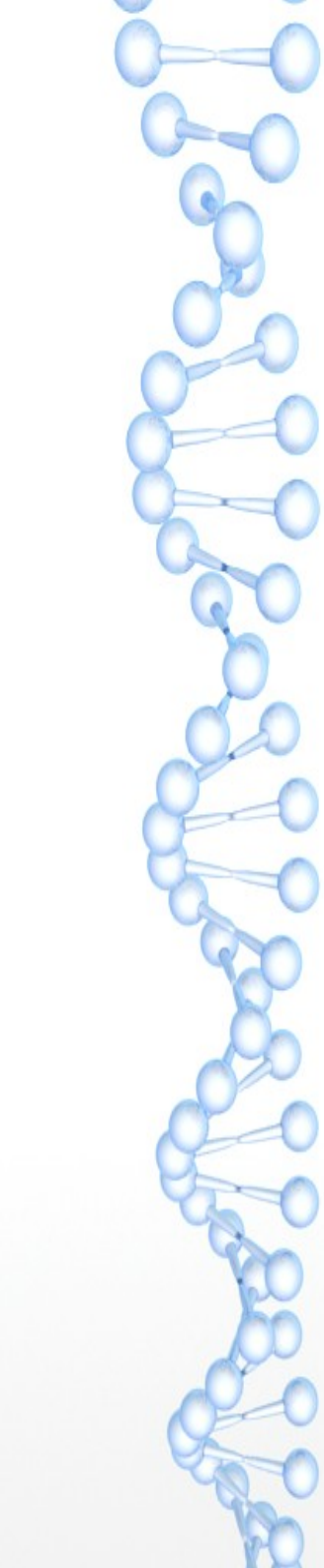


TopHat

Problem

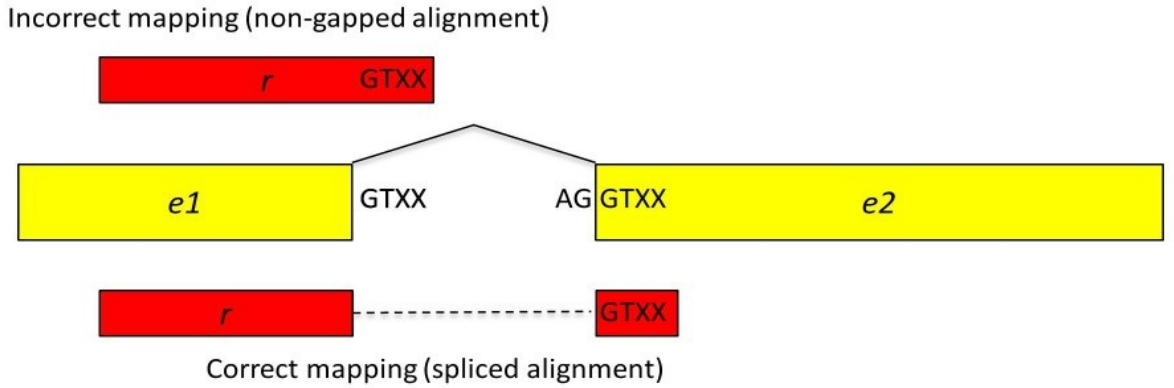
- We map the 'reads' to the reference transcriptome
- Transcriptomes are incomplete even for well-studied species

TopHat pipeline

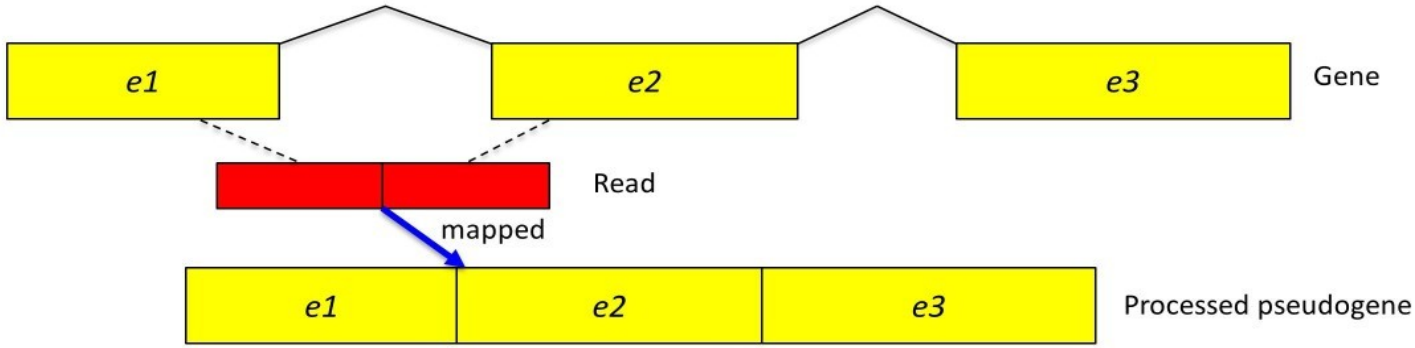
- First map non-junction reads (those contained within exons) using Bowtie
- Extracts the sequences for the resulting islands of contiguous sequence from the sparse consensus, inferring them to be putative exons

TopHat pipeline

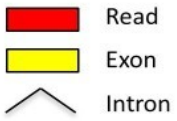
- TopHat includes a small amount of flanking sequence from the reference on both sides of each island (default=45 bp)
- Consider all pairings of these sites that could form canonical (GT–AG)
- TopHat has a parameter that controls when two distinct but nearby exons should be merged into a single exon. (default is 6 bp)
- By default potential introns longer than 70 bp and shorter than 20 000 bp

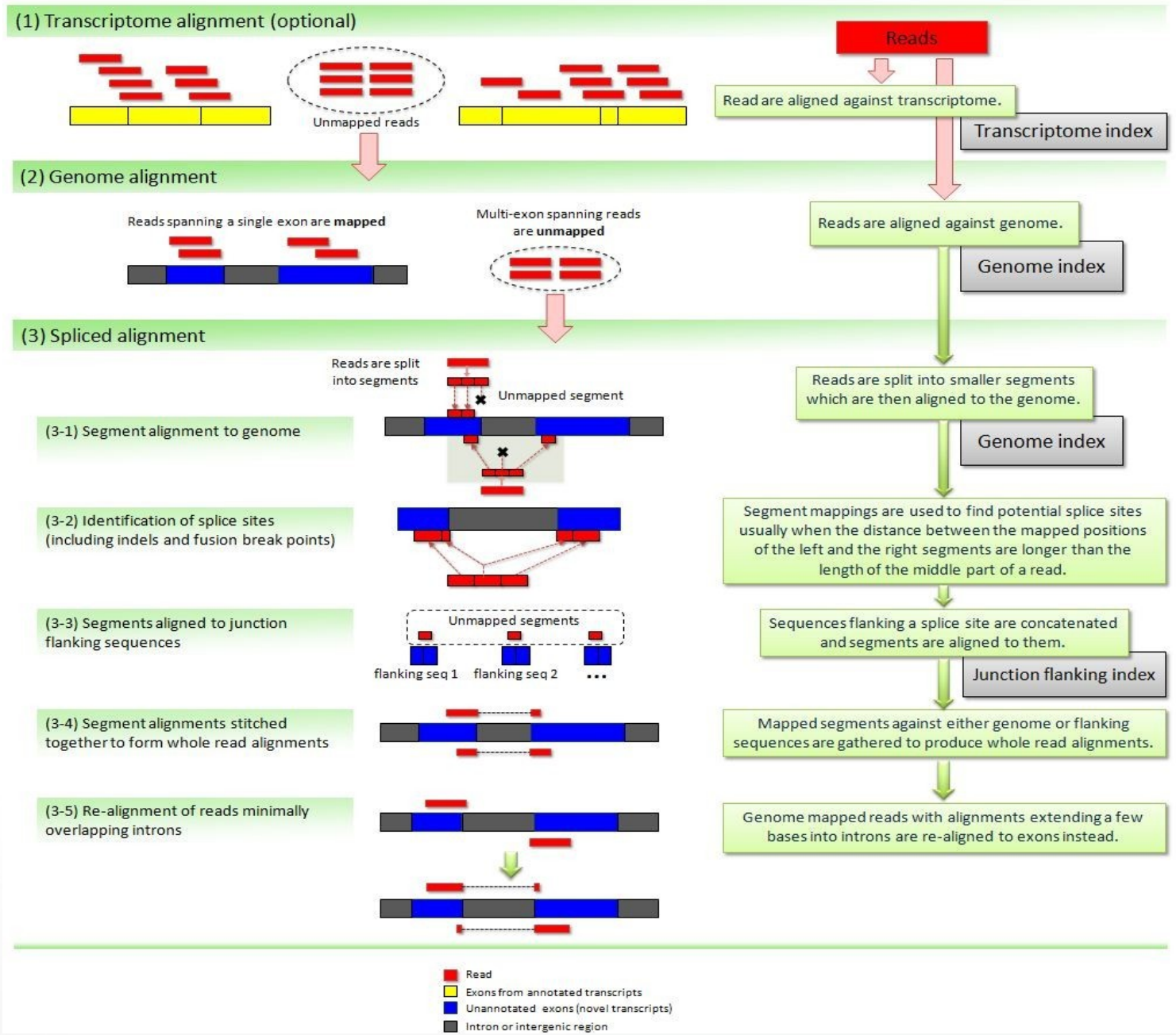


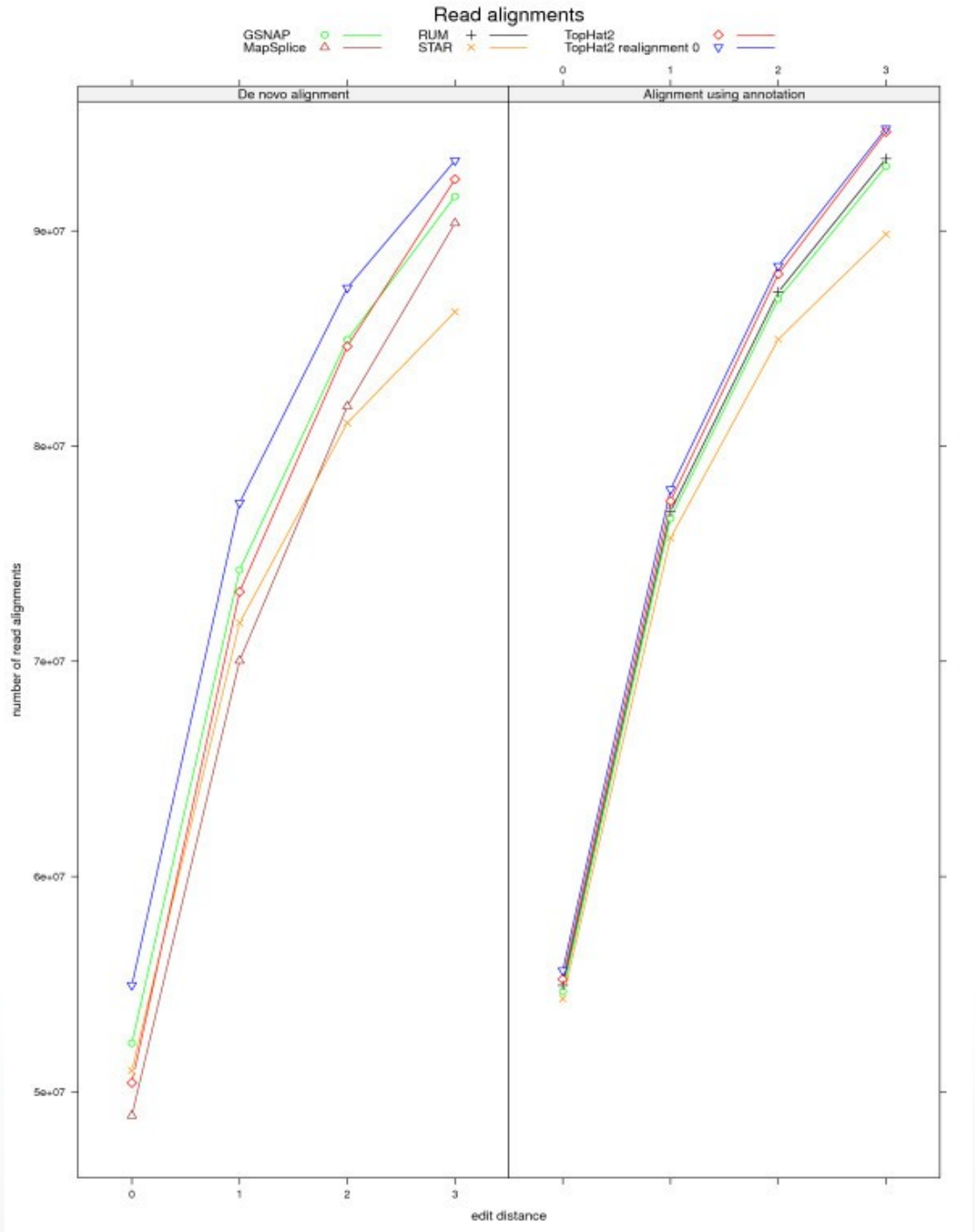
(1) Read *r* may be incorrectly mapped to the intron between exons *e1* and *e2*.

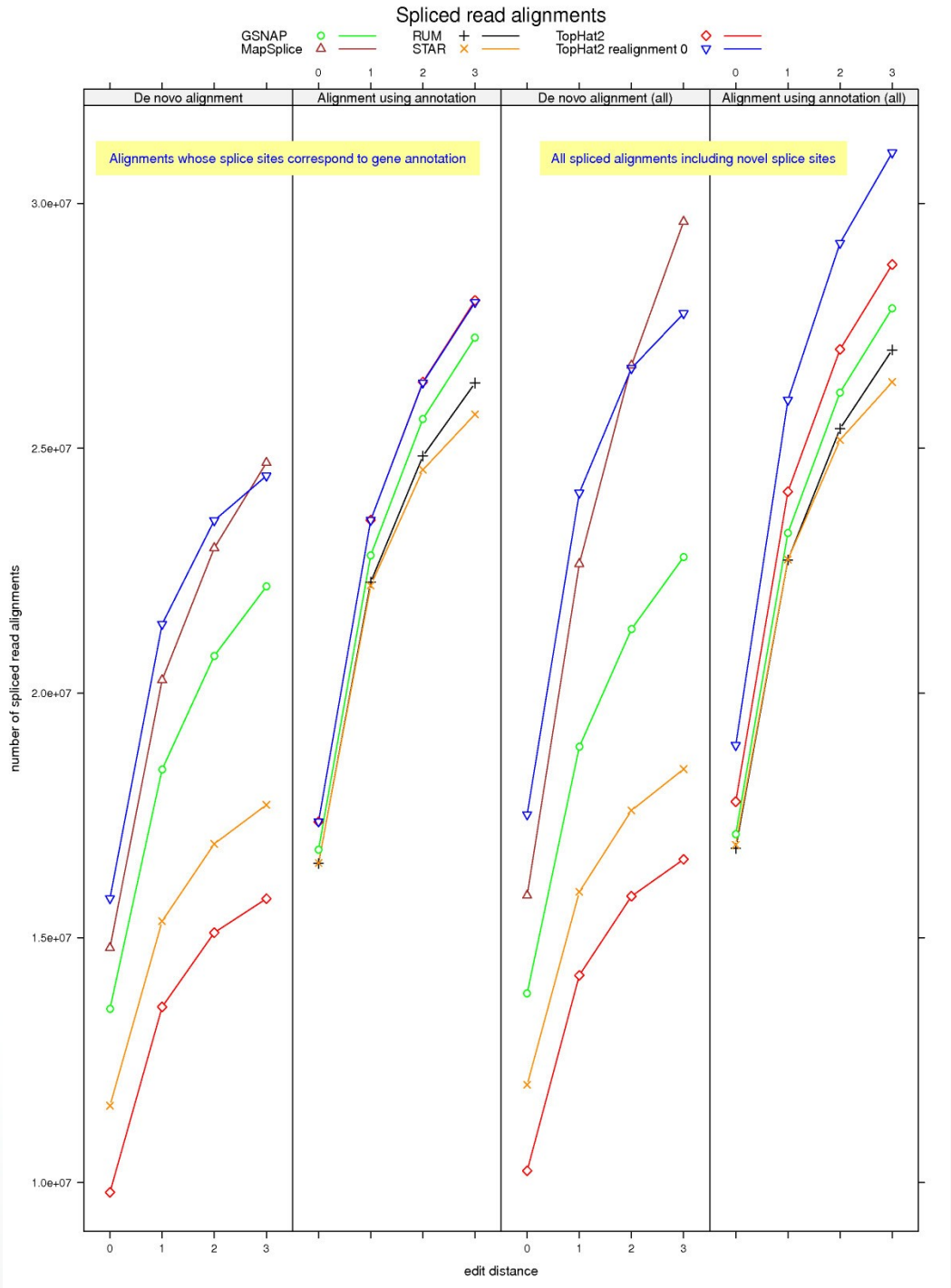


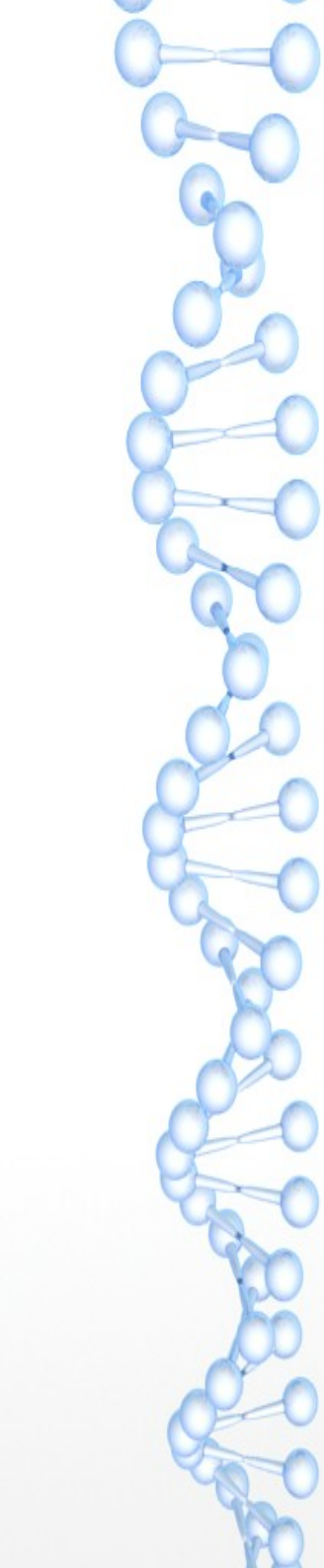
(2) Here, the read shown in red, which spans a splice junction, can be aligned end-to-end to a processed pseudogene.




(1) Transcriptome alignment (optional)
Unmapped reads
Reads
Read are aligned against transcriptome.
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Reads are aligned against genome.
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
Reads are split into smaller segments which are then aligned to the genome.
Genome index
(3-2) Identification of splice sites (including indels and fusion break points)
Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
Sequences flanking a splice site are concatenated and segments are aligned to them.
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.
(3-5) Re-alignment of reads minimally overlapping introns
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.
Read
Exons from annotated transcripts
Unannotated exons (novel transcripts)
Intron or intergenic region


Read alignments
GSNAP
MapSplice
RUM
STAR
TopHat2
TopHat2 realignment 0
De novo alignment
Alignment using annotation
number of read alignments
9e+07
8e+07
7e+07
6e+07
5e+07
0
1
2
3
edit distance


Spliced read alignments
GSNAP
MapSplice
RUM
STAR
TopHat2
TopHat2 realignment 0
De novo alignment
Alignment using annotation
De novo alignment (all)
Alignment using annotation (all)
Alignments whose splice sites correspond to gene annotation
All spliced alignments including novel splice sites
number of spliced read alignments
3.0e+07
2.5e+07
2.0e+07
1.5e+07
1.0e+07
0
1
2
3
edit distance



Cufflinks

Problem

- Transcript assembly
- Transcript abundance estimation

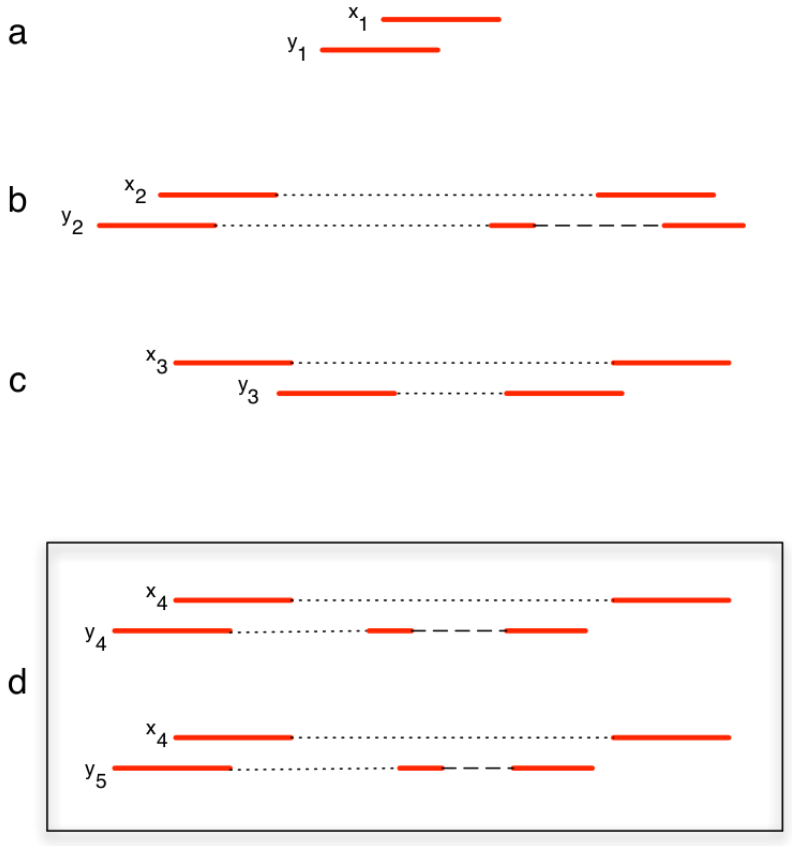
Assembling

1) Every fragment is consistent with at least one assembled transcript.
2) Every transcript is tiled by reads.
3) The number of transcripts is the smallest required to satisfy requirement (1).

Assembling

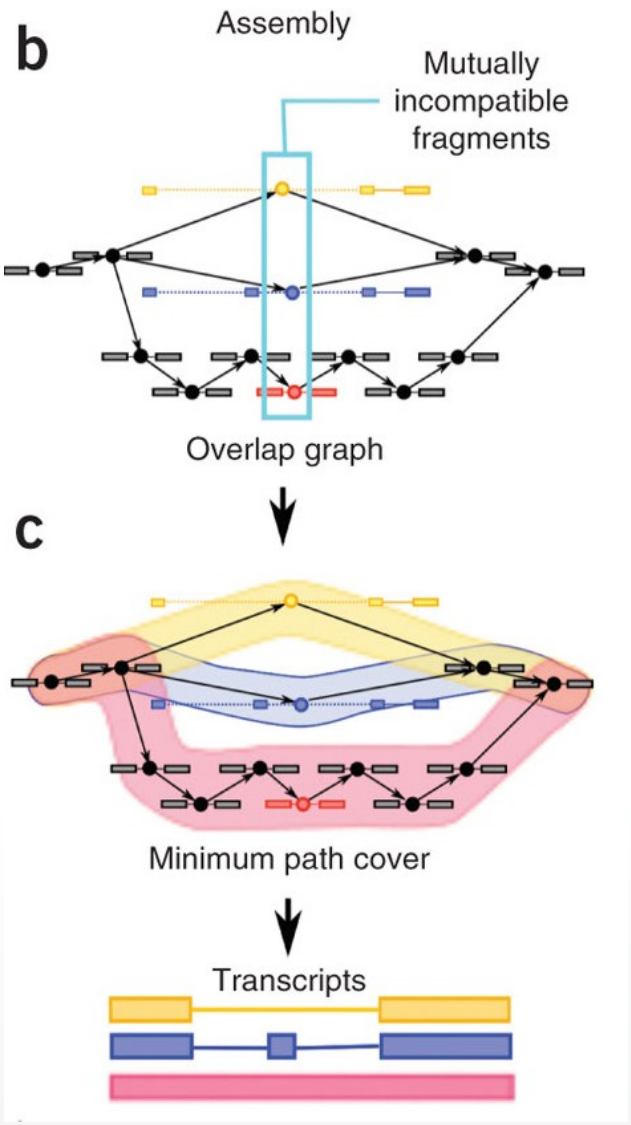
- Fragments were first divided into non-overlapping loci, and each locus was assembled independently of the others
- Two reads are compatible if their overlap contains the exact same implied introns (or none)
- Each fragment alignment was assigned a node in an “overlap graph” G .
- A directed edge (x , y) was placed between nodes x and y when the alignment for x started at a lower coordinate than y, the alignments overlapped in the genome, and the fragments were “compatible”

Compatible reads

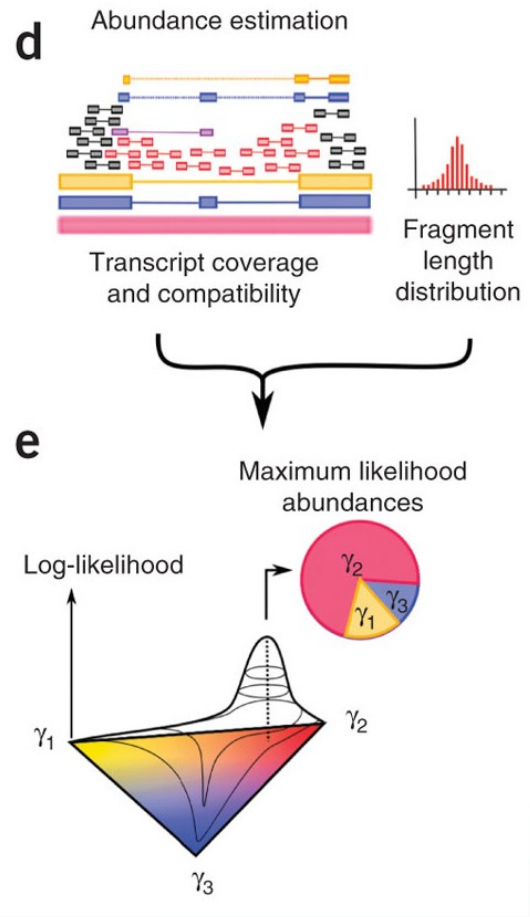


- Cufflinks then found a minimum path cover of G, meaning that every fragment node was contained in some path in the cover, and the cover contained as few paths as possible.

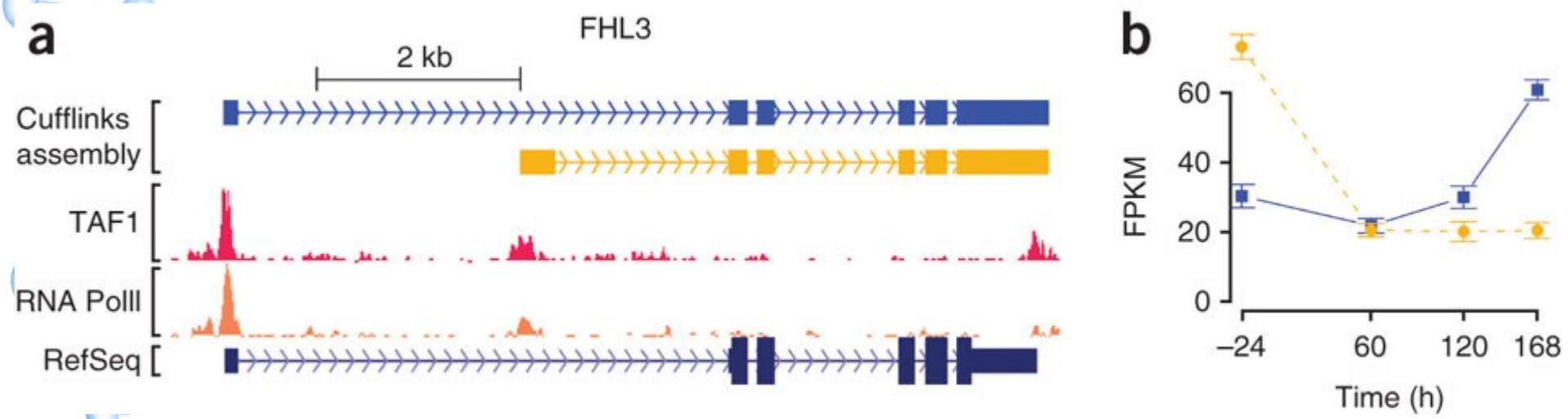
Assembly


b
Assembly
Mutually
incompatible
fragments
Overlap graph
c
Minimum path cover
Transcripts

Abundance


d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ1
γ2
γ3

New FHL3 isoform

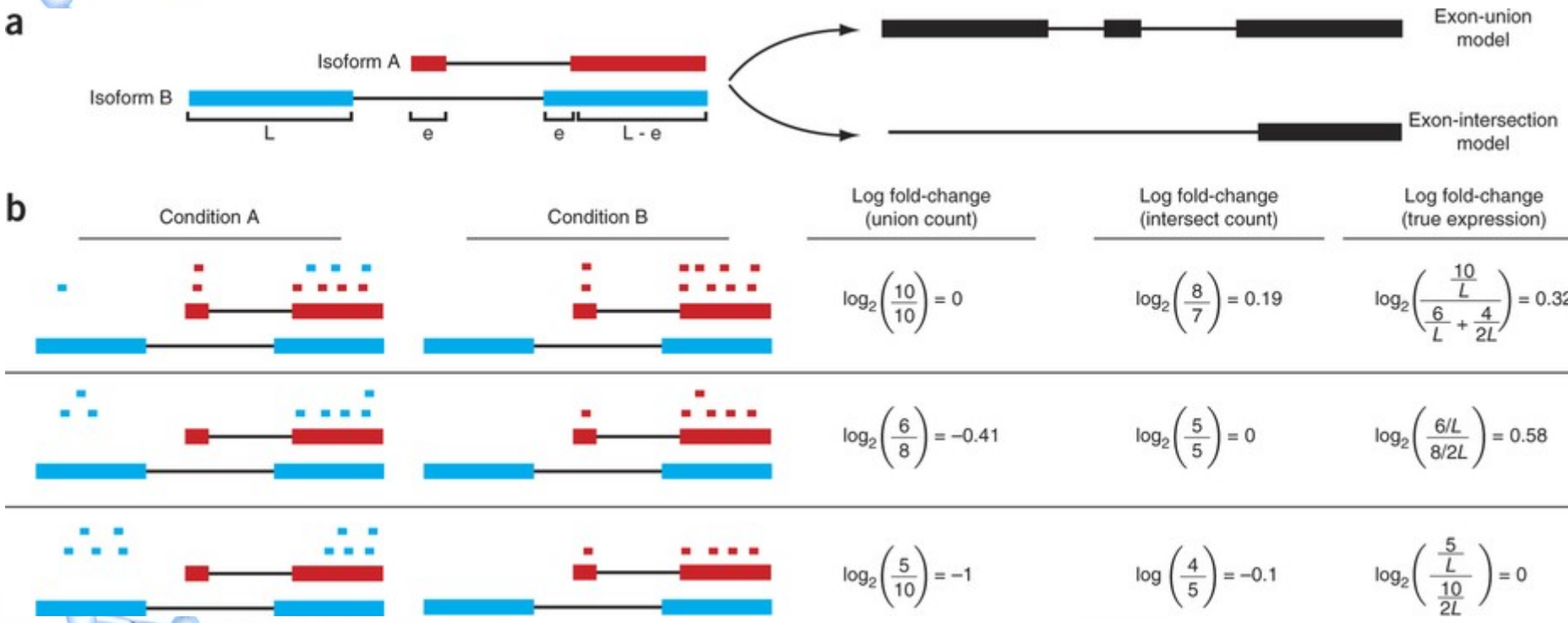


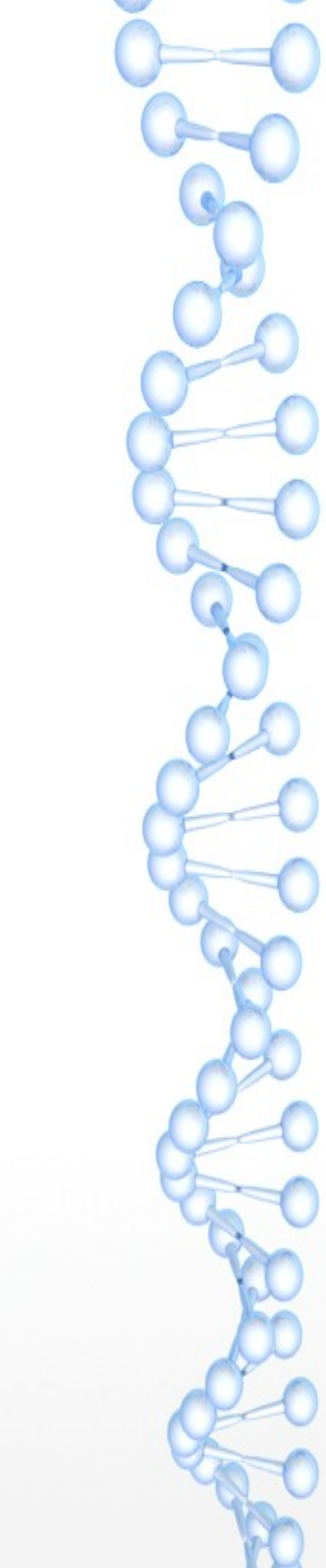
Cuffdiff 2

Current RNA-seq differential analysis methods focus on tackling one of two major challenges:

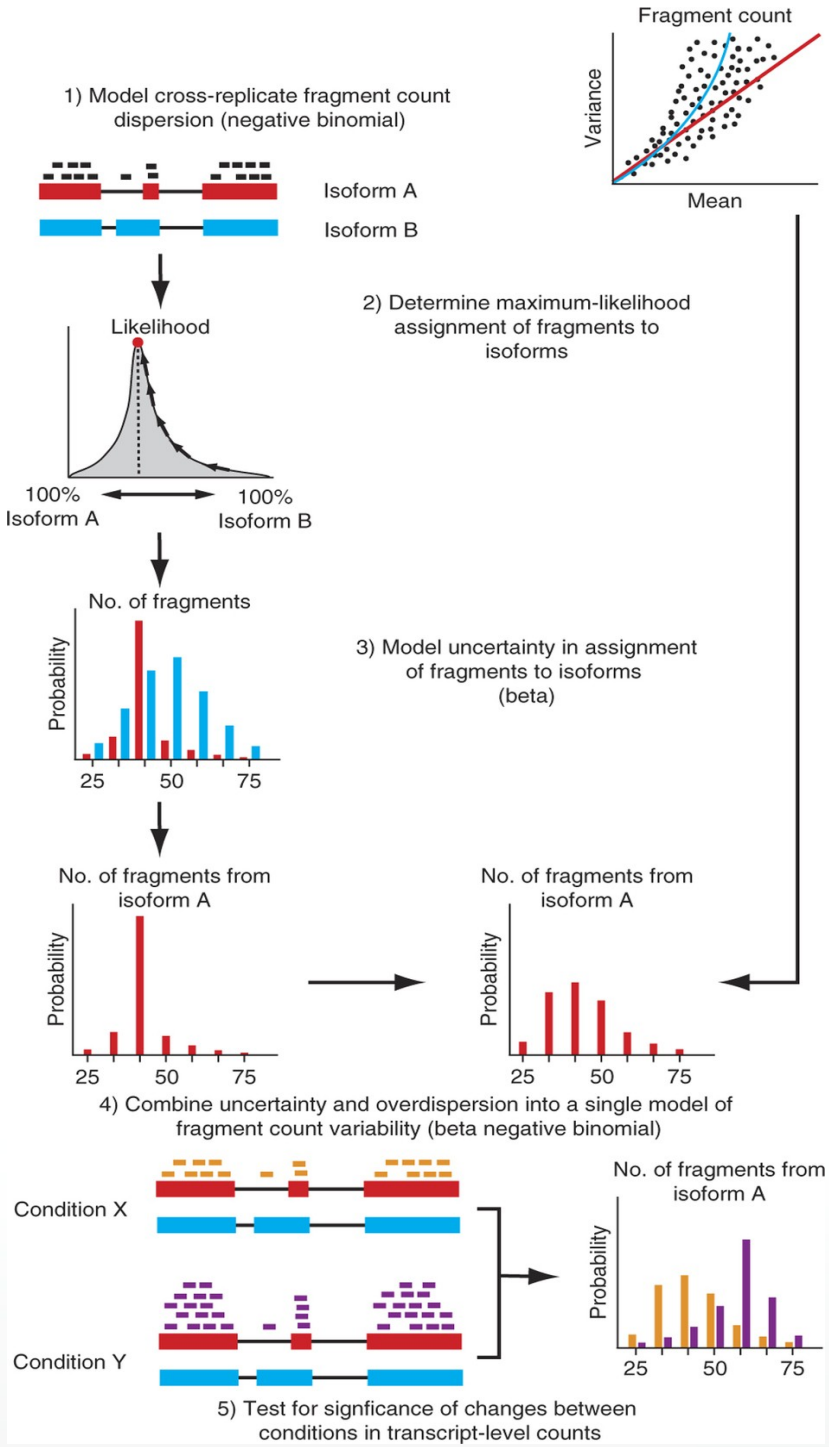
- The first is accurately deriving gene and isoform expression values from raw sequencing reads, which requires statistical computations at isoform-level resolution.
- The second is accounting for variability in measurements across biological replicates of an experiment.

Read counting problem





- Count uncertainty refers to the observation that in RNA-seq experiments it is common for up to 50% of reads to map ambiguously to different transcripts


Fragment count
Variance
Mean
1) Model cross-replicate fragment count dispersion (negative binomial)
Isoform A
Isoform B
Likelihood
100% Isoform A
100% Isoform B
2) Determine maximum-likelihood assignment of fragments to isoforms
No. of fragments
Probability
25
50
75
3) Model uncertainty in assignment of fragments to isoforms (beta)
No. of fragments from isoform A
Probability
25
50
75
No. of fragments from isoform A
Probability
25
50
75
4) Combine uncertainty and overdispersion into a single model of fragment count variability (beta negative binomial)
Condition X
Condition Y
No. of fragments from isoform A
Probability
25
50
75
5) Test for signficance of changes between conditions in transcript-level counts

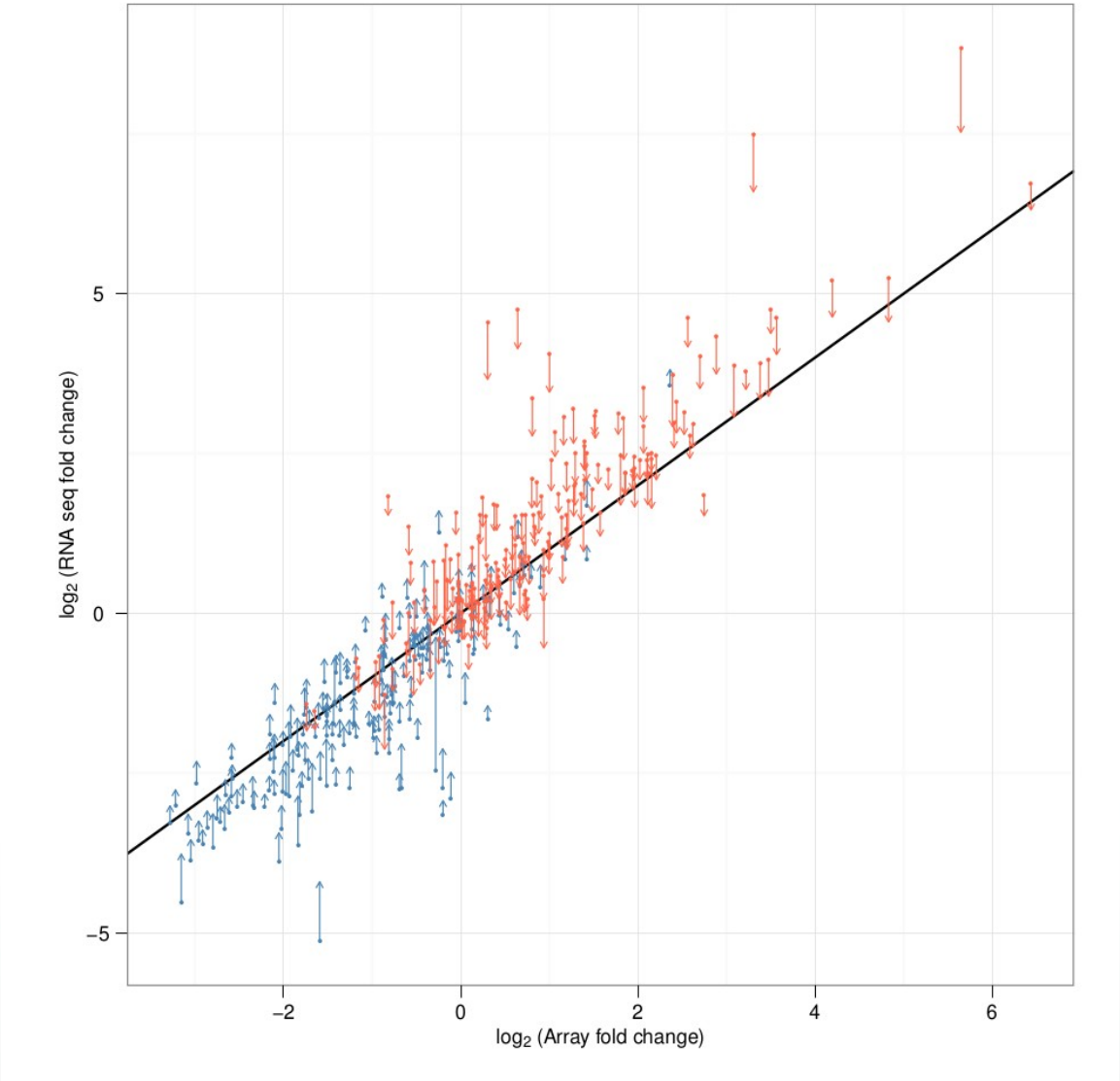
Cuffdiff 2

- Cuffdiff 2 estimates expression at gene- and transcript-level resolution
- Variance in the expression levels
- covariances between isoforms of the same gene from replicate experiments.
- This allows it to accurately estimate gene expression and perform differential analysis at gene-level

Response to loss of HOXA1

- Cuffdiff 2 – derived changes in gene expression in response to HOXA1 knockdown strongly agreed with values from microarrays
- Cuffdiff 2 improved concordance with the array measurements by 15% compared with the change in raw count

Response to loss of HOXA1



Gentleman's set

- Samtools
- Bowtie 2
- Bowtie 2 index
- Annotations known transcripts (GFF)
- TopHat 2
- Cufflinks 2

Articles

- TopHat: discovering splice junctions with RNA-Seq
- TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions
- Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms
- Differential analysis of gene regulation at transcript resolution with RNA-seq