

Alignment

Day 4 ECU-UNC Bioinformatics Course
2020

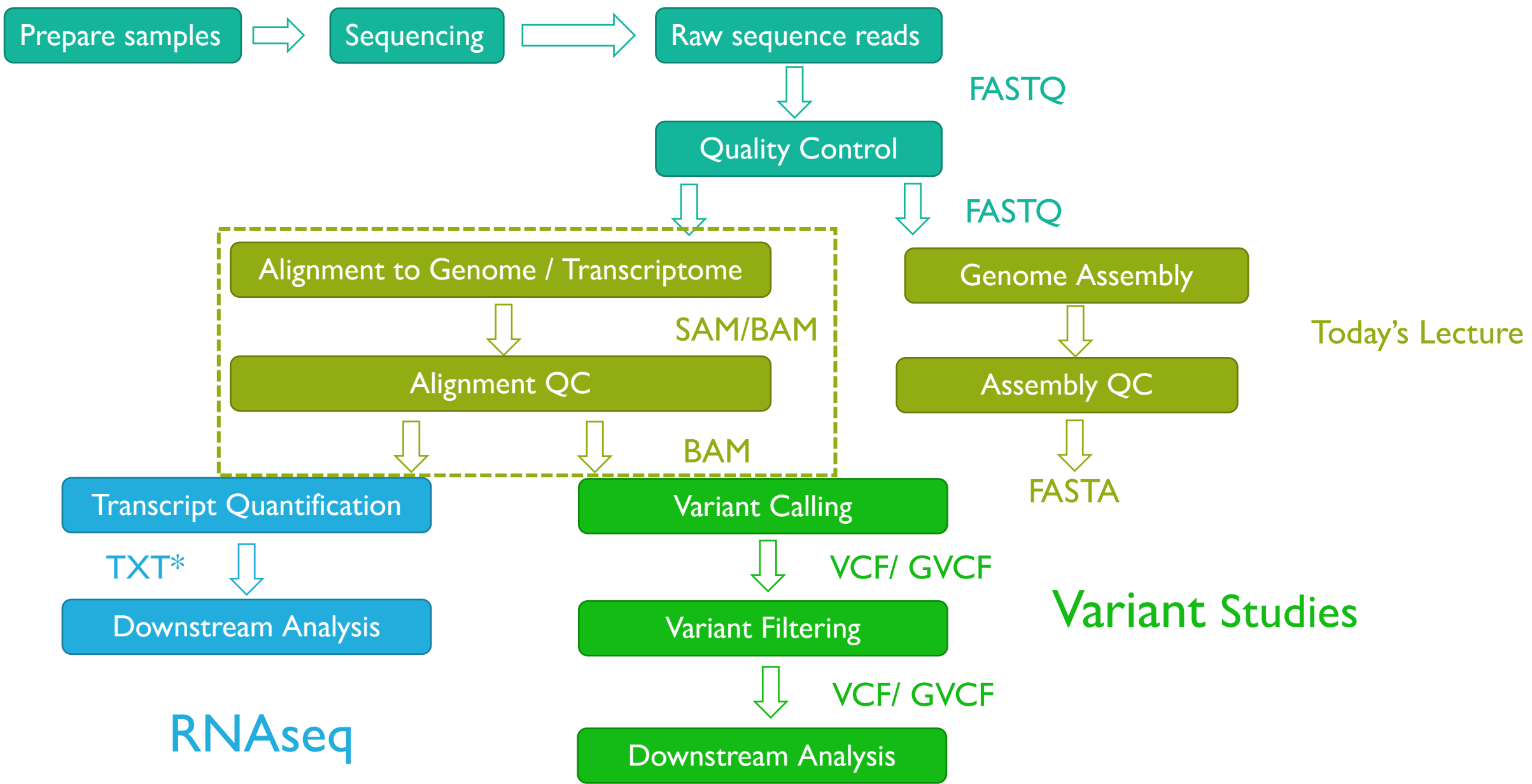
OUTLINE

10am Lecture

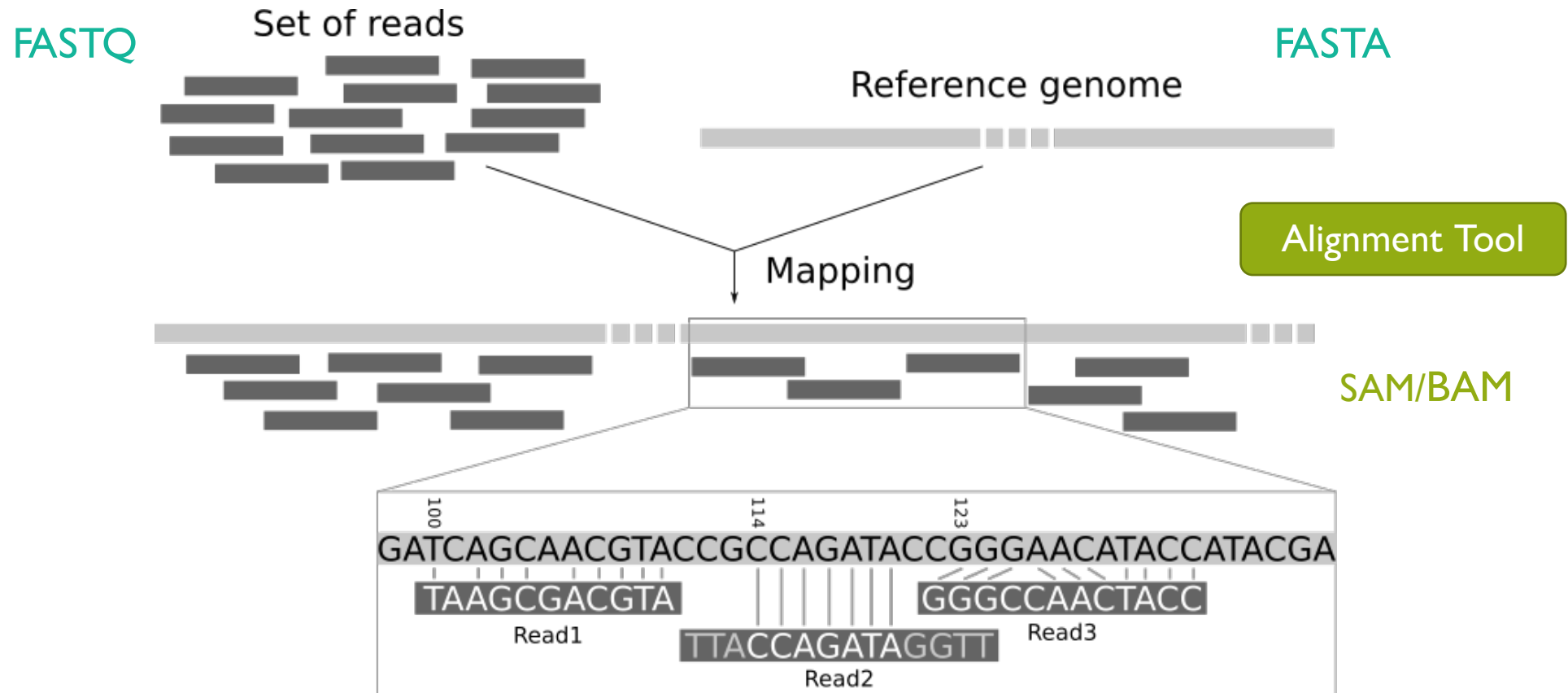
- Alignment strategies
- File formats
- QC

11am Tutorial

- Short-read mapping with BWA
- Splice-aware mapping with STAR



SEQUENCE ALIGNMENT



GENERAL CHALLENGES

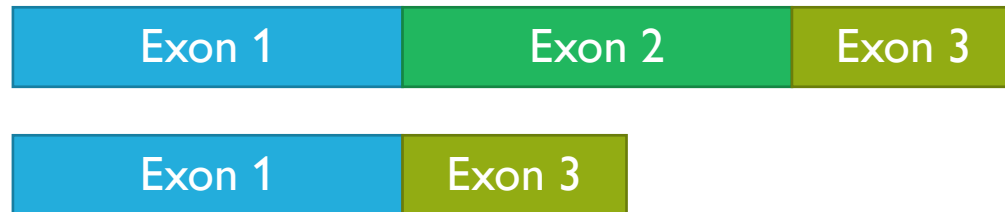
- Accommodating indels
- Distinguishing sequencing error from real variation
- Short reads (50-150bp) can map to multiple parts of the genome
- Millions of reads
- Large reference genome (huge search space)
- Potentially incomplete or low quality reference genome
- Repetitive sequences

RNASEQ CHALLENGES

Gene model in a reference genome



Transcript isoforms



A BIT ABOUT REFERENCES

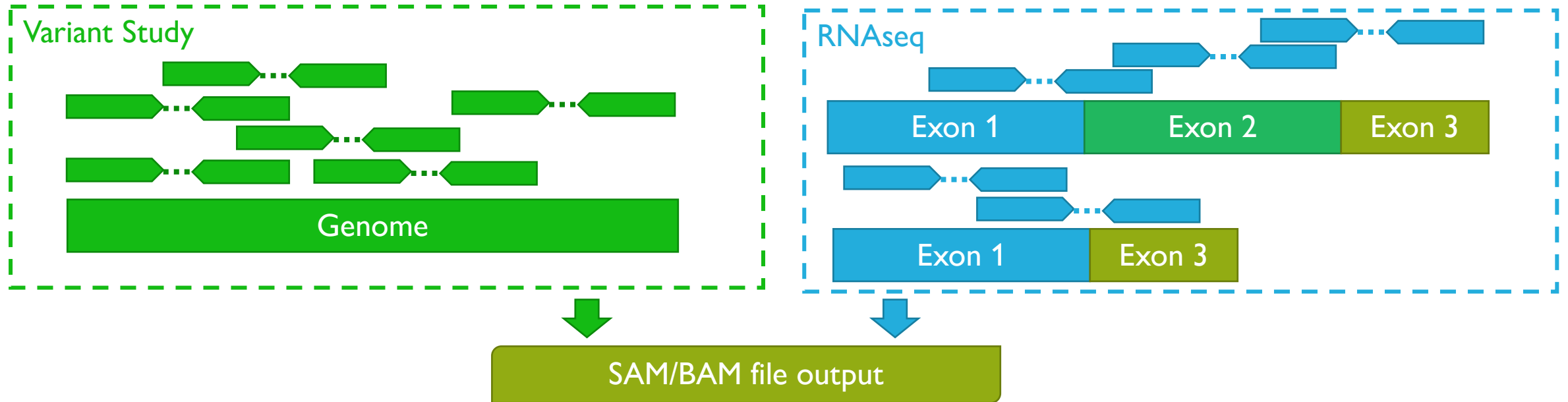
- Model with versions
- Annotated vs Not (check your version!)

Annotation (.gtf/.gff)	Exon 1	Exon 2	Exon 3
Reference (.fna)	ATGCAATTACGAATCAAGAAATTACCGACCTAATTGGAATCCTAACGATGAGACTATT		

- Annotations different on different databases (Chris' Lecture)
- Transcriptome = fasta with all known transcripts

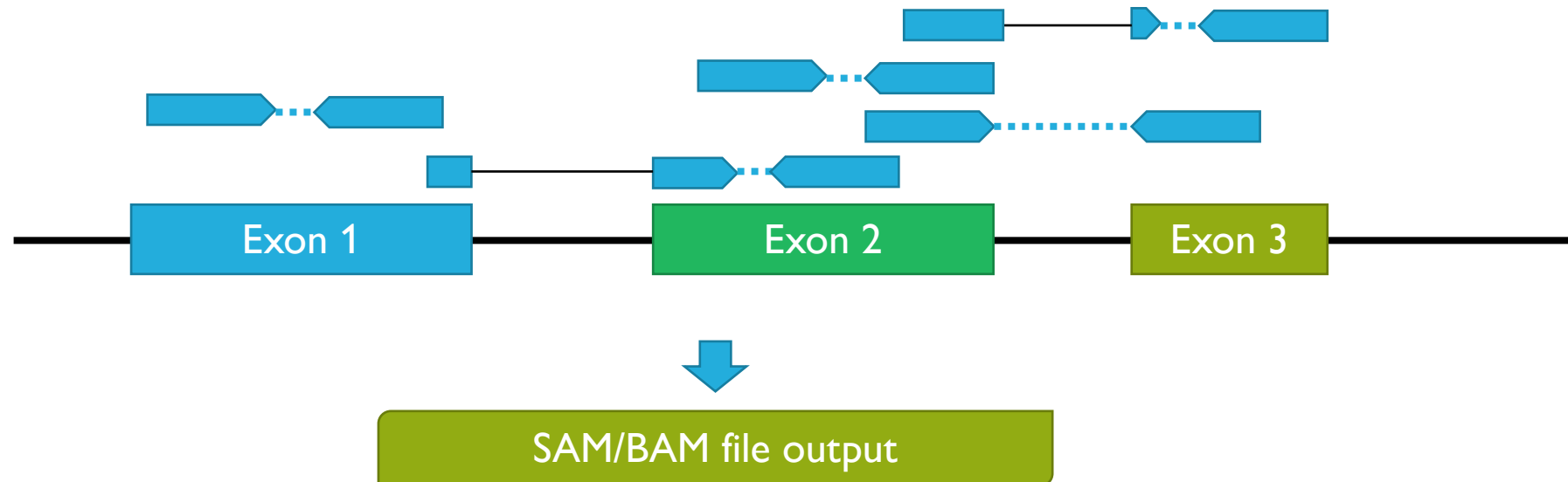
SHORT-READ ALIGNMENT

- Includes a gap penalty
- Standard alignment software: **Bowtie2** or **bwa** (also **stampy** and lots of others)



SPLICE-AWARE ALIGNMENT

- For RNAseq
- Align to genome (and sometimes annotation)
- Reads can be split up to accommodate very long introns & other splice junctions
- Standards are **HISAT2** and **STAR**



ALIGNMENT STEP 1: INDEX

- Index your reference prior to commencing
- Index allows alignment programs to efficiently search
- Ways of breaking the reference into bits

	Hash-based	Suffix-Arrays	Burrows-Wheeler (FM index)
Programs	BLAST, (Kallisto, Salmon)	STAR, Salmon	BWA, Bowtie2
Method	K-mer based	Sorted table of suffixes of a string	Sorted rearrangements of string
	Can be slow	Large memory requirements to index	Lower memory req. but reduced search efficiency

BWA

- `bwa index <INFILE.FA>`
- creates FM-index
- creates a bunch of files, .pac, .sa, .bwt., .ann .amb (burrows wheeler indices + suffix arrays)

Burrows Wheeler Transform

T: BANANAS → BWT: ANNB\$AA

	F		L
1	\$	BANAN	A
2	A	\$BANA	N
3	A	NA\$BA	N
4	A	ANANA\$	B
5	B	ANANA	\$
6	N	A\$BAN	A
7	N	ANASB	A

A	N	A
3	2	1

QUERY:
ANA

1. $[C[A] + 1, C[A + 1]] = [start, end] = [2, 4]$
2. $[C[N] + Occ(N, start - 1) + 1, C[N] + Occ(N, end)] = [6, 7]$
3. $[C[A] + Occ(A, start - 1) + 1, C[A] + Occ(A, end)] = [3, 4]$

[FM-Index Data Structure]

$C[c]$ of “ANNB\$AA”

c	S	A	B	N
$C[c]$	0	1	4	5

$Occ(c, k)$ of “ANNB\$AA”

	A	N	N	B	S	A	A
$\begin{smallmatrix} k \\ c \end{smallmatrix}$	1	2	3	4	5	6	7
S	0	0	0	0	1	1	1
A	1	1	1	1	1	2	3
B	0	0	0	1	1	1	1
N	0	1	2	2	2	2	2

STEP 2: ALIGN READS TO GENOME

- “bwa mem” (for short read applications), “bwa swa” for >200bp (Smith-Waterman)
- Finds matches to query in index.
- Assigns them a MQ (map quality) score based on PHRED base score + mismatches + gaps



Chr10:1020 AGTACCAGAAGTATCTCGACTCTAAGGAATTGAGTCA

|||||
AGTA^ACTCGACT^TTA

Mismatches 1

MQ=10

Chr2:20231 TCATATGGTAGTAACTGCACGTTAAGTAAATTAGCAAT

|||||
AGTA^ACTCGACT^TTA

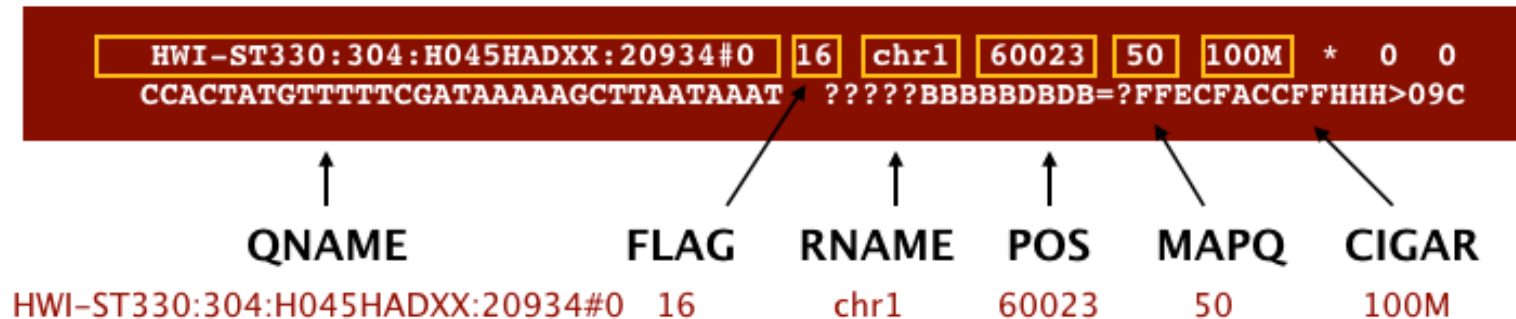
Mismatches 2

MQ=1



SAM FILE FORMAT

- **Flags:** alignment information (pairs mapped? Secondary alignment? Or both?)
<https://broadinstitute.github.io/picard/explain-flags.html>
- **MAPQ:** Reflects $-10\log_{10}(\text{probability the mapping is wrong})$
 - Different aligners score a bit differently – don't compare MAPQ between aligners.
 - Perfect mapping in bwa mem=60, perfect mapping in bowtie2=42
 - Unique mapping in STAR is 255
 - 0 is ambiguous mapping – equal quality in multiple locations.





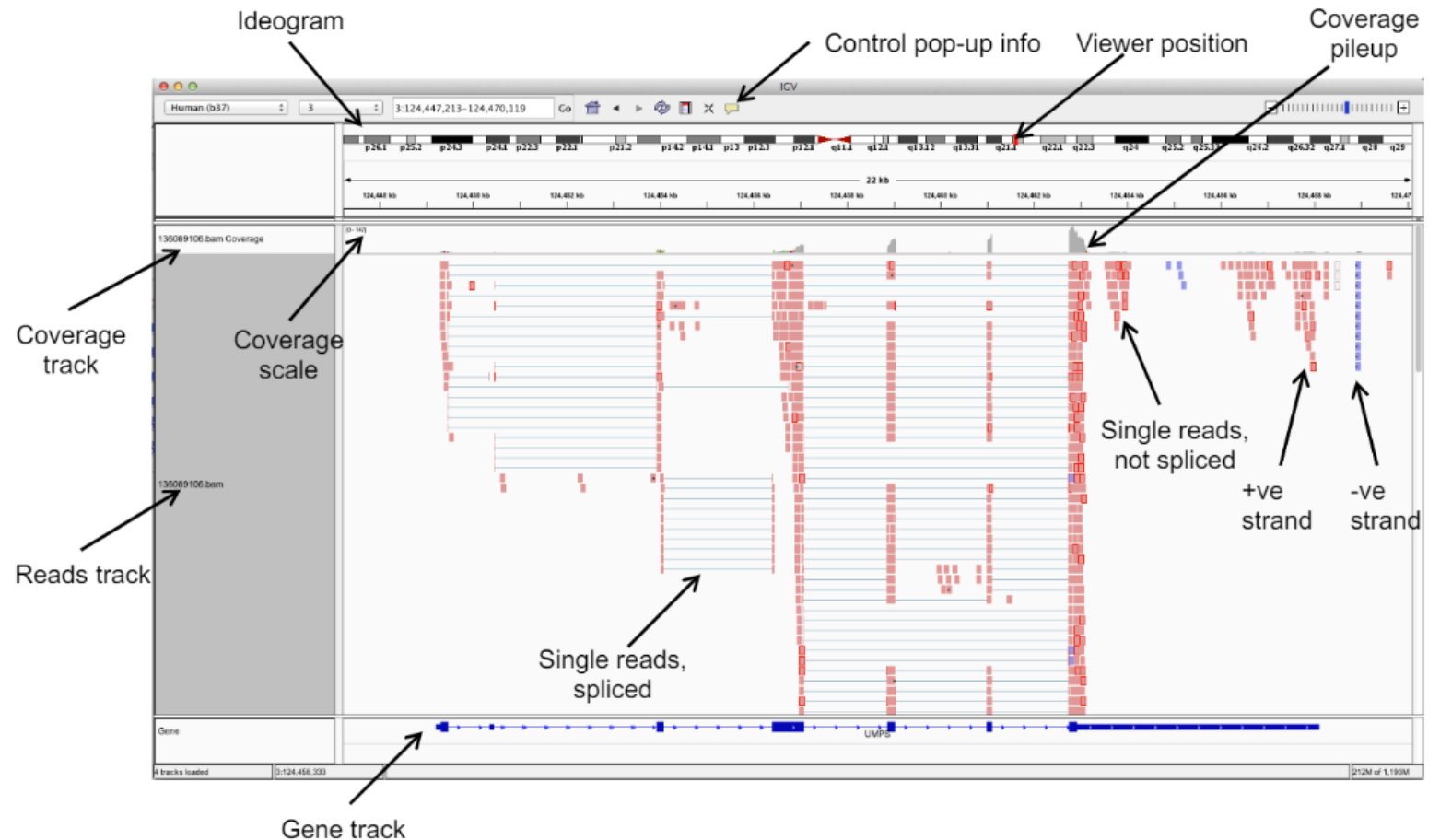
SAM FILE FORMAT

- 1 SAM/BAM per FASTQ
- SAM format is the FASTQ information plus additional information about position of read and its mapping quality
- BAM is a binary compressed version of a SAM file (much smaller file size!)
- **samtools** software can convert SAM to BAM
- **STAR** can give you BAM directly

CHECK IT OUT



Integrative
Genomics
Viewer



STEP 3: SORT AND INDEX BAM

- Convert sam to bam

```
samtools view -b -S -o <OUTFILE.bam> <INFILE.sam>
```

- Sort sam file

```
samtools sort -o <OUTFILE.bam.sorted> <INFILE.bam>
```

- Index bam file (creates a .bai)

```
samtools index <INFILE.bam.sorted>
```

- If PCR step in library prep – Mark PCR duplicates (molecules with same UMI OR exact length and sequence)

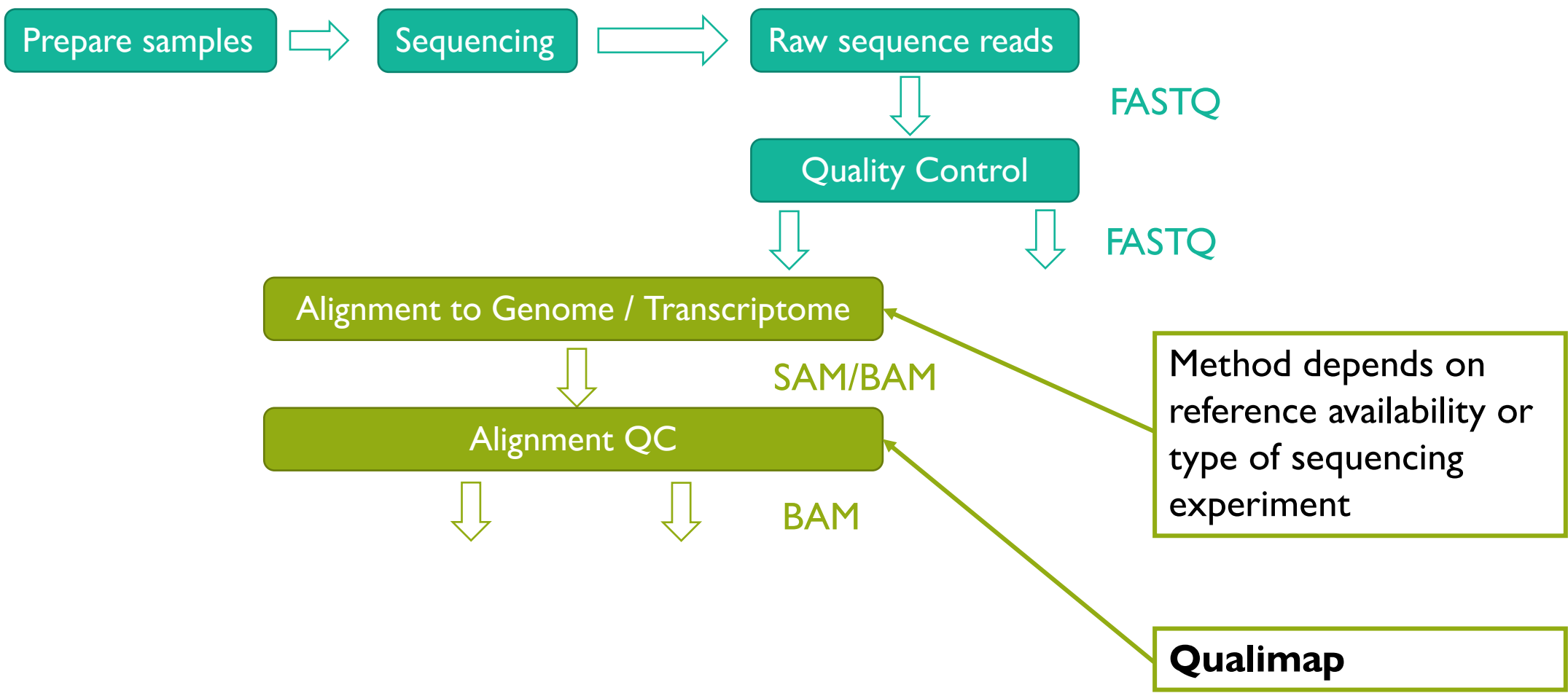
```
java -jar picard.jar MarkDuplicates I=input.bam O=marked_duplicates.bam, M=marked_dup_metrics.txt
```

ALIGNMENT QUALITY

- **Qualimap** program
 - Can assess short-read alignments “`qualmap bamqc`” and RNAseq alignments “`qualimap rnaseq`”
- Should have similar mapping rates across all samples – may need to remove a sample
- If mapping to a heterospecific reference can use to adaptively explore good mapping parameters to increase number of mapped reads or correct for other biases.
- Repetitive regions = way more mapped reads (not RNAseq)

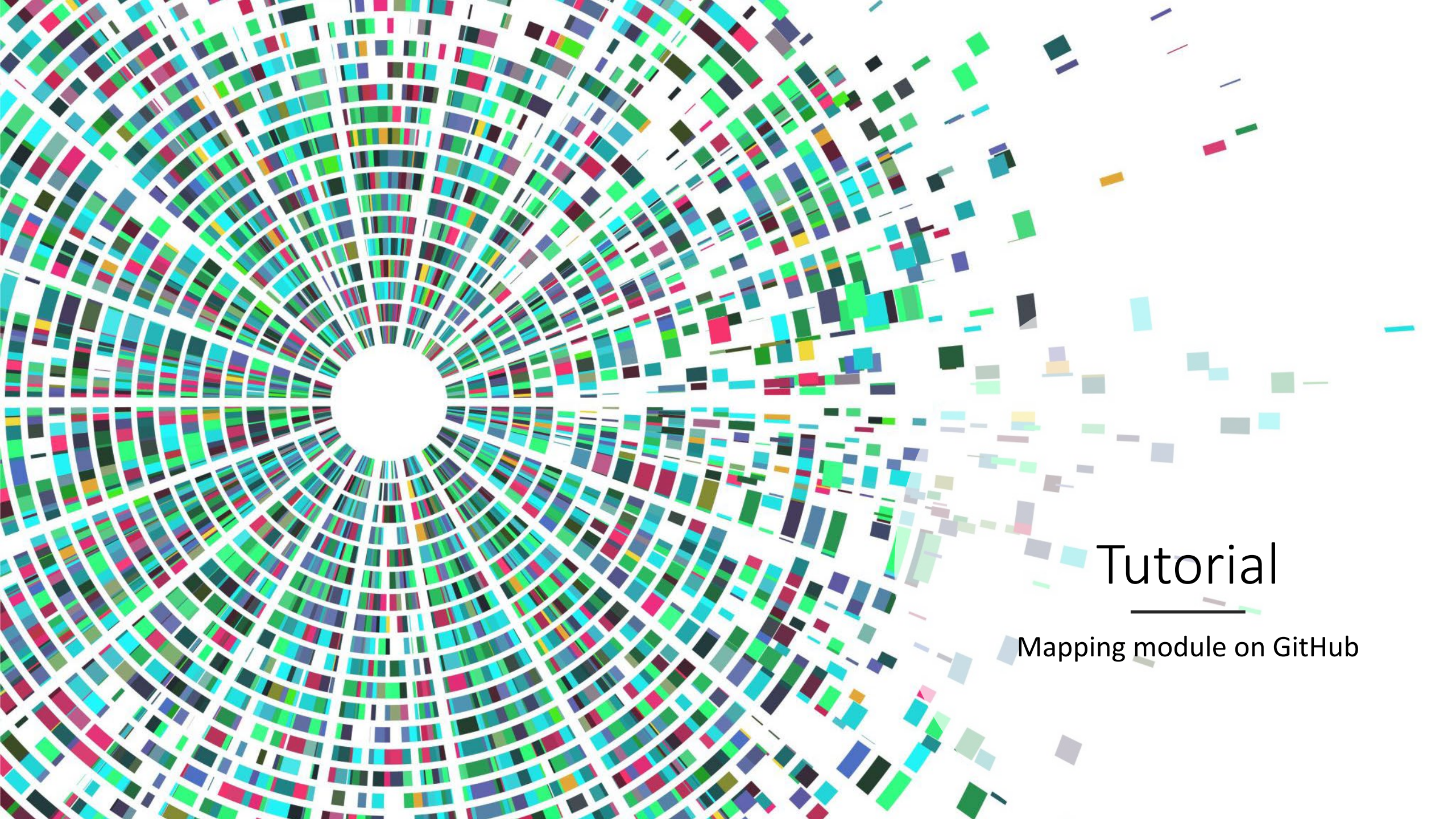
For RNAseq

- Good quality mapped to conspecific reference > 60% uniquely mapped reads (will vary depending on species and genome quality)!
- Want lots of reads mapping to exons – if not might be DNA contamination or new transcripts



FURTHER READING & REFERENCES

- Li and Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25: 1754-1760
- Sheng et al (2017) Multi-perspective quality control of Illumina RNA sequencing data analysis. *Briefings in Functional Genomics*
- McGill iGEM (2020) Introduction to Burrows-Wheeler Alignment and Samtools for Cancer Mutation Calling Bioinformatics. https://www.youtube.com/watch?v=P_YKQKFI4Lk
- Griffith Lab RNA-seq Bioinformatics Course Lecture (<https://rnabio.org/course/>)
- HBC training Tutorials
 - https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lectures/alignment_quantification.pdf
 - https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/03_QC_STAR_and_Qualimap_run.html
- <http://www.acgt.me/blog/2014/12/16/understanding-mapq-scores-in-sam-files-does-37-42>
- <https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/>
- <https://samtools.github.io/hts-specs/SAMv1.pdf>



Tutorial

Mapping module on GitHub