Sequence Alignment

HackBio

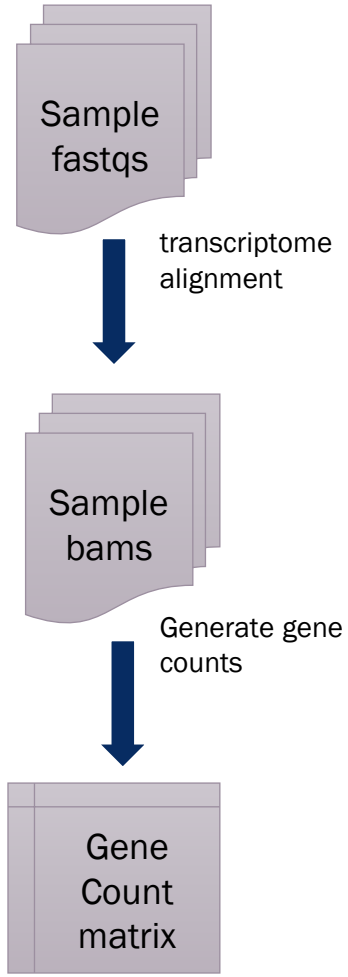
Melyssa Minto

West Lab, Duke Neurobiology

Computational Biology and Bioinformatics

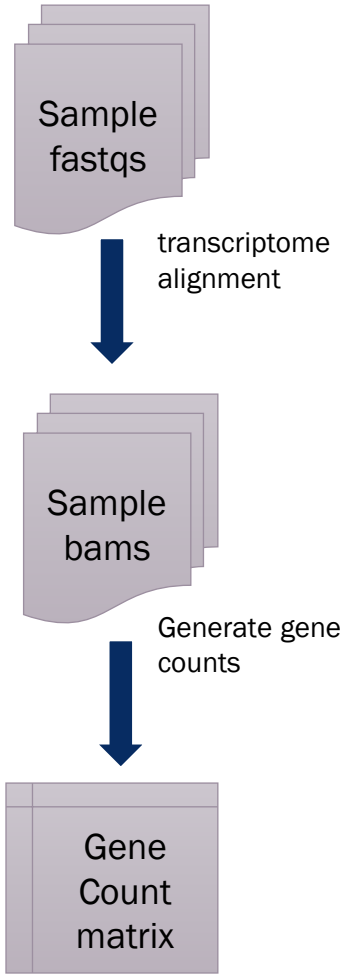
Transcriptomics pre-proccesing workflow

Preprocessing



Transcriptomics pre-proccesing workflow

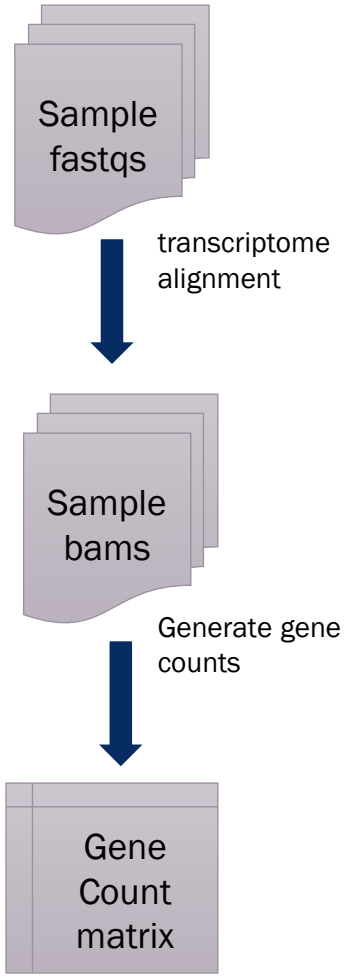
Preprocessing

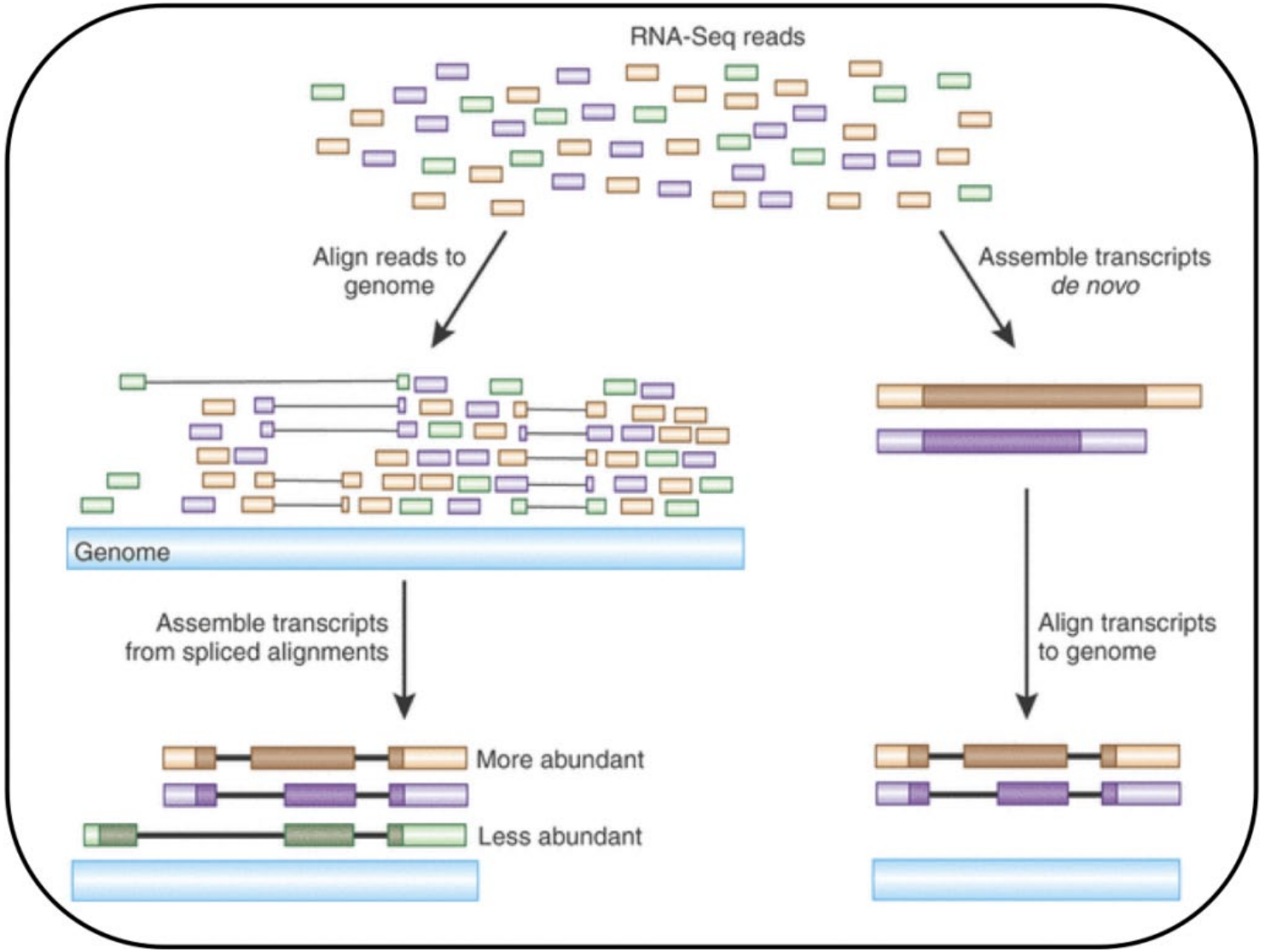



```
@A00257:355:HK7CTDRXX:1:2101:3522:1204 1:N:0:GACTACGA
CNCTTGAATGCTGAGATTACAGATGTGCTCATAGACAACAGTAGCCACATC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3577:1204 1:N:0:GACTACGA
CNGGGAGAACCAGGTTAAAATTGAAGGTAGAAAACACTATAAGATGGAGGA
+
F#FFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3703:1204 1:N:0:GACTACGA
CNTATCCATATAAGAATTCAACAGAGAAACGGCAGGAAGACCCTTACCACT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Transcriptomics pre-proccesing workflow

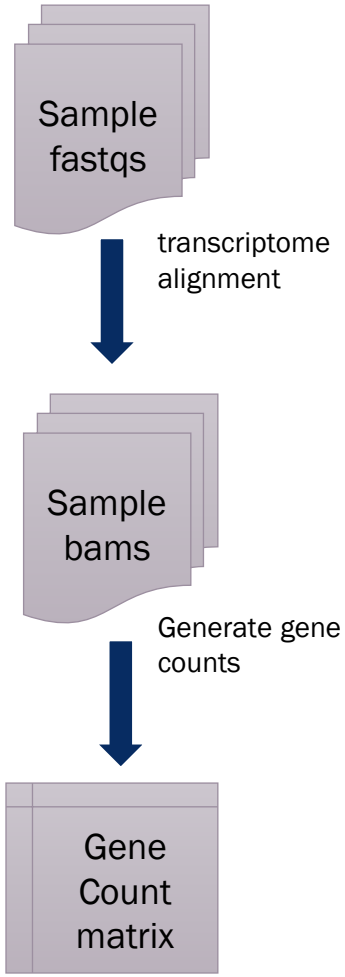
Preprocessing

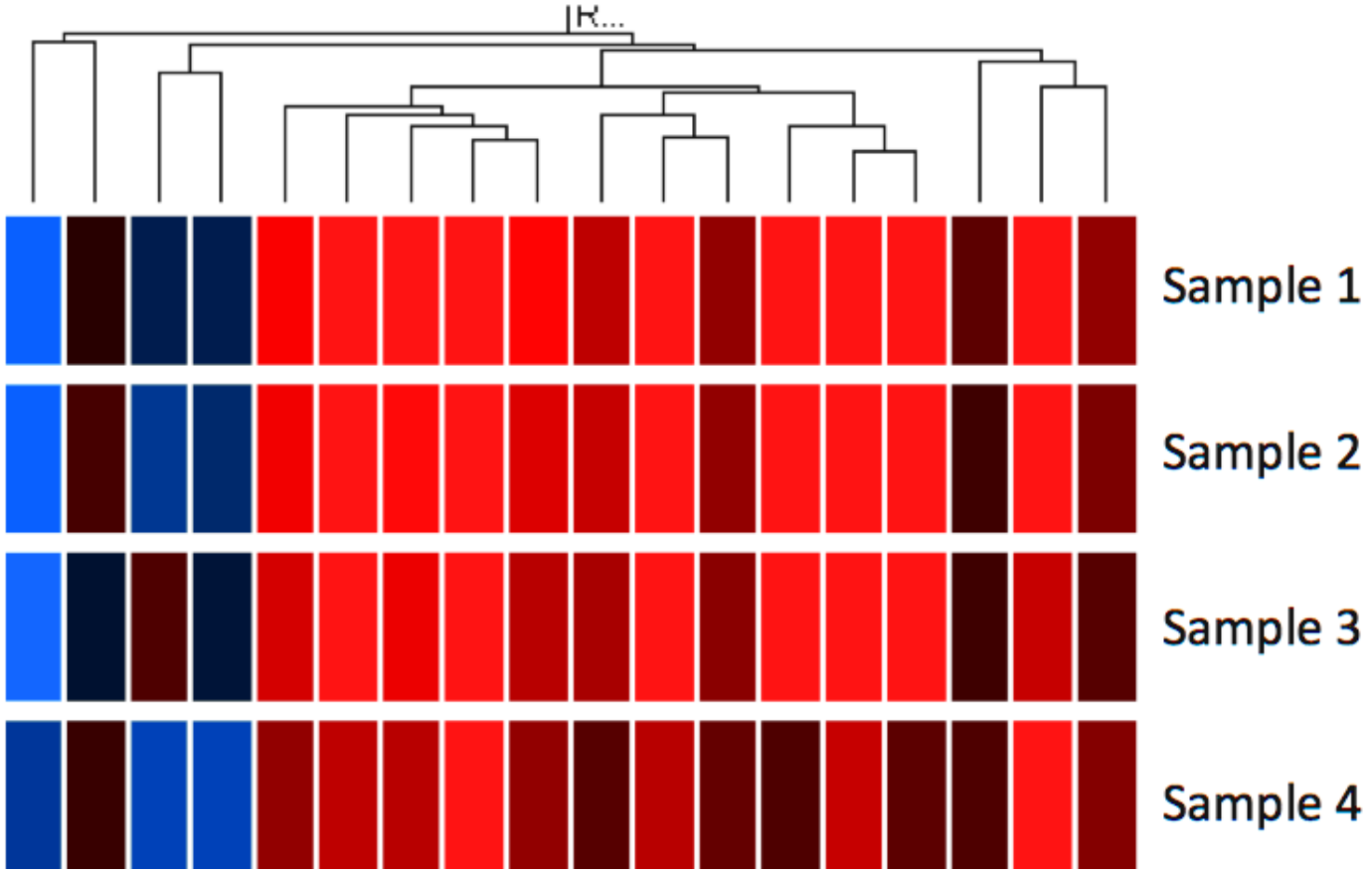




Transcriptomics pre-proccesing workflow

Preprocessing





Genomic Sequence Alignment

- Comparing two sequences and matching up by similarity

Genomic Sequence Alignment

- Comparing two sequences and matching up by similarity
  - Hamming Distance: # of nonmatching elements between sequences

ATTCGGATCTCA
ATTCGGCACTGA

Genomic Sequence Alignment

- Comparing two sequences and matching up by similarity
  - Hamming Distance: # of nonmatching elements between sequences

ATTCGGATCTCA
ATTCGGCACTGA

Hamming distance = 3

Genomic Sequence Alignment

- Comparing two sequences and matching up by similarity
  - Hamming Distance: # of nonmatching elements between sequences

ATTCGGATCTCA
ATTCGGCACTGA

  - Computationally inefficient for DNA alignment

Genomic Sequence Alignment

- Comparing two sequences and matching up by similarity
  - Hamming Distance: # of nonmatching elements between sequences

ATTCGGATCTCA
ATTCGGCACTGA

- Computationally inefficient for DNA alignment
- Also, DNA sequencing is *messy!*

ATTC-GATCTCA
ATTCGGAACT-A

The alignment problem

Input: DNA sequences

$$X = x_1 x_2 \ldots x_m$$

$$Y = y_1 y_2 \ldots y_n$$

The alignment problem

Input: DNA sequences $X = x_1 x_2 \ldots x_m$

$Y = y_1 y_2 \ldots y_n$

Output: optimal alignment of the sequences, while counting for *sequence variations* such as

- Frame shifts
- Substitution
- Inversion
- Transposition
- Duplication

The alignment problem

Input: DNA sequences $X = x_1 x_2 \ldots x_m$

$Y = y_1 y_2 \ldots y_n$

Output: optimal alignment of the sequences, while counting for *sequence variations* such as

- Frame shifts
- Substitution
- Inversion
- Transposition
- Duplication

$\begin{bmatrix} x_i \\ y_j \end{bmatrix}$ or $\begin{bmatrix} x_i \\ - \end{bmatrix}$ or $\begin{bmatrix} - \\ y_j \end{bmatrix}$

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

ATCGGCT
ATCGGCT

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

```
ATCGGCT
ATCGGCT
--CGGCTCGA
```

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

```
ATCGGCT
ATCGGCT
  --CGGCTCGA
TT--ATCGGGT
```

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

```
      ATCGGCT
      ATCGGCT
       --CGGCTCGA
TT--ATCGGGT
      ATCCCCT
```

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

```
      ATCGGCT
      ATCGGCT
       --CGGCTCGA
TT--ATCGGGT
      ATCCCCT
       AT-GGC-NTCCNN
```

*We need an algorithm that can align sequences while **penalizing** sequence variations to get the optimal alignment*

```
      ATCGGCT
      ATCGGCT
       --CGGCTCGA
TT--ATCGGGT
      ATCCCCT
       AT-GGC-NTCCNN
```

Different alignment tools

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

OSI certified

Burrows-Wheeler Aligner

Introduction

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

Sequence analysis

Advance Access publication October 25, 2012

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

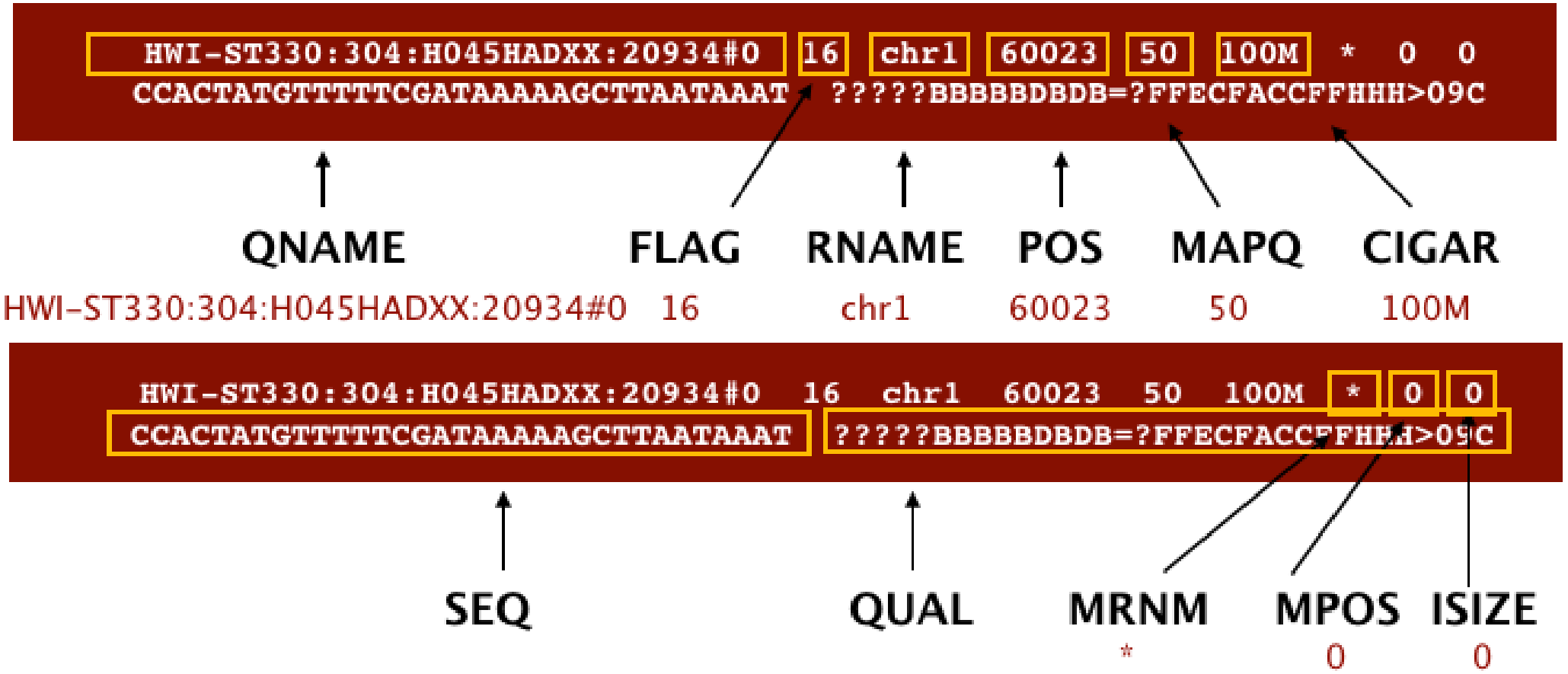
Associate Editor: Inanc Birol

Alignment file types

- .sam and .bam files are the typical alignment files
- They include
  - Header
  - Alignments & Alignment Info

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Assessing alignment quality



Aligning with STAR

1.Configure reference genome

2. Align!

```
# genome generate
STAR --runThreadN 6 \
--runMode genomeGenerate \
--genomeDir /home/transcriptomics/STAR_mm10/\
--genomeFastaFiles /home/transcriptomics/mm10.fa \
--sjdbGTFfile /home/transcriptomics/mm10.gtf

# algin
STAR --genomeDir /home/transcriptomics/STAR_mm10/ \
--runThreadN 6 \
--readFilesIn sample.fq \
--outFileNamePrefix sample \
--outSAMtype BAM SortedByCoordinate \
--outSAMunmapped Within \
--outSAMattributes Standard
```