Mapping RNA-seq reads

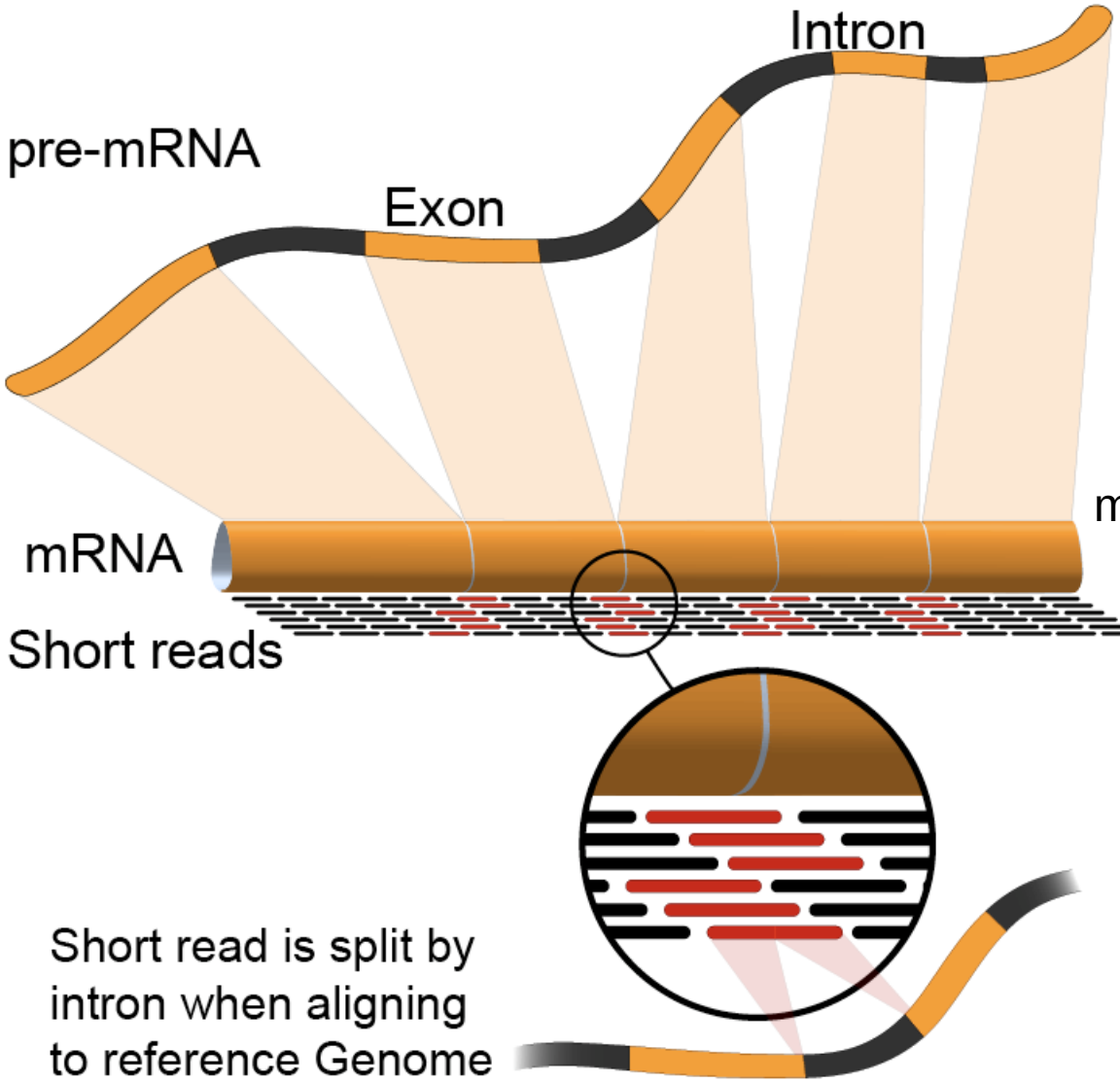
Alexander Dobin
CSHL

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

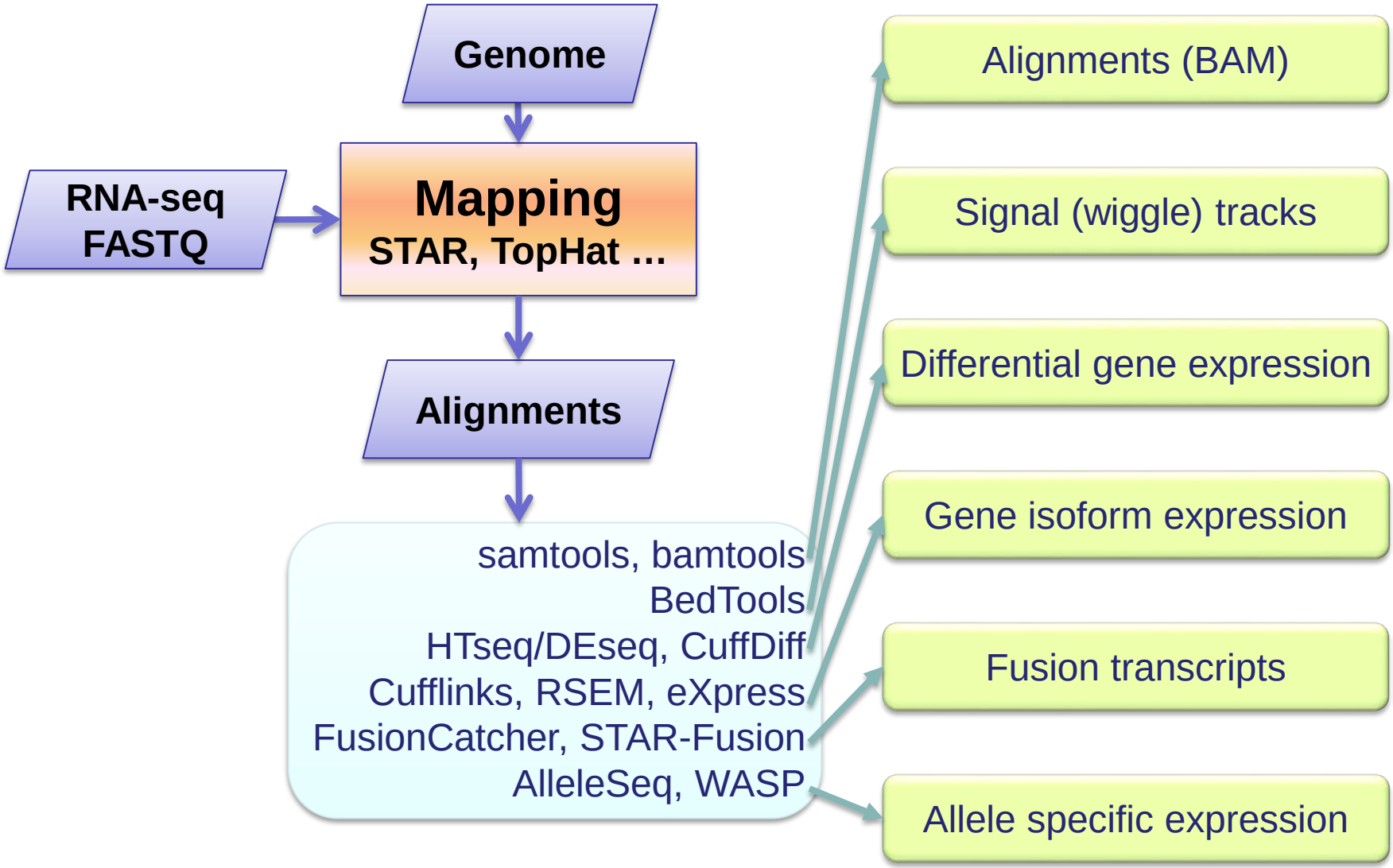


mRNA:0.3-10kb, median=1.4kb

short reads: 50-300b

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

RNA-seq pipeline



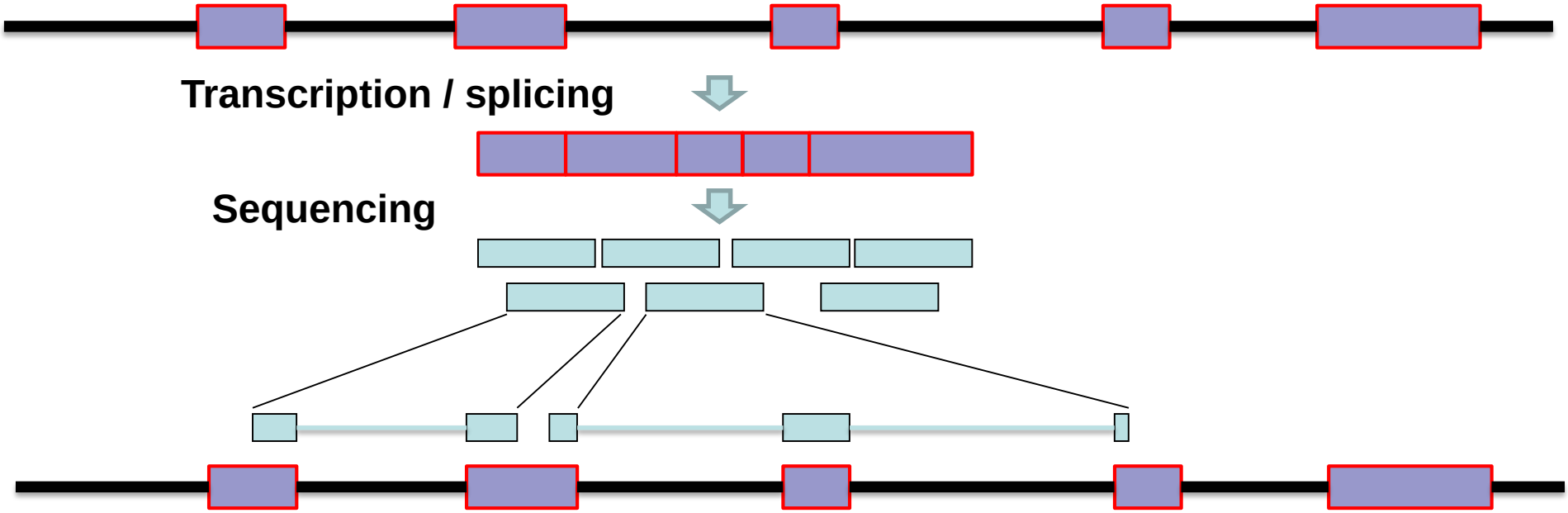
Mapping RNA-seq reads sto the genome

Short reads aligners

DNA	
BWA	
Bowtie(2)	

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

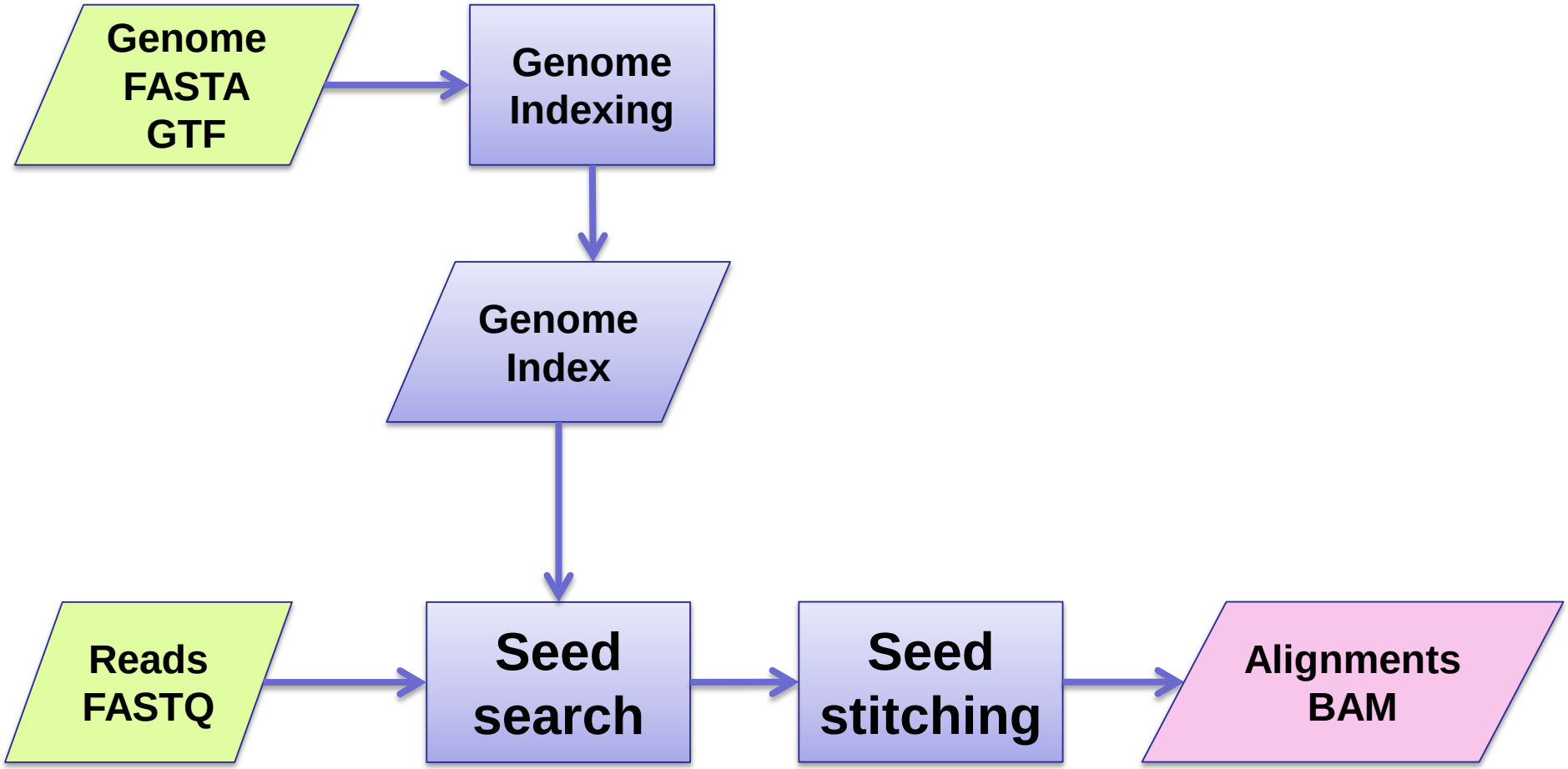
Challenges of RNA-seq mapping



- **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

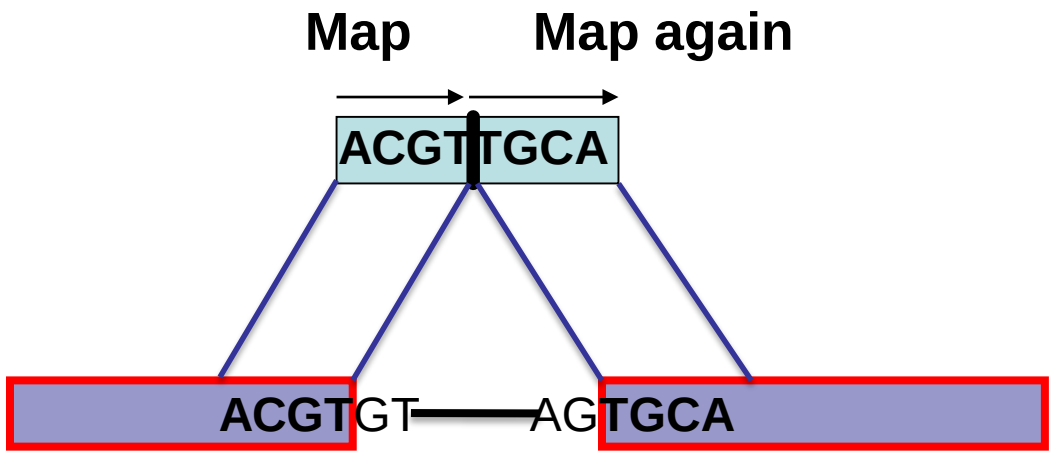
STAR mapping workflow



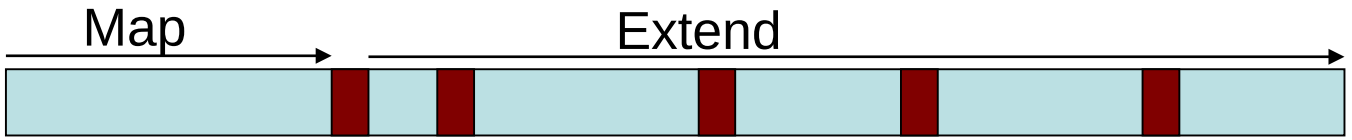
Seed search: basic idea

- “Consecutive maximal exact prefix search”

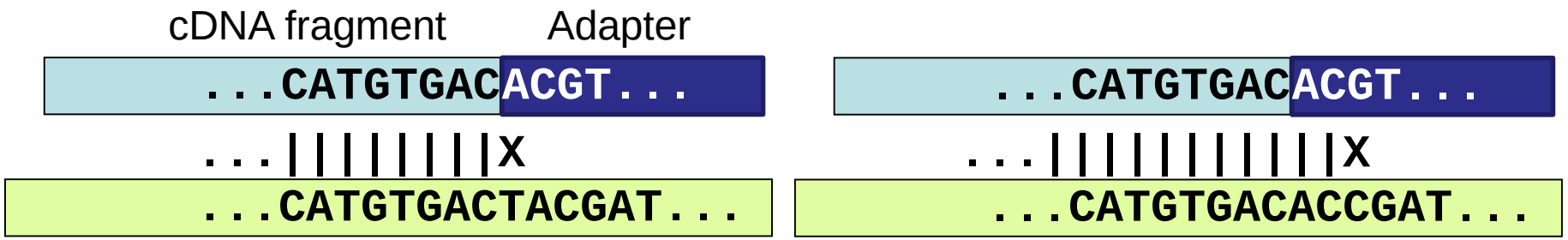
- MEM, Maximal Exact Match: Mummer, MAUVE
- BWA-MEM, Cushaw2, GEM



Mismatches and tails



Adapter trimming

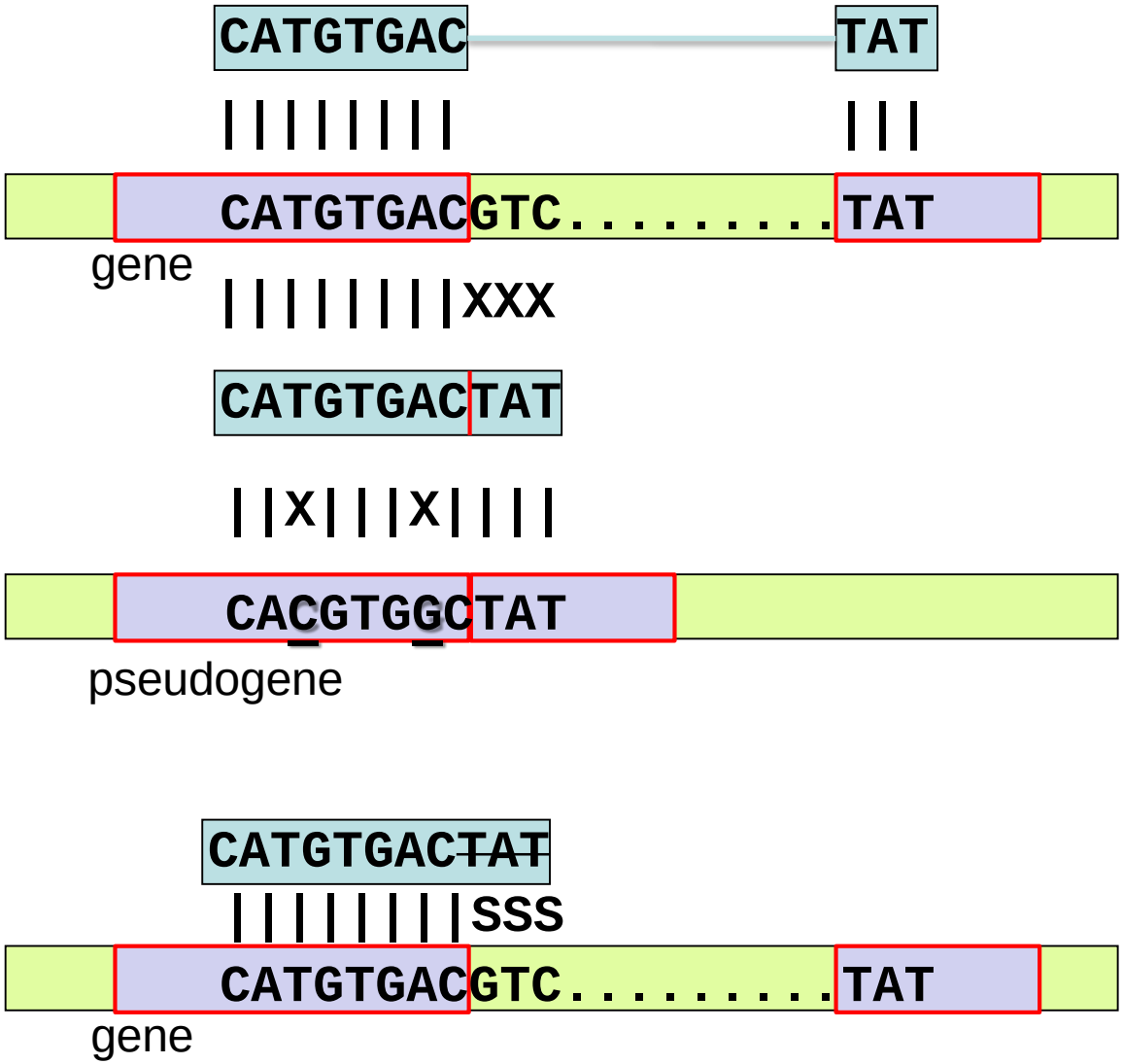


Locus 1:
only the cDNA sequence
maps to the genome

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

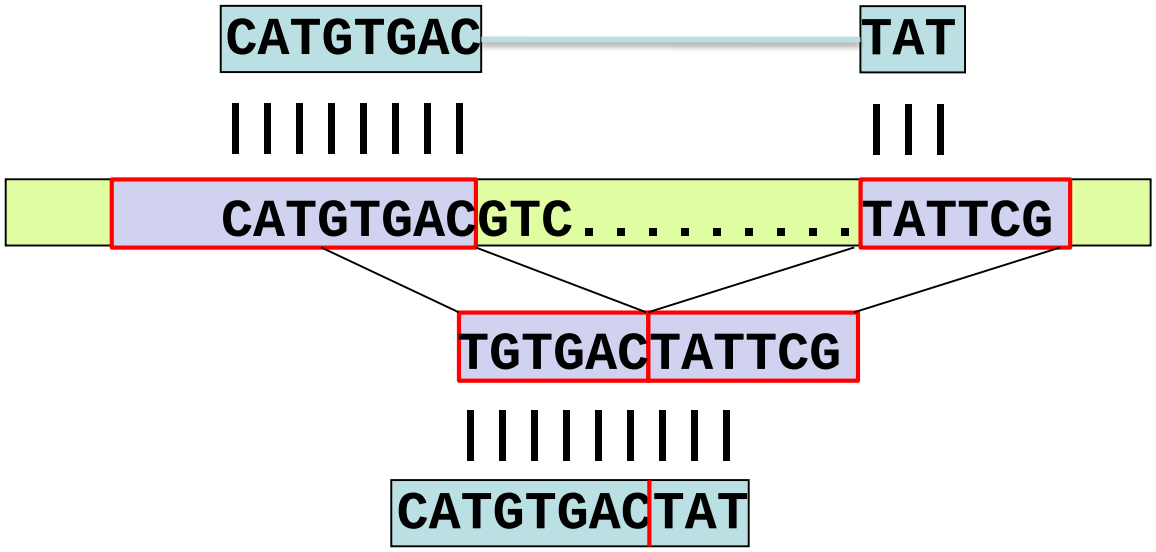
False alignment!

Soft-clip unmatched bases

Soft-clipping penalty < MM penalty

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

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

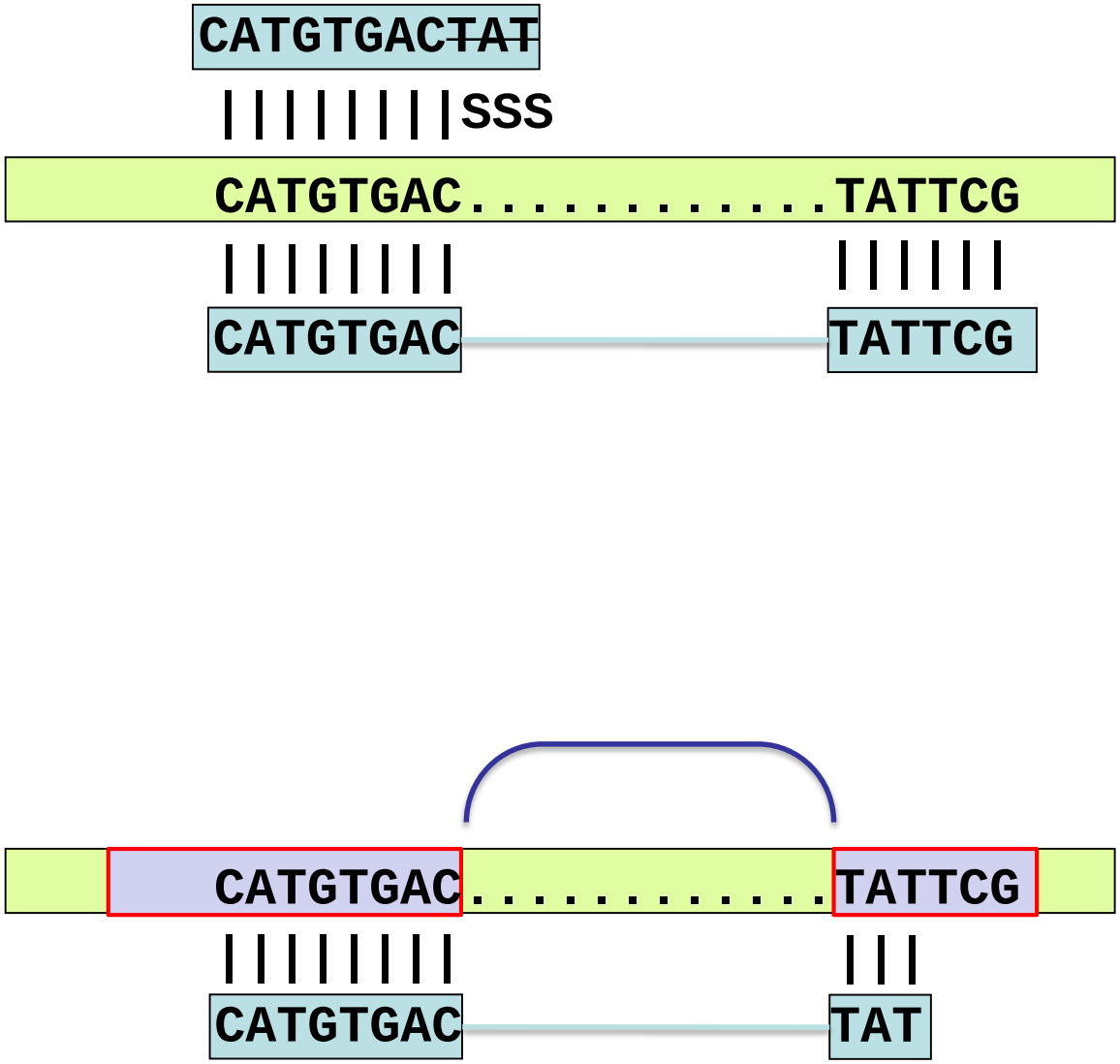
%	BLAT	Tophat	STAR	STAR + Annot
base FPR	2.9	5.4	2.0	0.1
base FPR: wrong loci	2.7	5.3	1.9	0.1
base FPR: wrong locus, missed splice	2.2	4.5	1.8	0.1
% of all reads mapping to processed pseudogenes				
all	0.7	1.6	0.8	0.1
False Positive, wrong locus, missed splice	81.8	82.3	71.1	26.0

~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



1st pass:
Reads with short overhangs are soft-clipped

Read with long overhangs identify novel junctions

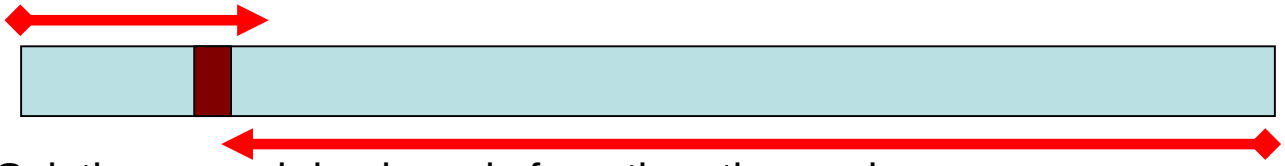
2nd pass:
Junctions from the 1st pass are added to the search space

Read with short overhangs map spliced to novel junctions

Increasing search sensitivity

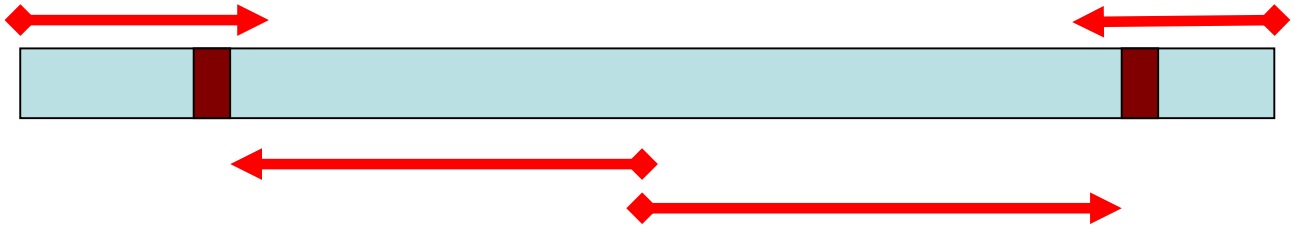
- One mismatch near one of the ends

  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>`**

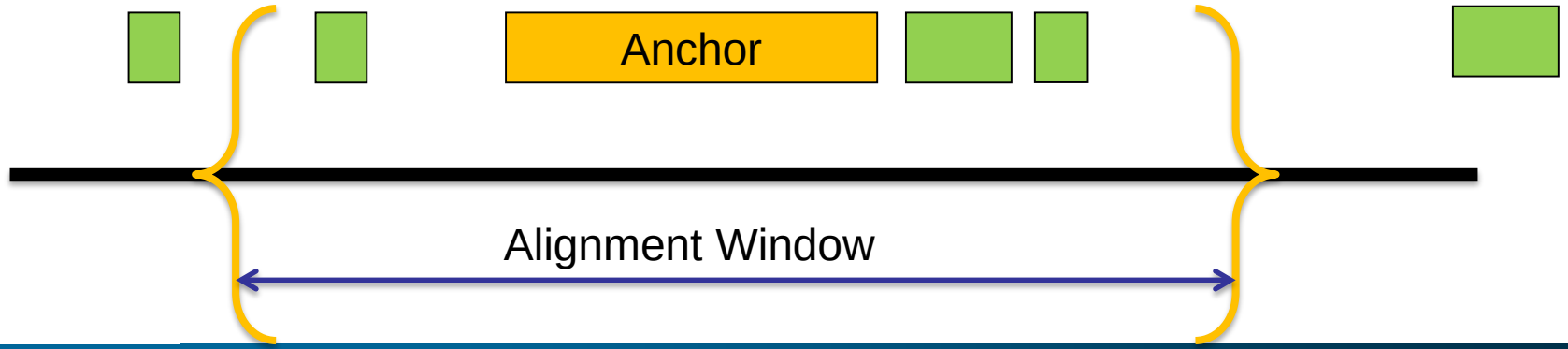
  “Anchors”: seeds that map <N times: =50 by default

- “Alignment windows”: genome regions around anchors

  All seeds inside alignment windows are stitched together

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

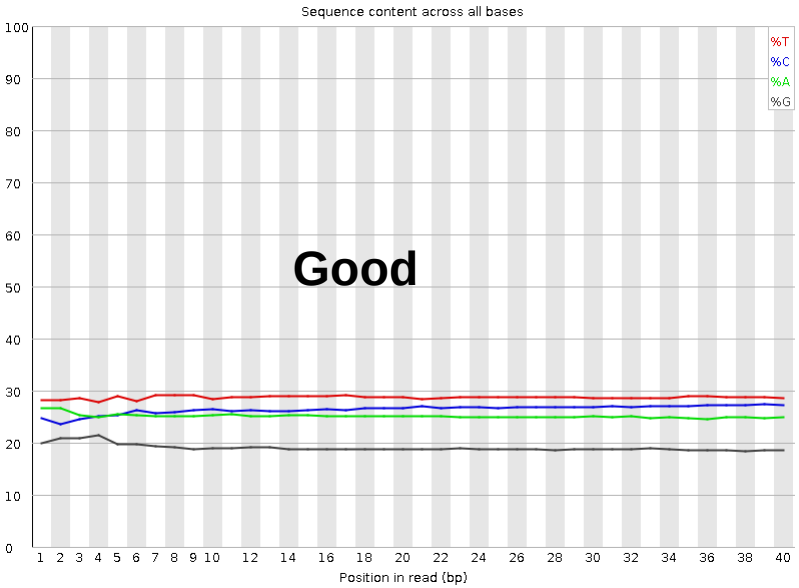
  `--alignIntronMax <N>, --alignMatesGapMax <N>`

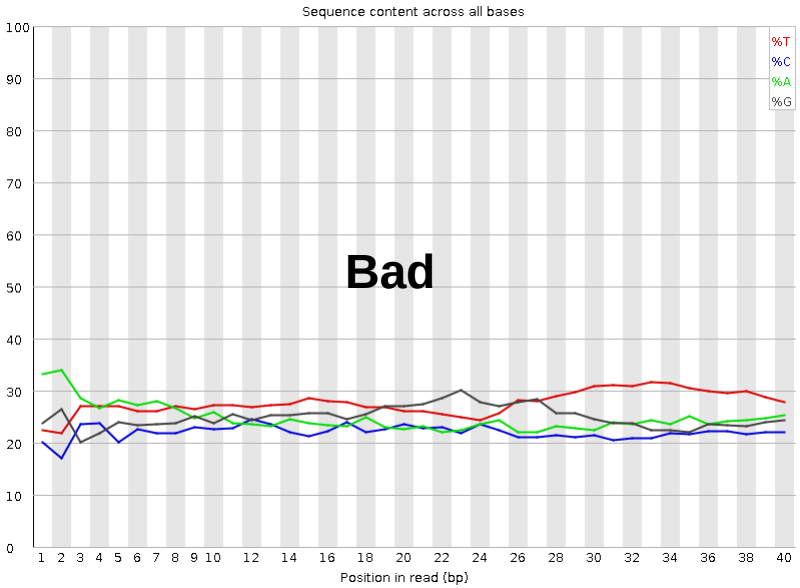


Pre-mapping QC with FASTQC

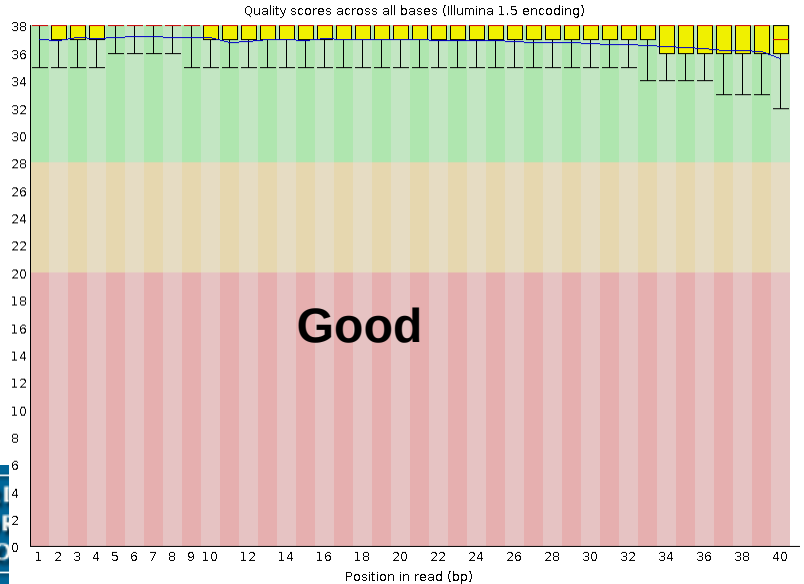
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

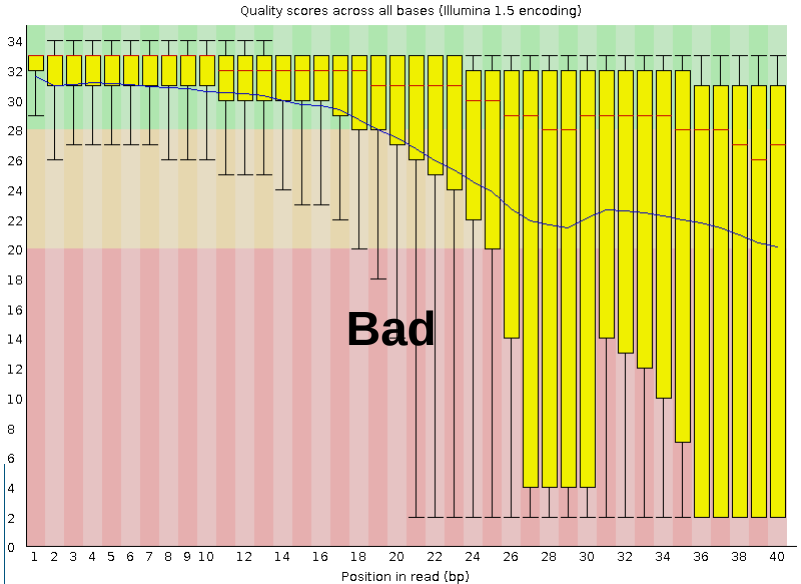
Per-base sequence content





Per-base sequence quality





Understanding mapping results

STAR's Log.final.out file:

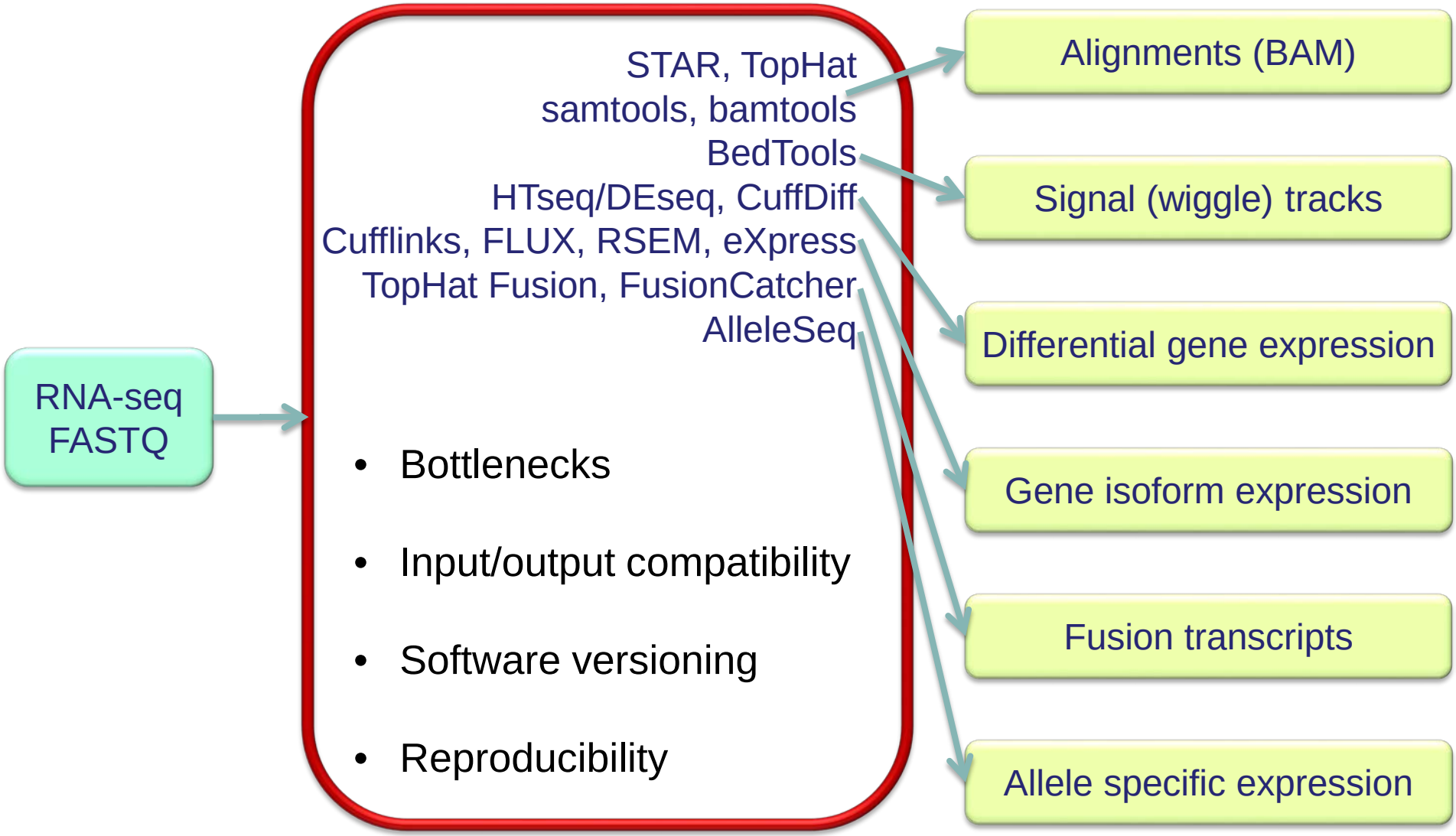
```
                          Average input read length |       202
                                       UNIQUE READS:
                         Uniquely mapped reads % |       90.08%
                           Average mapped length |       201.98
                        Mismatch rate per base, % |       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?

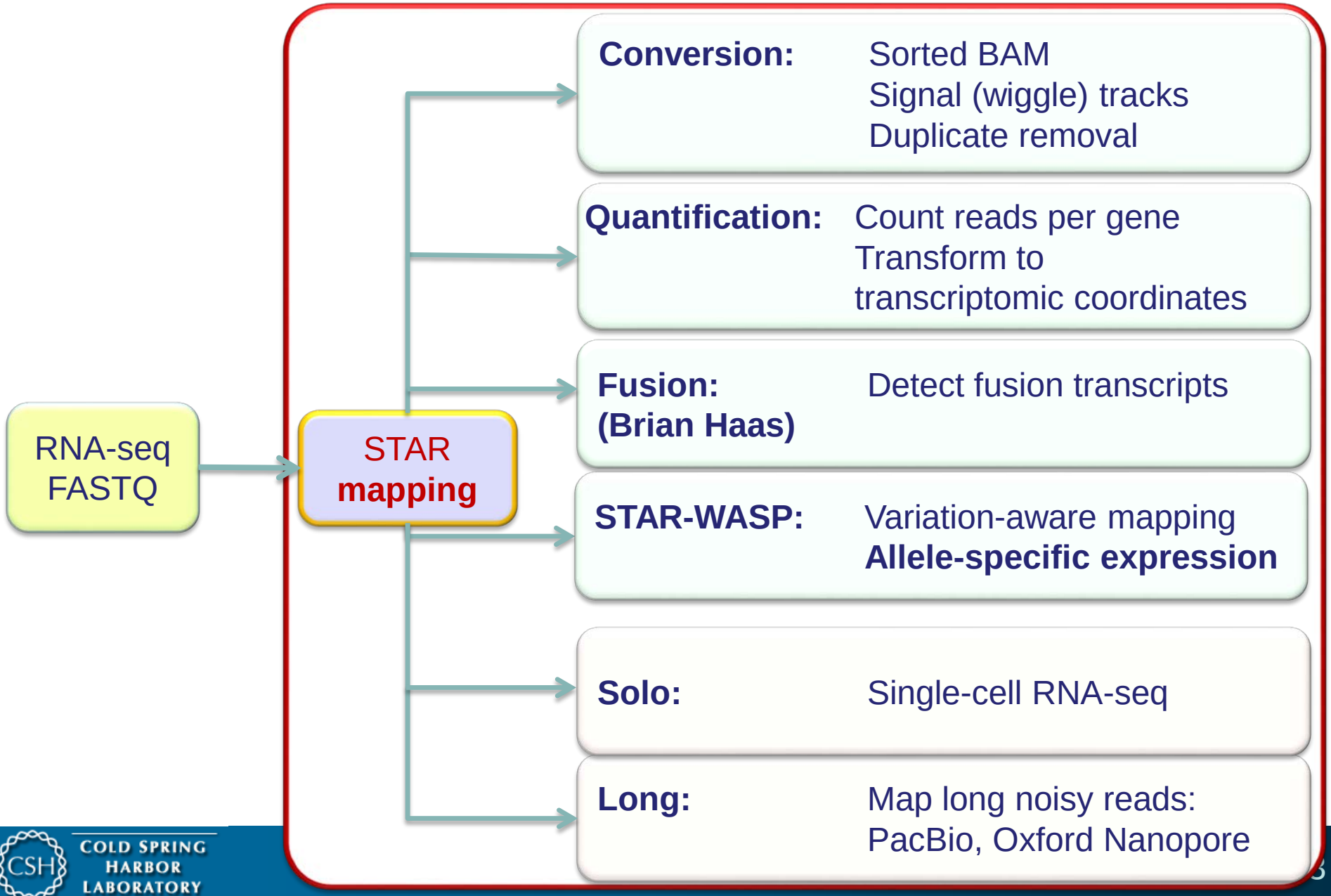
Possible problem	Checks/Solutions
File formatting mix-up:	Ensure the same order of read1/2
read1/read2 order broken	Map read1/2 separately
Poor quality of sequencing	Plot quality scores vs read length
Tails	
Poor quality	Trim by quality
Adapter - short insert	Trim adapter
rRNA insufficient depletion	Include rRNA sequences in the reference
Contamination with other species	BLAST unmapped reads

STARtools: post-mapping analyses at no extra cost

RNA-seq pipeline



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,

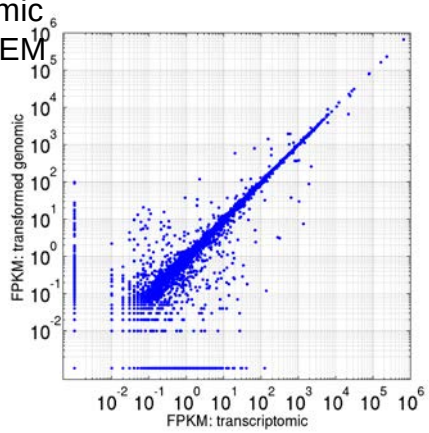
  and at the same time

  converts genomic alignments to transcriptomic

  no extra computational time required

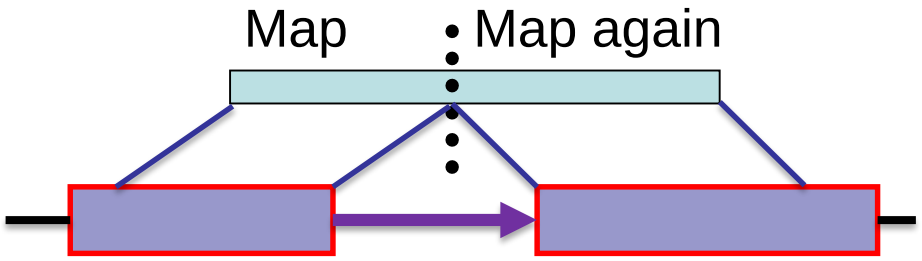
  `--quantMode TranscriptomeSAM`

Expression genomic => transcriptomic STAR / RSEM

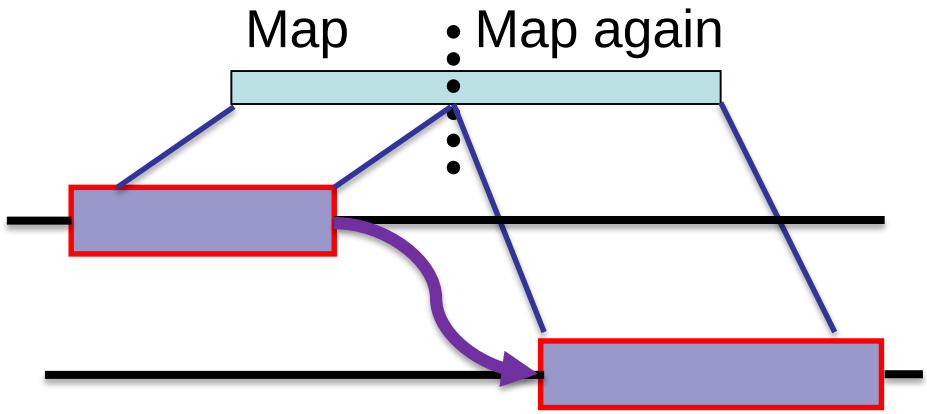


Expression using transcriptomic alignments Bowtie / RSEM

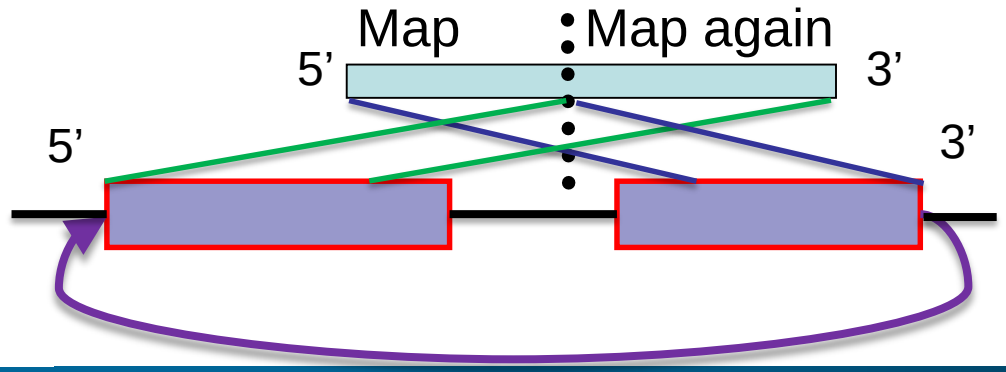
Chimeric and circular junctions



Linear junction

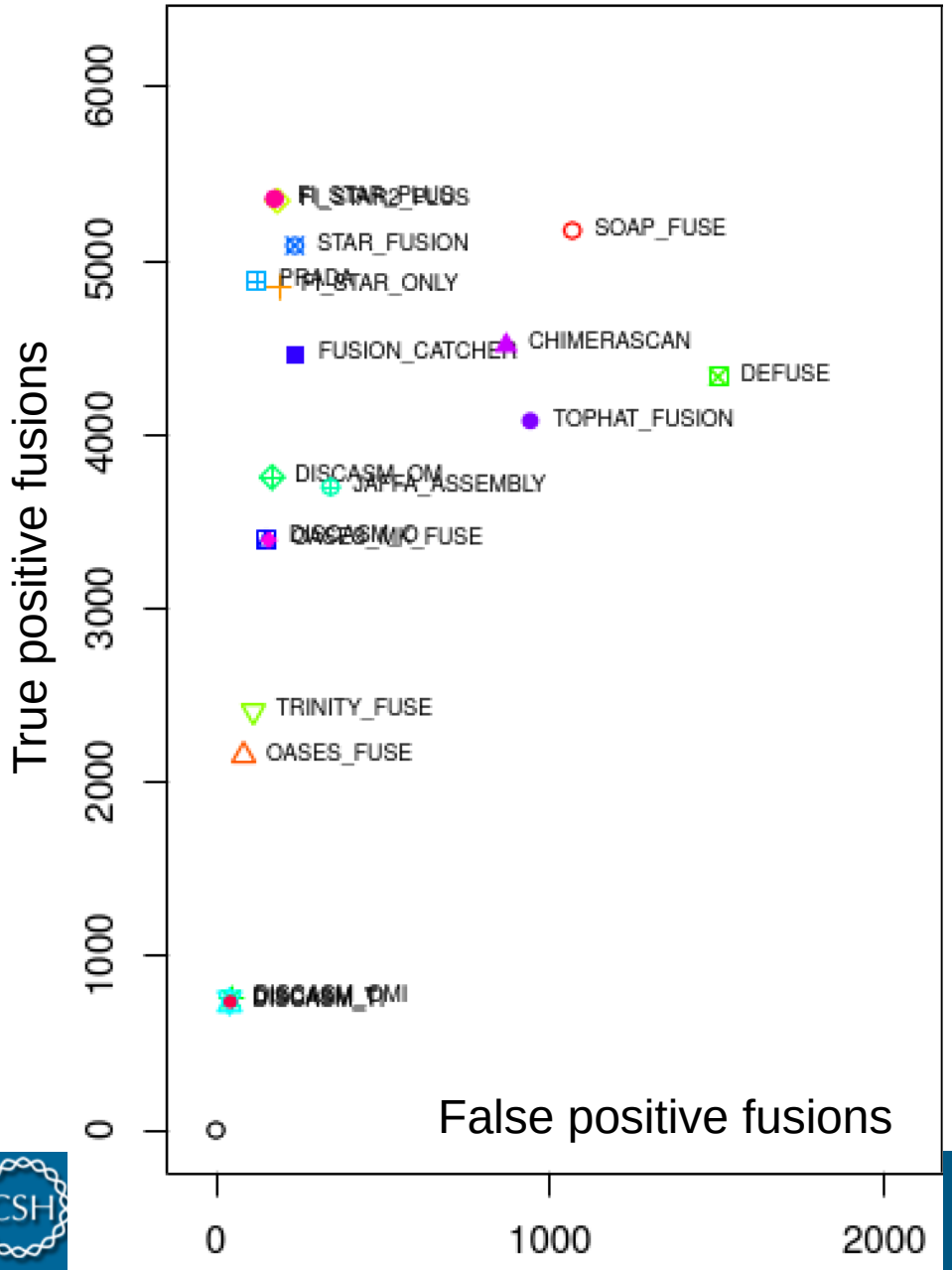


Chimeric junction



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