

# Mapping RNA-seq reads

Alexander Dobin  
CSHL

random]plasmd

bioRxiv preprint doi: <https://doi.org/10.1101/000000>; this version posted January 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Chemically, RNA is a linear polymer of nucleotides. Each nucleotide consists of a phosphate group, a sugar, and a nitrogenous base. The bases are adenine (A), guanine (G), cytosine (C), and uracil (U). The bases are connected to each other by hydrogen bonds, forming the double helix structure of DNA. RNA is single-stranded, but it can form local secondary structures through base pairing. The sequence of bases in RNA determines the sequence of amino acids in a protein, which is then translated into a functional protein.

RNA-seq is a high-throughput sequencing technology that allows for the quantification of RNA levels in a sample. It involves the conversion of RNA into a library of cDNA fragments, which are then sequenced using high-throughput sequencing (HTS) technology. The resulting sequencing reads are mapped to a reference genome, and the abundance of each transcript is estimated based on the number of reads that map to it.

Mapping RNA-seq reads to a reference genome is a complex task due to the high degree of similarity between many transcripts. This is particularly challenging for genes that contain alternative splicing, where different exons are joined together in different ways to produce different isoforms of a protein. Mapping reads to the correct isoform is essential for accurate quantification of gene expression.

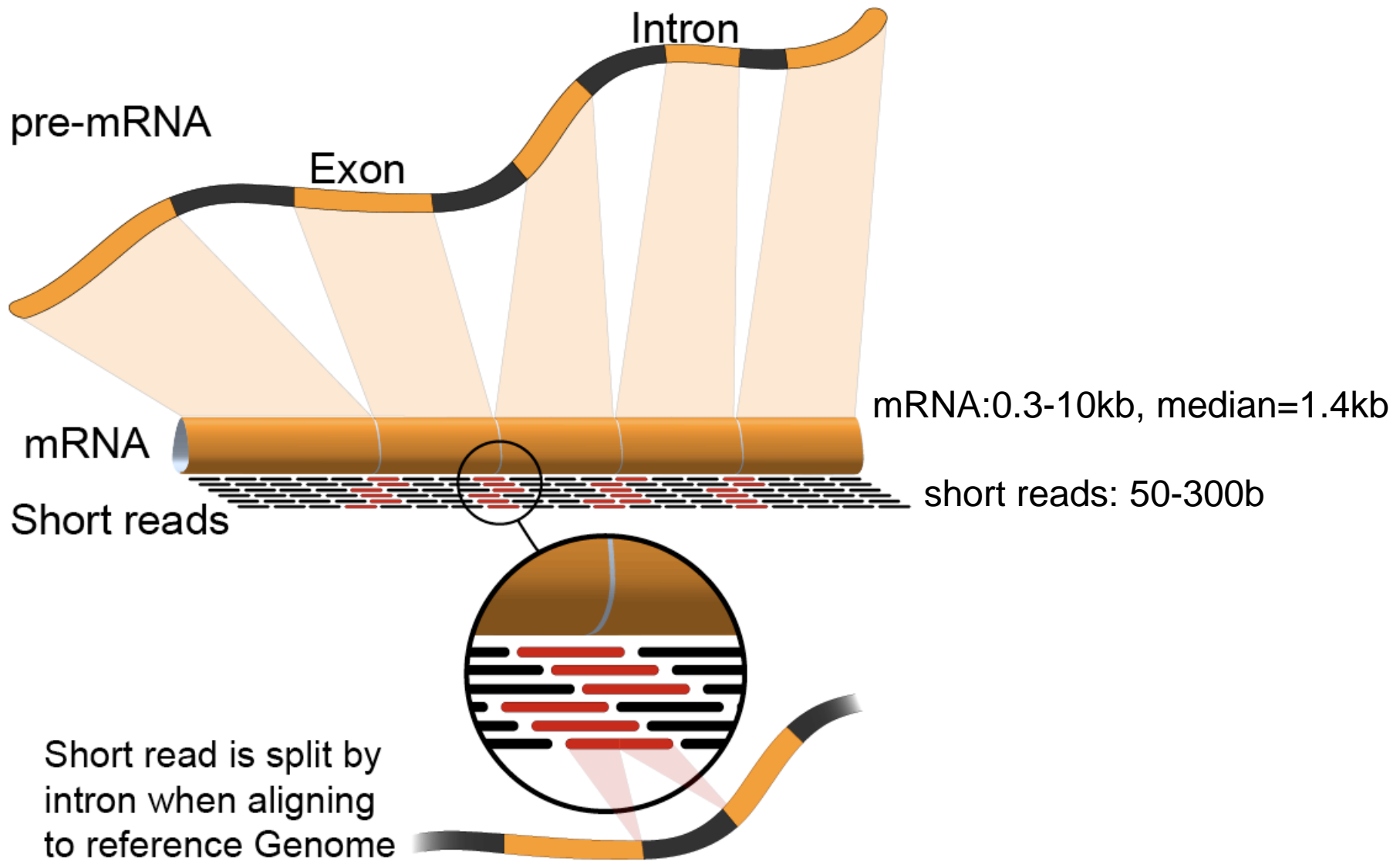


# Outline

- Introduction: RNA-seq technology, analyses, pipelines
- Mapping of RNA-seq reads to the genome
- STARtools: basic post-mapping analyses

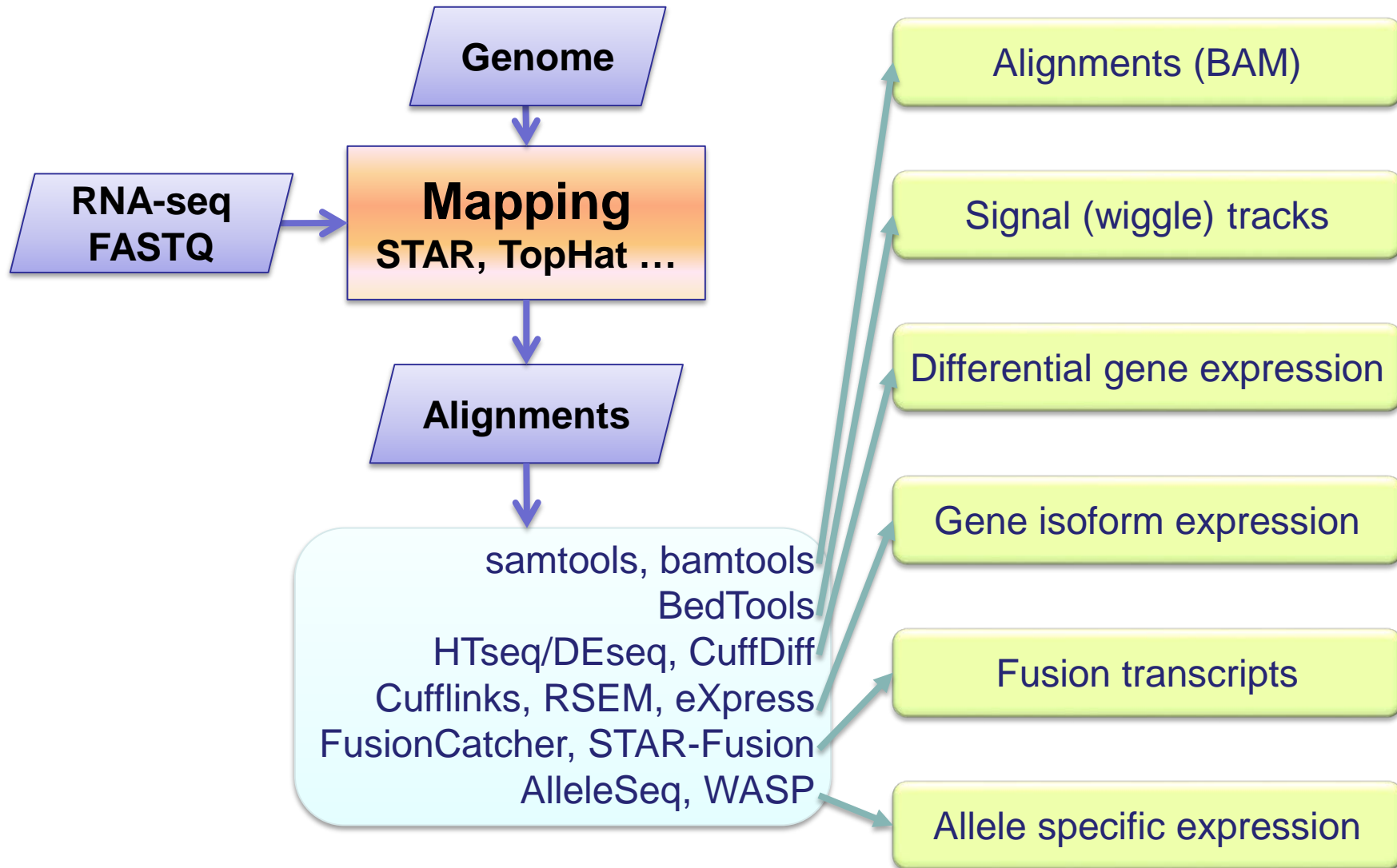
# Introduction: RNA-seq technology, analyses, pipelines

# RNA-seq



<https://en.wikipedia.org/wiki/RNA-Seq#/media/File:RNA-Seq-alignment.png>

# RNA-seq pipeline



# Mapping RNA-seq reads sto the genome

# Short reads aligners

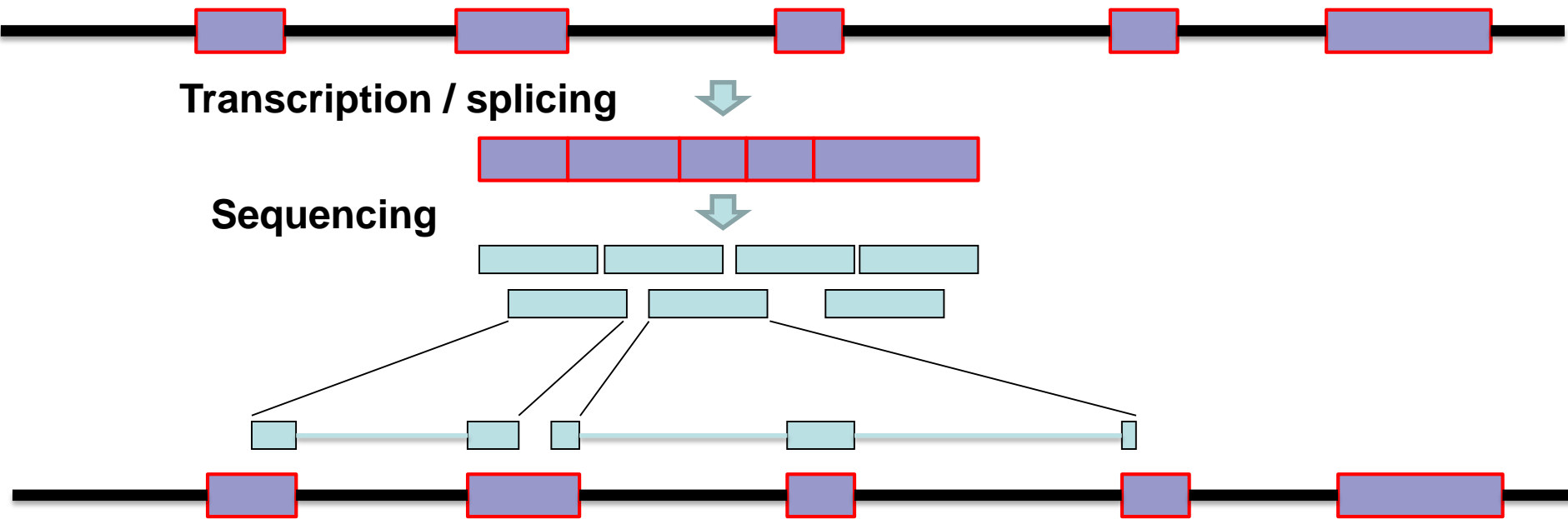
DNA	
BWA	
Bowtie(2)	

RNA	
TopHat(2)	Slow
STAR	Fast, many features
HISAT	Fast, low RAM
GSNAP	Slow, accurate

# Challenges of RNA-seq mapping

Transcription / splicing

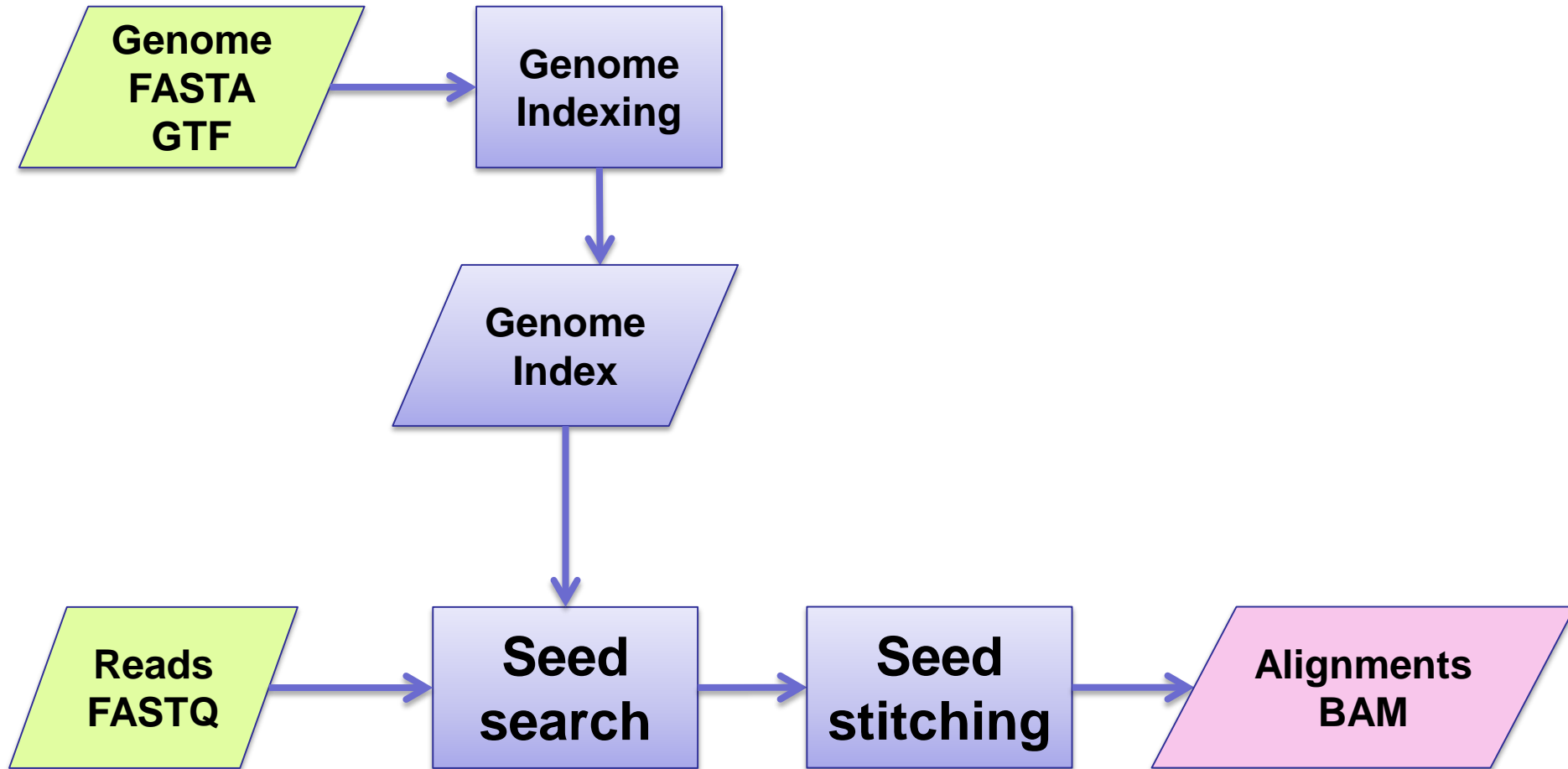
Sequencing



- **Most long RNAs are spliced**
- Short reads map non-contiguously, may contain >1 splice junction
- Large introns: ~0.1-1,000 kb in mammals
- Multi-mappers are important (expression of repeats, paralogs, pseudogenes)
- Highly expressed loci create mapping artifacts
- Genomic variations: SNPs, indels, SVs
- RNA editing

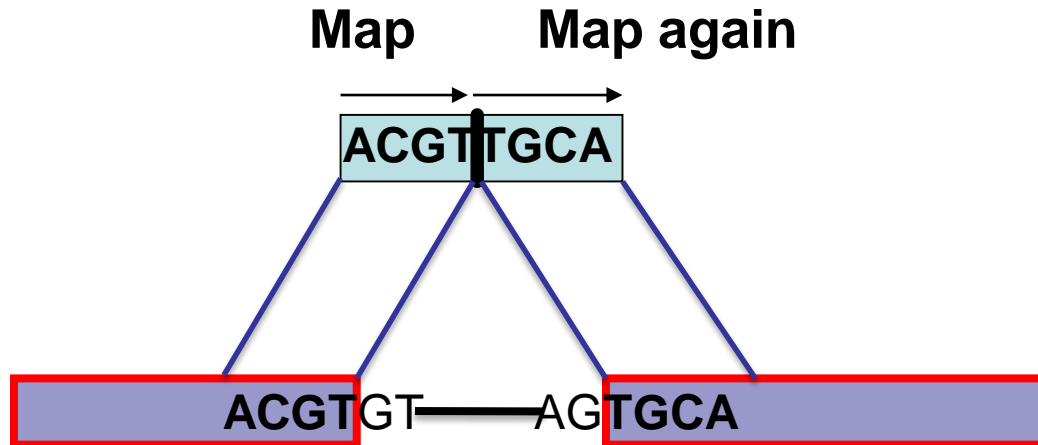


# STAR mapping workflow

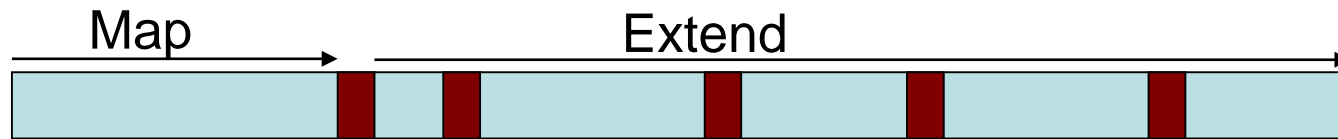


# Seed search: basic idea

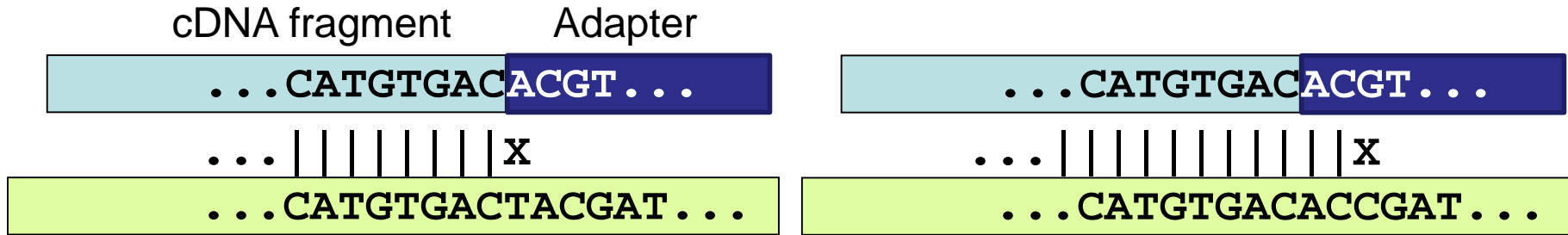
- “Consecutive maximal exact prefix search”
- MEM, Maximal Exact Match: Mummer, MAUVE
- BWA-MEM, Cushaw2, GEM



# Mismatches and tails



# Adapter trimming



Locus 1:

only the cDNA sequence  
maps to the genome

Locus 2:

cDNA sequence + 2 adapter bases  
map to the genome

- Adapter at 3' of the read sequence if  
fragment ("insert") length < sequence length
- Untrimmed adapter can turn multi-mappers into unique mappers
- Trimming software: Cutadapt, Trimmomatic, FASTX, etc.  
take care not to mess up the read order for paired-end reads
- Basic aggressive 3' adapter trimming in STAR with

```
--clip3pAdapterSeq <sequence> --clip3pAdapterMMp 0.1
```

# Soft-clipping vs. End-to-End alignment



True alignment: spliced

Hard to find because of short overhang: 3b will randomly occur every 64b in the genome

End-to-end alignment: 3MM

May be filtered out

End-to-End alignment to the pseudogene: 2MM

**False alignment!**

**Soft-clip unmatched bases**

Soft-clipping penalty < MM penalty

Also helps to catch A-tails, end modifications, adapters, poor quality tails

gene  
|||||||xxx

CATGTGACTAT

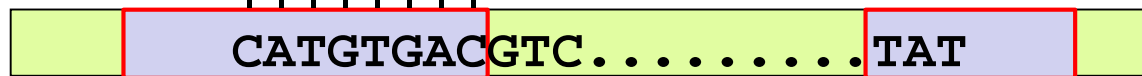
||x|||x|||



pseudogene

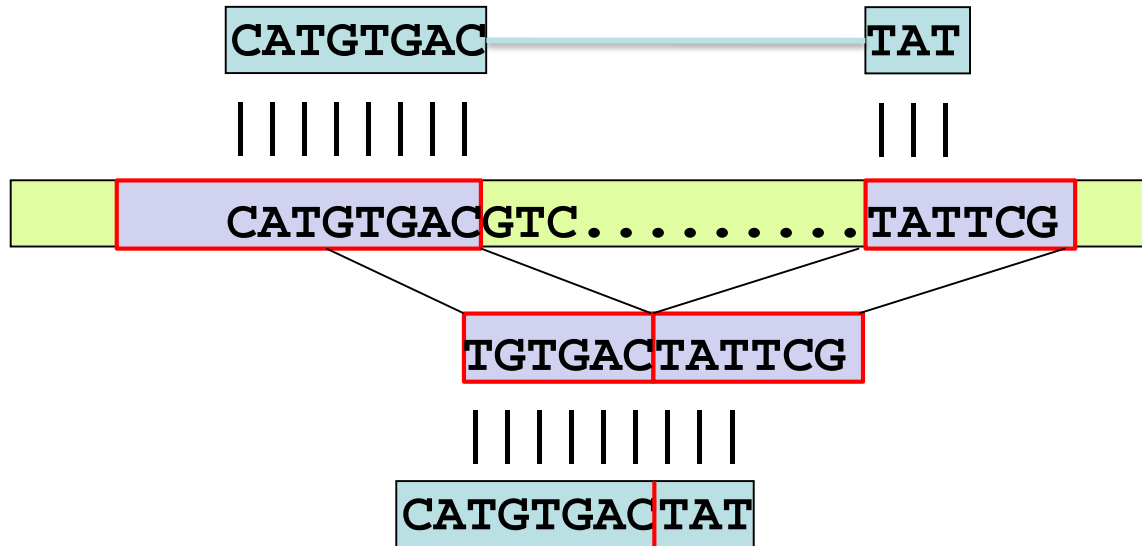
CATGTGACTAT

|||||||SSS



gene

# Mapping short splice overhangs



True alignment: spliced

Hard to find because of short overhang: 3b will randomly occur every 64b in the genome

## Solution:

use annotated junctions

STAR concatenates exon sequences and adds them to the search space

%	BLAT	Tophat	STAR	STAR + Annot
<i>base FPR</i>	2.9	5.4	2.0	0.1
<i>base FPR: wrong loci</i>	2.7	5.3	1.9	0.1
<i>base FPR: wrong locus, missed splice</i>	2.2	4.5	1.8	0.1
<u>% of all reads mapping to processed pseudogenes</u>				
<i>all</i>	0.7	1.6	0.8	0.1
<i>False Positive, wrong locus, missed splice</i>	81.8	82.3	71.1	26.0

~80% of false positive alignments arise from alignments missing a splice junctions and mapping to a wrong locus

~30% of false positive alignments map to processed pseudogenes

**~80% of pseudogene alignments are false positive**

# 2-pass mapping

CATGTGACTAT

|||||||sss

CATGTGAC.....TATTCG

|||||||

CATGTGAC

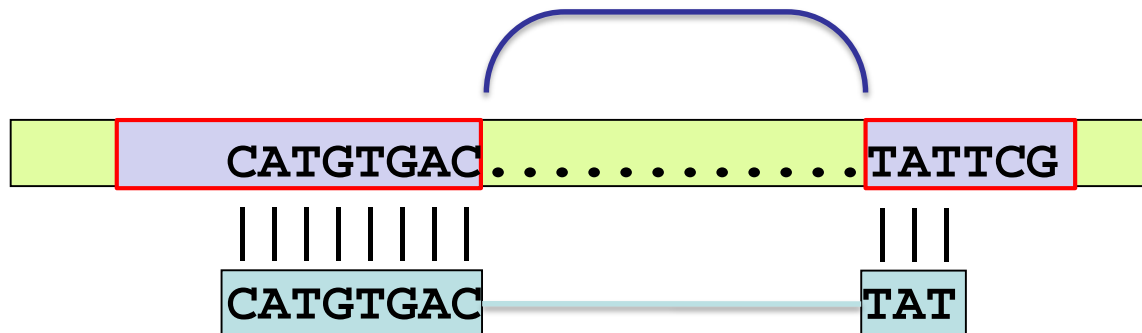
|||||||

TATTCG

## 1<sup>st</sup> pass:

Reads with short overhangs are soft-clipped

Read with long overhangs identify novel junctions



## 2<sup>nd</sup> pass:

Junctions from the 1<sup>st</sup> pass are added to the search space

Read with short overhangs map spliced to novel junctions

# Increasing search sensitivity

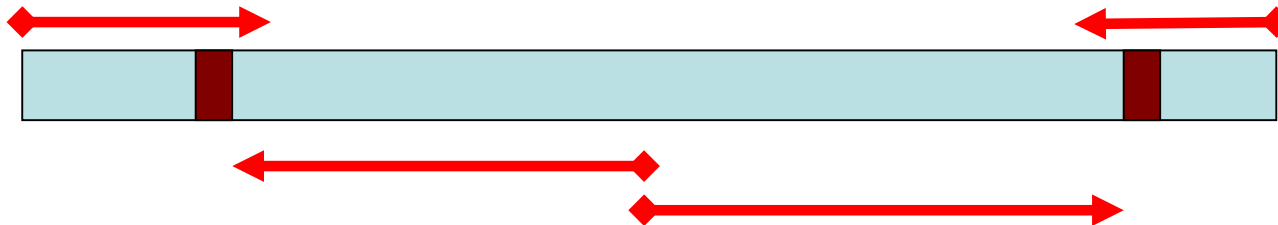
- One mismatch near one of the ends

Seed is too short – max exact search does not stop at the mismatch and maps to a wrong locus



Solution: search backwards from the other end

- Two mismatches near the ends



Solution: start search from the middle of the read

- `--seedSearchStartLmax <N>`

user defined parameter to start search as often as needed

- Reducing N will increase sensitivity, but reduce mapping speed
- Default N= 50b, works well for Illumina reads



# Anchor seeds and windows

- **--seedMultimapNmax <N>**

All seeds that map <N> times are recorded: =10,000 by default  
10-mers map 10,000 on average in human genome

- **--winAnchorMultimapNmax <N>**

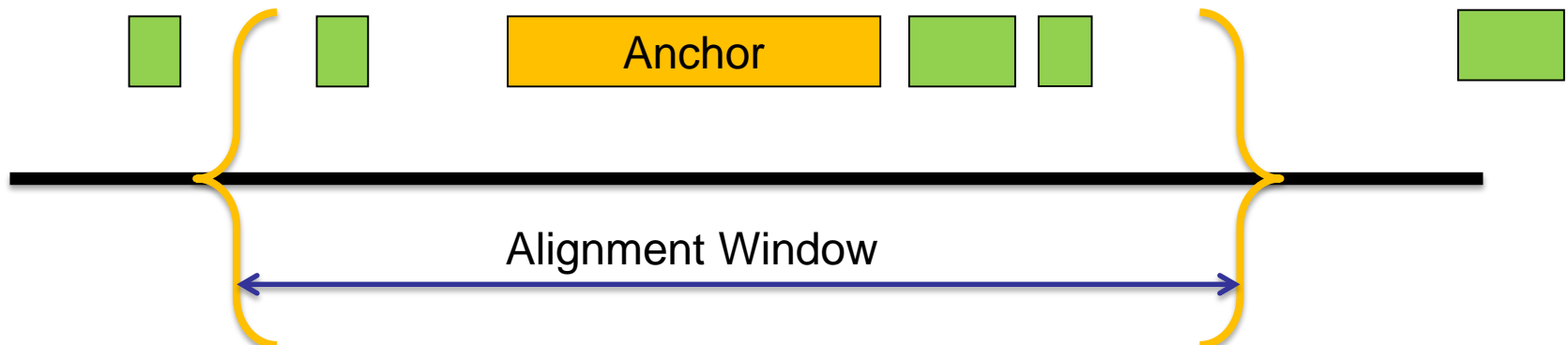
“Anchors”: seeds that map <N> times: =50 by default

- “Alignment windows”: genome regions around anchors

All seeds inside alignment windows are stitched together

Size of the window ~ maximum intron size, ~1Mb for human

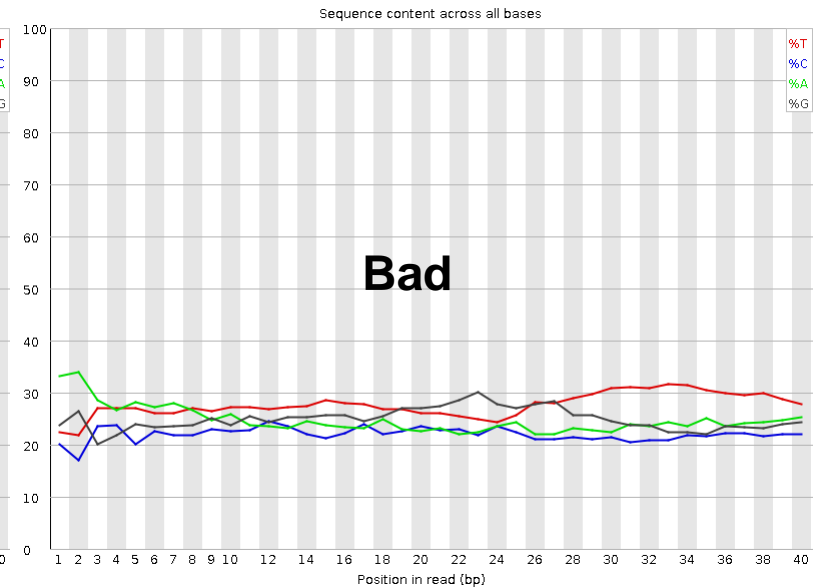
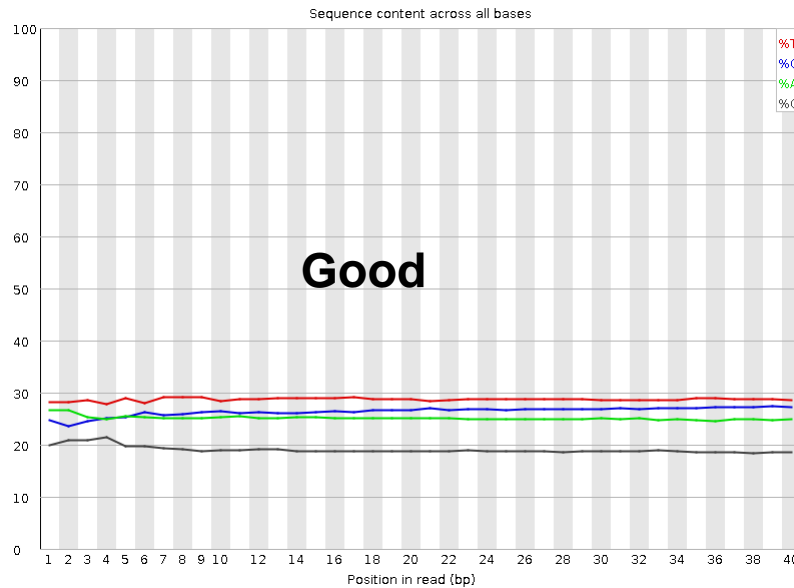
**--alignIntronMax <N>, --alignMatesGapMax <N>**



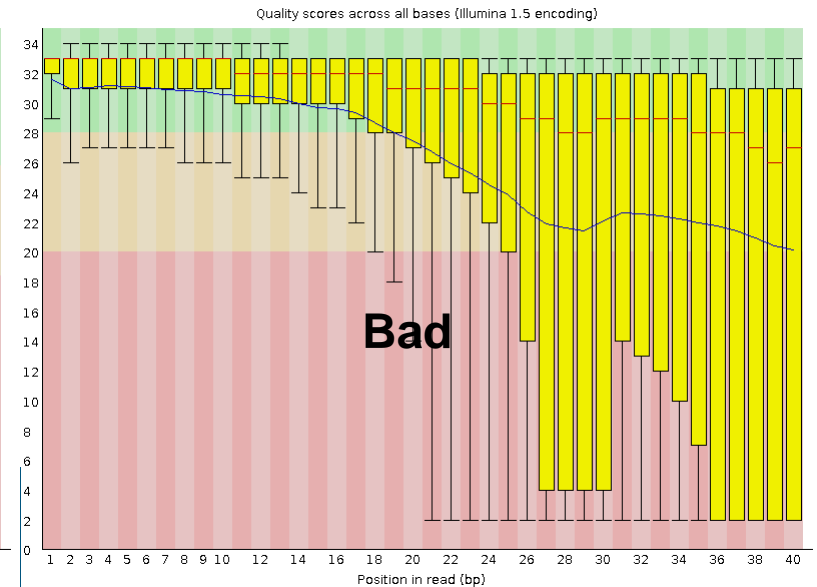
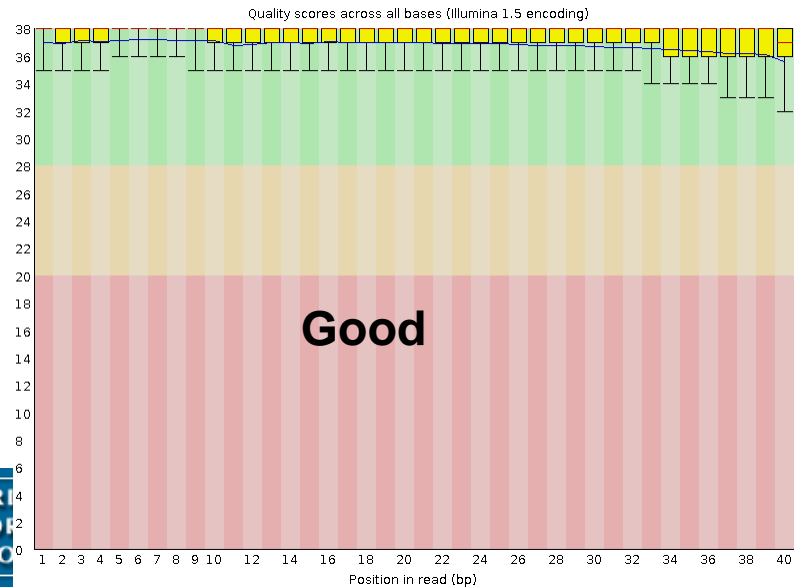
# Pre-mapping QC with FASTQC

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

**Per-base  
sequence  
content**



**Per-base  
sequence  
quality**



# Understanding mapping results

## STAR's Log.final.out file:

Average input read length	202
---------------------------	-----

### UNIQUE READS:

Uniquely mapped reads %	90.08%
-------------------------	--------

Average mapped length	201.98
-----------------------	--------

Mismatch rate per base, %	0.30%
---------------------------	-------

Deletion rate per base	0.02%
------------------------	-------

Insertion rate per base	0.01%
-------------------------	-------

### MULTI-MAPPING READS:

% of reads mapped to multiple loci	3.55%
------------------------------------	-------

% of reads mapped to too many loci	0.02%
------------------------------------	-------

### UNMAPPED READS:

% of reads unmapped: too many mismatches	2.82%
--	-------

% of reads unmapped: too short	3.44%
--------------------------------	-------

% of reads unmapped: other	0.08%
----------------------------	-------

# Why my mapping rate is low?

## Possible problem

- File formatting mix-up:  
read1/read2 order broken
- Poor quality of sequencing
- Tails  
Poor quality  
Adapter - short insert
- rRNA insufficient depletion
- Contamination with other species

## Checks/Solutions

Ensure the same order of read1/2  
Map read1/2 separately

Plot quality scores vs read length

Trim by quality  
Trim adapter

Include rRNA sequences in the reference

BLAST unmapped reads

# STARtools: post-mapping analyses at no extra cost

# RNA-seq pipeline

RNA-seq  
FASTQ

STAR, TopHat  
samtools, bamtools  
BedTools  
HTseq/DEseq, CuffDiff  
Cufflinks, FLUX, RSEM, eXpress  
TopHat Fusion, FusionCatcher  
AlleleSeq

- Bottlenecks
- Input/output compatibility
- Software versioning
- Reproducibility

Alignments (BAM)

Signal (wiggle) tracks

Differential gene expression

Gene isoform expression

Fusion transcripts

Allele specific expression

# STARtools

RNA-seq  
FASTQ

STAR  
mapping

**Conversion:** Sorted BAM  
Signal (wiggle) tracks  
Duplicate removal

**Quantification:** Count reads per gene  
Transform to  
transcriptomic coordinates

**Fusion:**  
(Brian Haas) Detect fusion transcripts

**STAR-WASP:** Variation-aware mapping  
**Allele-specific expression**

**Solo:** Single-cell RNA-seq

**Long:** Map long noisy reads:  
PacBio, Oxford Nanopore

# Quantification tools

- Count how many reads overlap any of the isoforms of each gene
- Used in differential gene expression analysis
- HTseq: produces counts from BAM alignments
- 109M reads, 2x101b:

STAR map: 22 min

HTseq count: 250 min

**STAR map+count: 22 min**

**--quantMode GeneCounts**

- eXpress, RSEM ...
- maximum likelihood estimation of isoform expression
- need alignments in “transcriptomic” coordinates
- mapping to transcriptome with BWA, Bowtie...
- STAR maps to the genome,

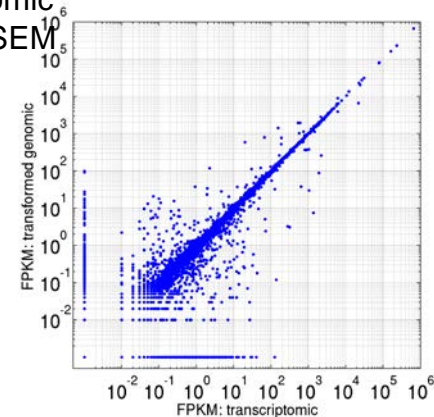
and at the same time

**converts genomic alignments to transcriptomic**

**no extra computational time required**

**--quantMode TranscriptomeSAM**

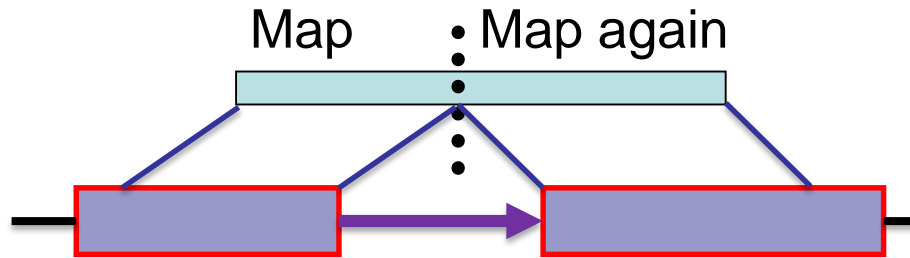
Expression  
genomic =>  
transcriptomic  
STAR / RSEM



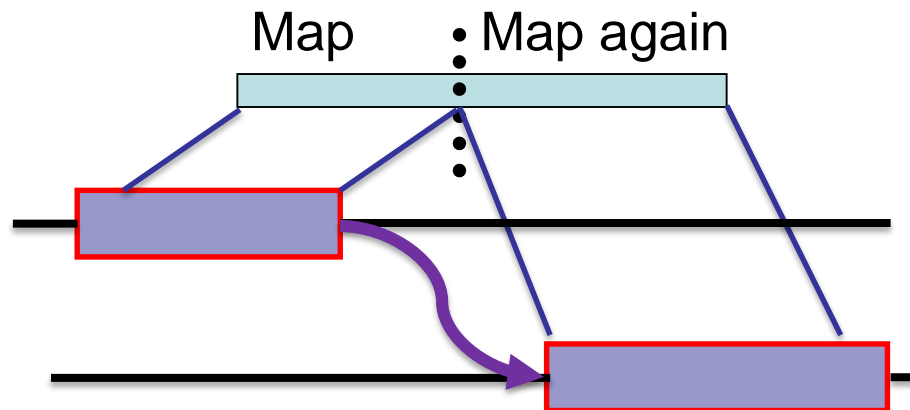
Expression using  
transcriptomic alignments  
Bowtie / RSEM



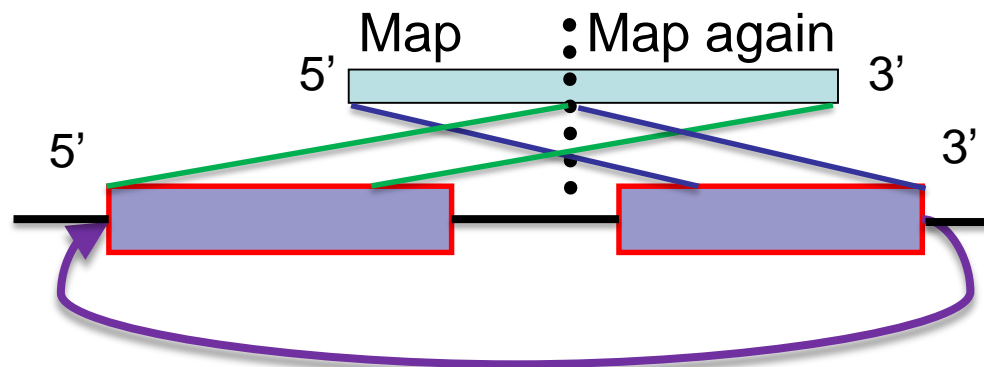
# Chimeric and circular junctions



**Linear junction**



**Chimeric junction**



**Circular junction**

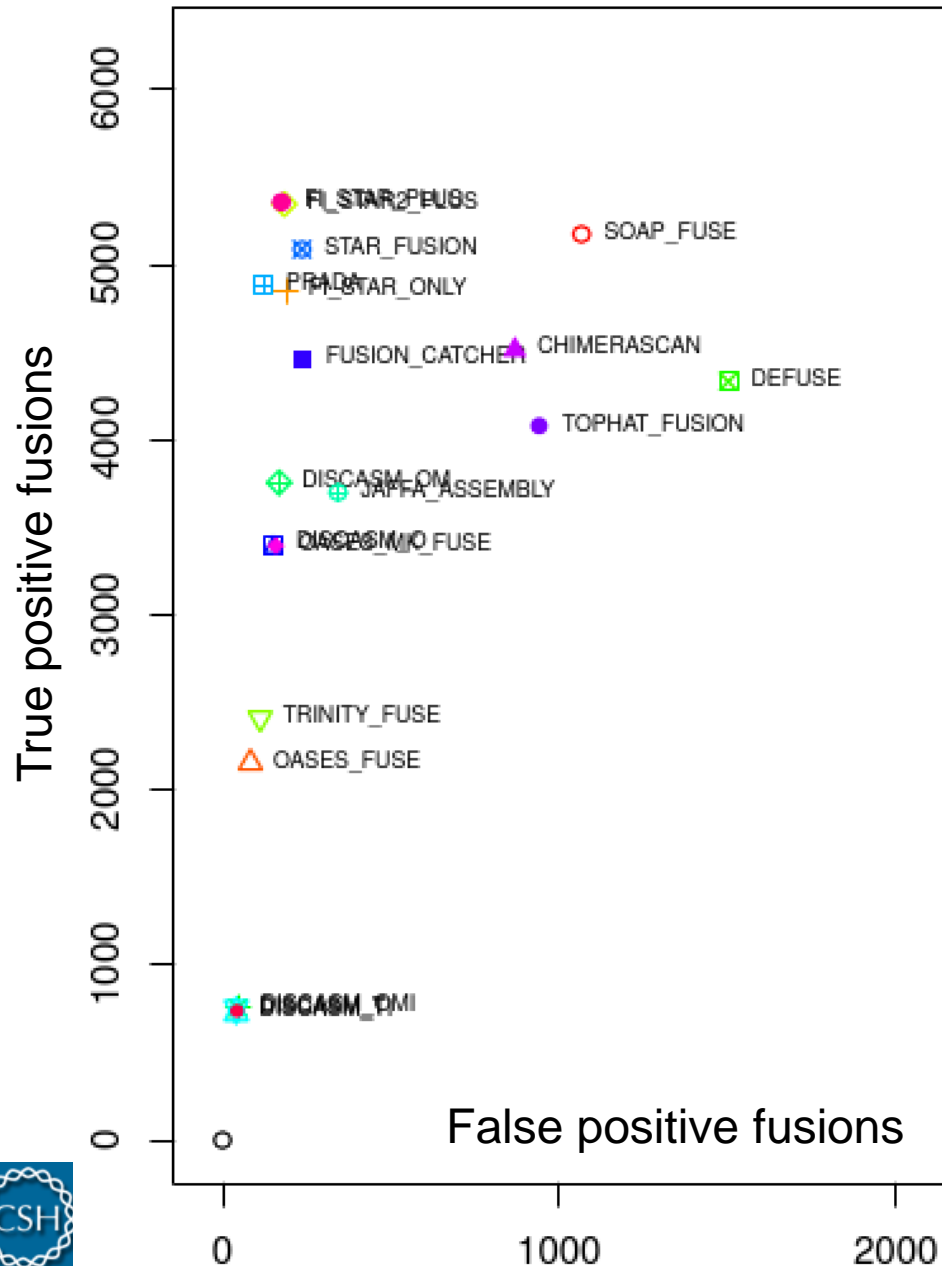
# STAR-Fusion

## STAR-Fusion FusionInspector

developed by **Brian Haas**  
(Broad Institute)

Analyzes STAR chimeric  
alignments to detect fusion  
transcripts in RNA-seq data

<https://github.com/STAR-Fusion/STAR-Fusion>



# Summary

- RNA-seq pipelines
- RNA-seq alignment challenges
- Tweaking mapping parameters
- Post-mapping: STARtools

<https://github.com/alexdobin/STAR>

<https://groups.google.com/forum/#!forum/rna-star>