1,\ldots,R_m$

This is weighted sum ..

$$P(R \mid \Psi) = \prod_{i=0}^{m} \sum_{j=0}^{n} \Psi_j P(R = R_i \mid T = T_j)$$

This is BAYES again!

$$L(\Psi \mid R) \propto P(R \mid \Psi) P(\Psi)$$

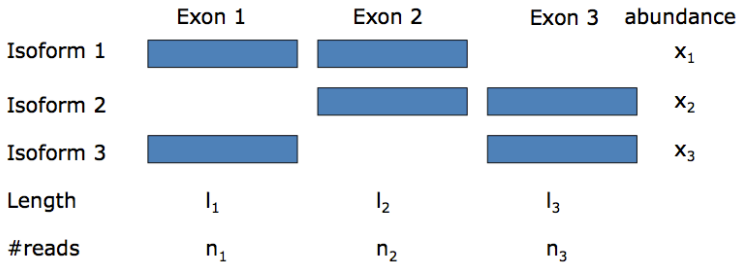
Our optimization.

$$\Psi = \underset{\Psi}{\operatorname{argmax}} L(\Psi \mid R)$$

- Can estimate mRNA expression of each isoform using total number of reads that map to a gene and ψ

Isoform Inference is akin to a "system of equations"

- If given known set of isoforms



- Estimate x to maximize the likelihood of observing n

24

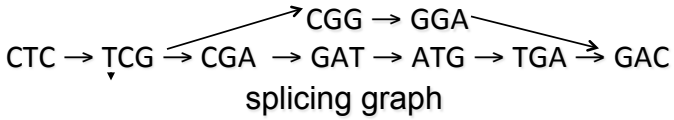
Splicing graph construction

If a reference genome sequence is not used:

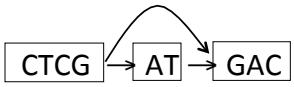
- Break sequences into *k*-mers (20-mers).
- Build graph using *k*-mers as vertices, connect them iff they occur consecutively in a sequence [Pevzner et al., 2001].

Example (3-mers):

Sequences: CTCGATGAC, CTCGGAC
Vertices: {CTC, TCG, CGA, GAT, ATG, TGA, GAC, CGG, GGA}

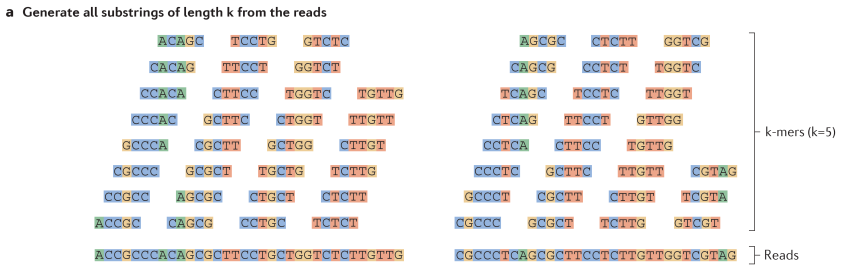


splicing graph



collapsed splicing graph

Graph-based transcript assembly



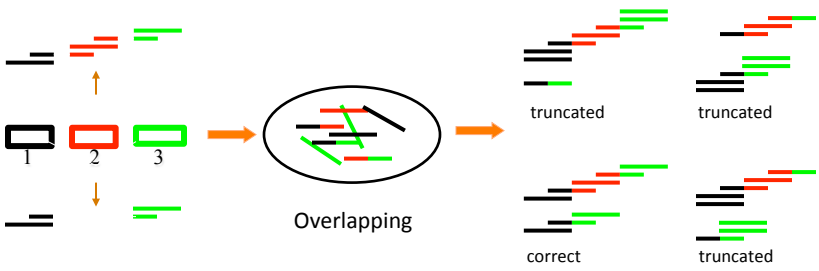
Martin and Wang *Nat Rev Genet* 12:671 (2011)

Graph-based transcript assembly

Martin and Wang *Nat Rev Genet* 12:671 (2011)

How to assemble multiple alternative spliced transcripts?

In the presence of AS, conventional assembly may be erroneous, ambiguous, or truncated.

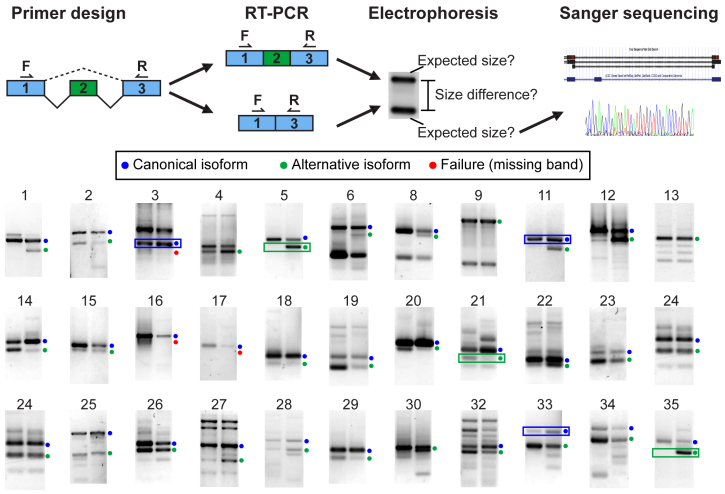


Common questions: how reliable are expression predictions from RNA-seq?

- Are novel exon-exon junctions real?
  - What proportion validate by RT-PCR and Sanger sequencing?
- Are differential/alternative expression changes observed between tissues accurate?
  - How well do DE values correlate with qPCR?

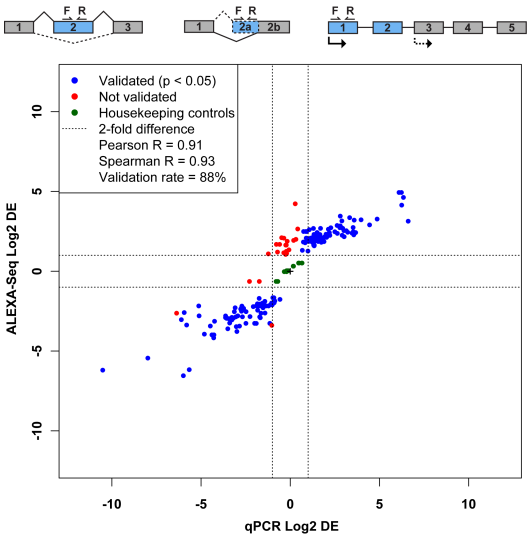
- 384 validations
  - qPCR, RT-PCR, Sanger sequencing
- See ALEXA-Seq publication for details:
  - Also includes comparison to microarrays
  - Griffith et al. *Alternative expression analysis by RNA sequencing*. Nature Methods. 2010 Oct;7(10):843-847.

Validation (qualitative)



33 of 192 assays shown. Overall validation rate = 85%

Validation (quantitative)



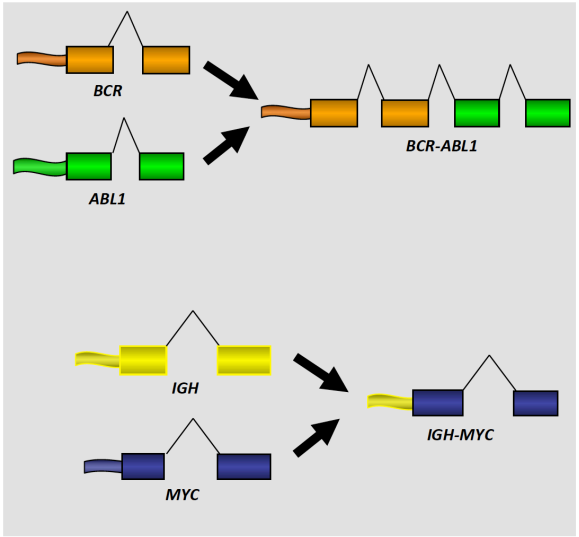
qPCR of 192 exons identified as alternatively expressed by ALEXA-Seq

Validation rate = 88%

Gene fusions

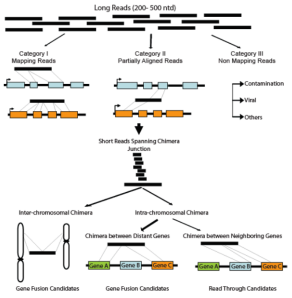
I. Generation of a fusion gene

II. Activation of proto-oncogenes by relocation in the vicinity of active regulatory elements



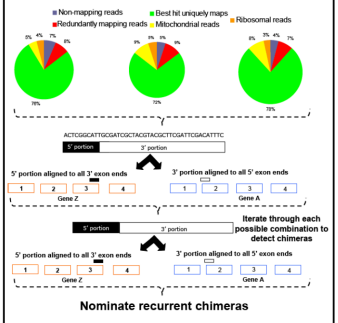
Evolution of chimera detection strategies

Long single read (454) + Short single read (Illumina)



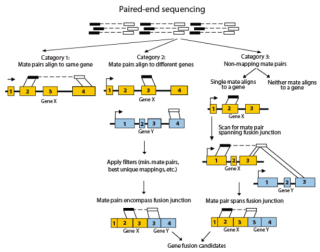
(Nature - Maher et al., 2009)

Short single read (Illumina)



(Nature - Maher et al., 2009)

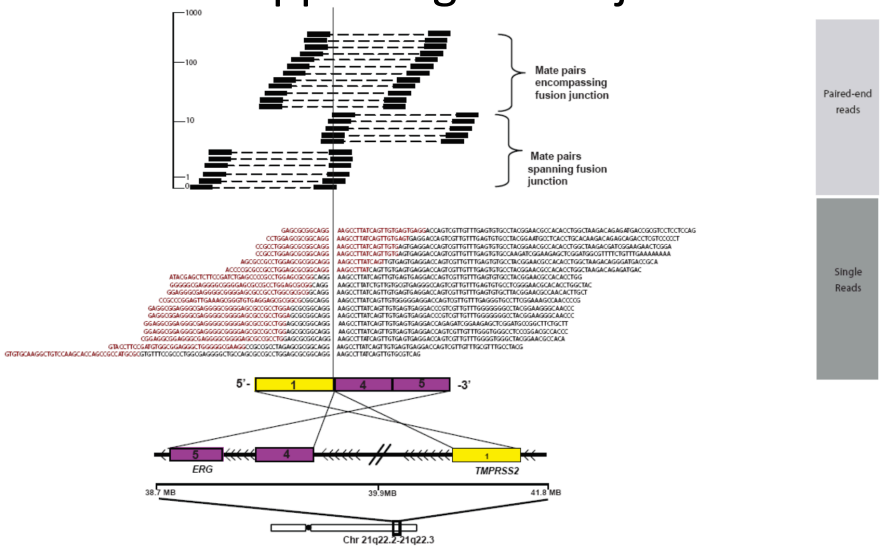
Paired-end reads (Illumina)



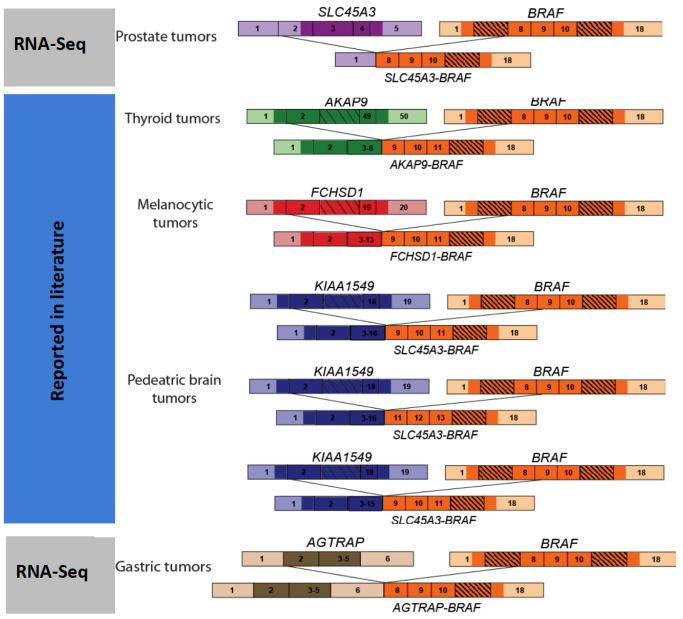
(PNAS - Maher et al., 2009)
(Bioinformatics – Iyer et al., 2011)

Depth of Coverage

Paired end strategies are particularly useful at supporting fusion junctions.



(PNAS - Maher et al., 2009)



Certain fusions have been detected in multiple cancers.