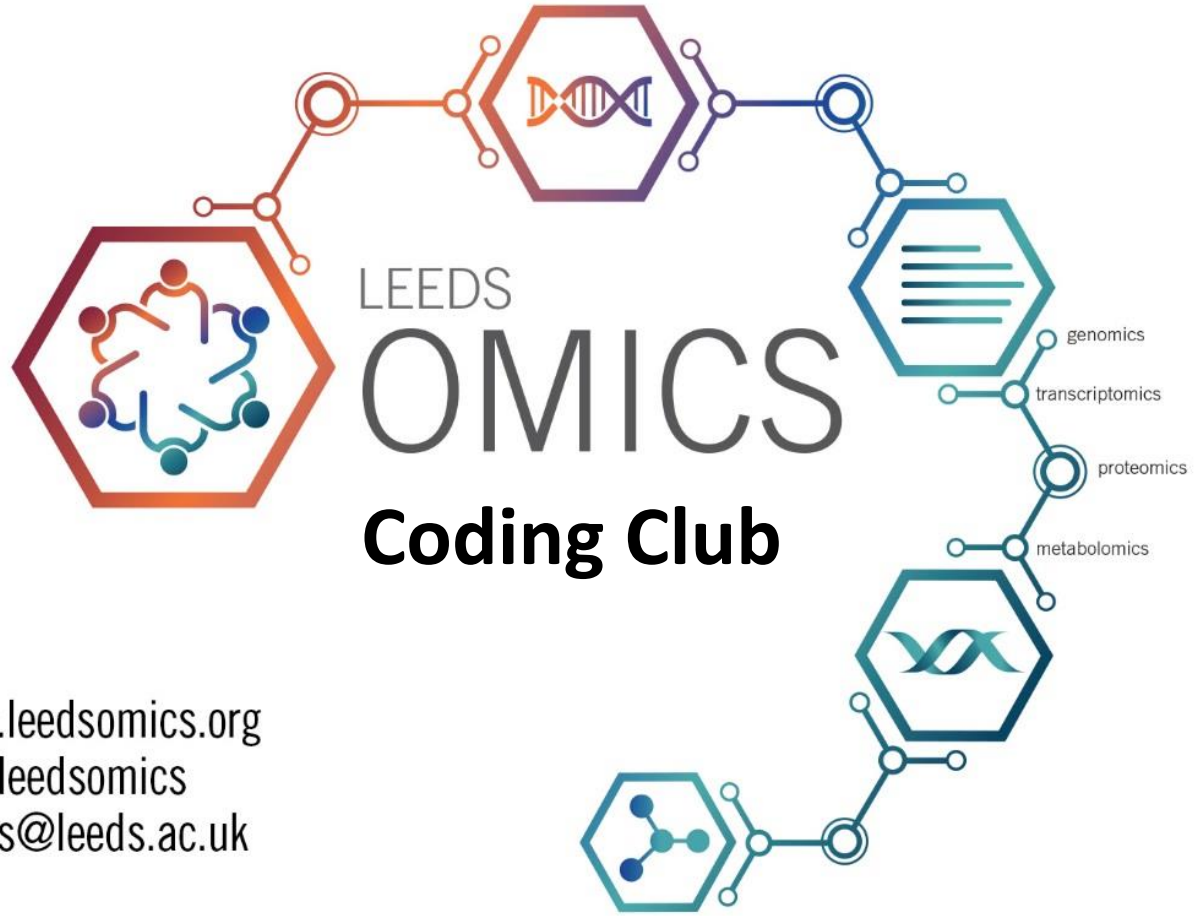


Coding Club

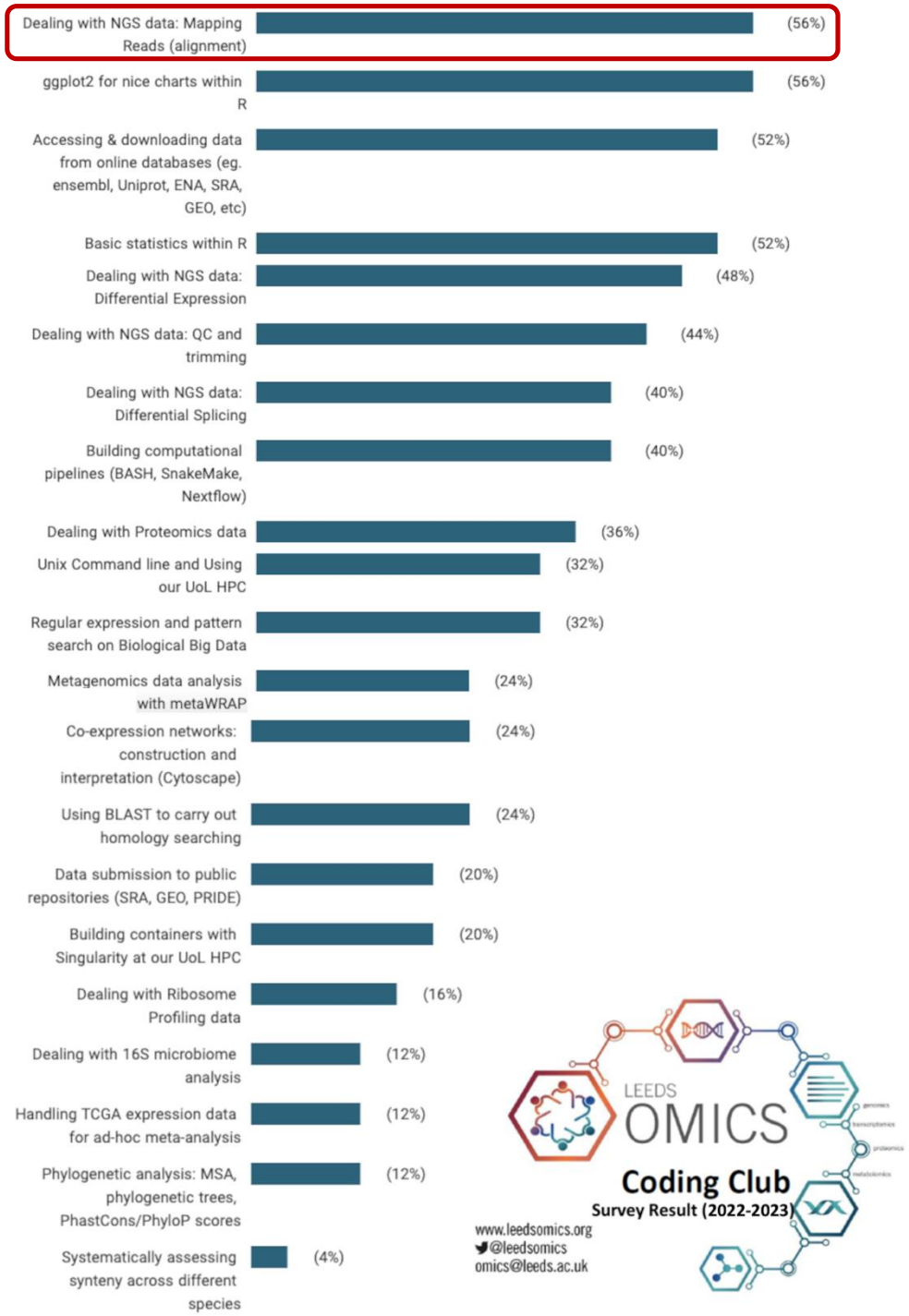
www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Dealing with NGS data:
Aligning/Mapping reads

Club Moderators: Elton Vasconcelos and Euan McDonell

Topics to be addressed on the **2022-23** season - **Survey Result**

1st session



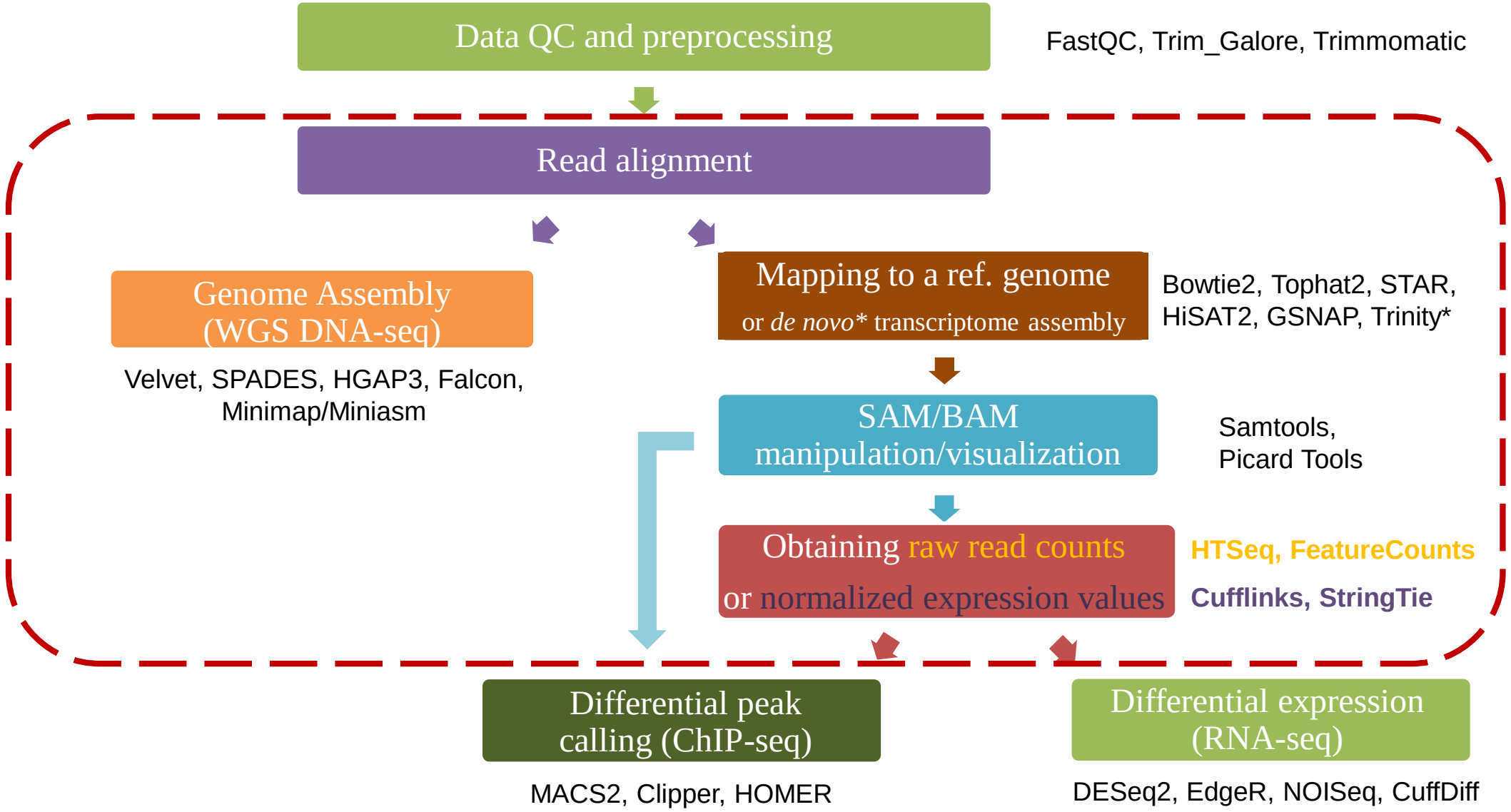
LEEDS OMICS

Coding Club

Survey Result (2022-2023)

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

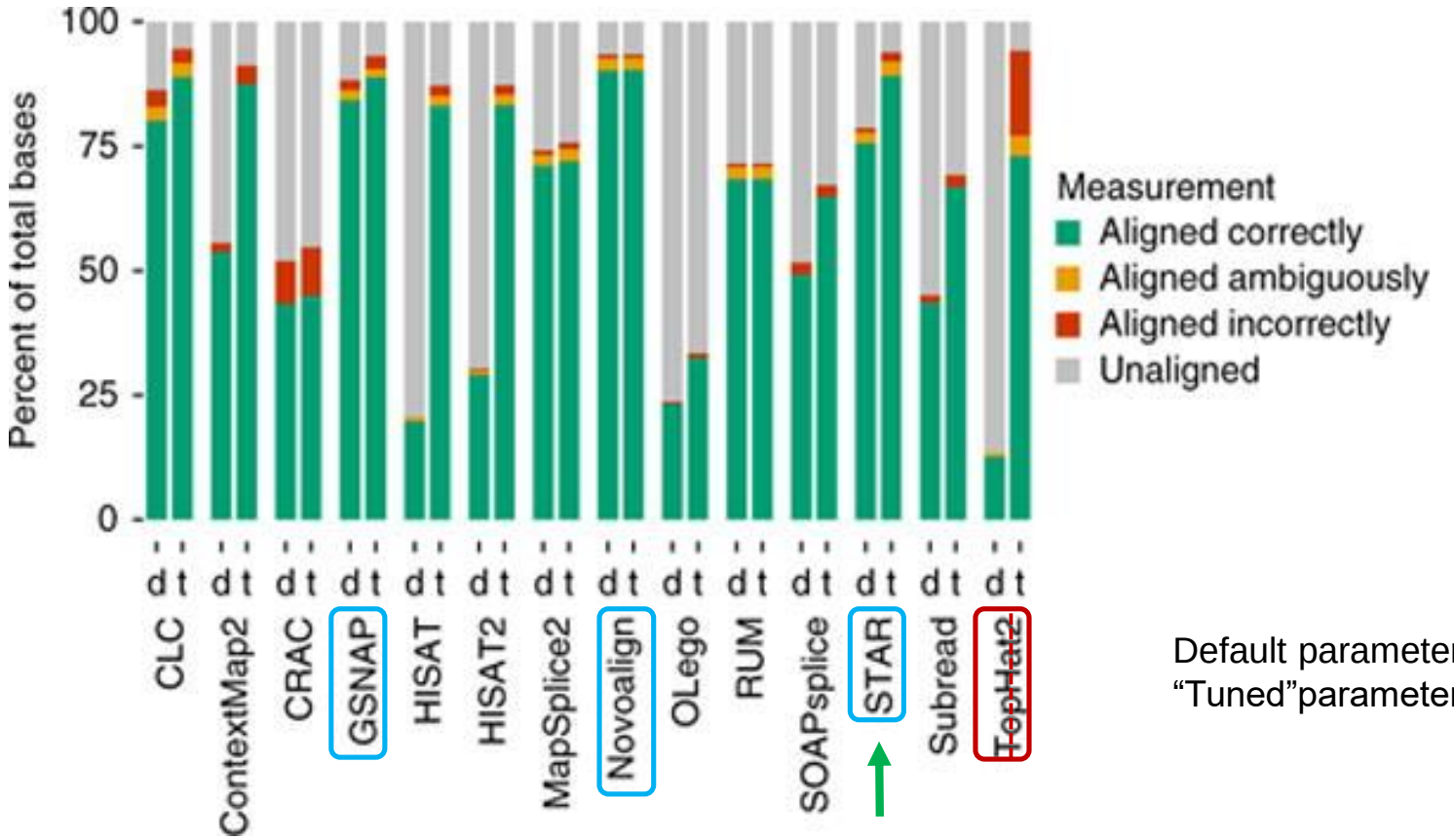
Important steps on NGS data analysis workflow



Which aligner should I run on my **RNA-seq samples**?

Tool	Functionality	Organisms
Bowtie2	Non-splice-aware local alignment against a ref. genome	lacking introns
Tophat2	Splice-aware local alignment against a ref. genome	Any
GSNAP	Splice-aware local alignment against a ref. genome	Any
STAR	Splice-aware local alignment against a ref. genome	Any
HISAT2	Splice-aware local alignment against a ref. genome	Any
Trinity	DBG *de novo* assembly	lacking a ref. genome
Salmon	Pseudoalignment against a ref. transcriptome	with a robust/reliable transcript isoforms annotation
Kallisto	Pseudoalignment against a ref. transcriptome	with a robust/reliable transcript isoforms annotation

Aligners performance on **RNA-seq data**



Default parameters (d)
"Tuned"parameters (t)

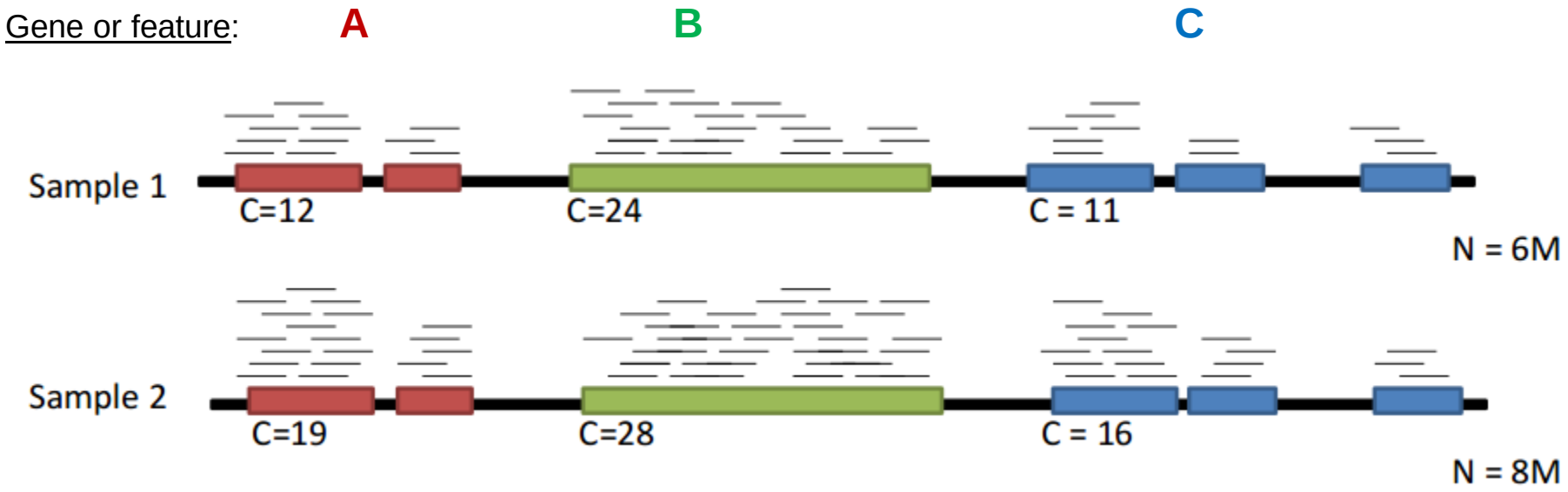
Baruzzo et al., 2017 – *Nat. Methods*

STAR command line

1. Generate your indexed reference genome
2. Run the alignment: sequencing reads (fastq format) -vs- reference genome

```
#$ -cwd -V
#$ -l h_rt=12:00:00,h_vmem=4G,nodes=1,ppn=24

## Bringing STAR executable to your own HPC user environment
export PATH=/nobackup/leedsomics_tools/STAR-2.7.10a/bin/Linux_x86_64_static/:$PATH

## Generating your indexed reference genome
STAR --runThreadN 24 --runMode genomeGenerate --genomeDir HsapSAindexedGenome --genomeFastaFiles Homo_sapiens.GRCh38.dna.primary_assembly.fa --sjdbGTFfile Homo_sapiens.GRCh38.105.gtf

## Aligning one paired-end sequencing sample (e.g. sampleX)
STAR --runMode alignReads --genomeDir HsapSAindexedGenome/ --readFilesIn sampleX_R1.fastq.gz sampleX_R2.fastq.gz --readFilesCommand zcat --outFileNamePrefix sampleX-vsGRCh38_  --outSAMtype BAM Sor
tedByCoordinate --outSAMattributes All --runThreadN 24

## In case there are several fastq files to be aligned, a for loop is more appropriate
for i in *_R1*.fq.gz; do STAR --runMode alignReads --genomeDir HsapSAindexedGenome/ --readFilesIn $i `echo $i | sed 's/_R1-/_R2-/g'` --readFilesCommand zcat --outFileNamePrefix `echo $i | sed 's/_
R[12].*$/-vsGRCh38_/g'` --outSAMtype BAM SortedByCoordinate --outSAMattributes All --runThreadN 24; done
```

Read Counts and Normalization Metrics



- **Reads or Fragments per kilobase per million (RPKM or FPKM)**

$$\textbf{RPKM} = \left(\frac{\#\text{aligned reads onto feature } (C)}{\text{feature length in kb} \times \#\text{total reads on sample } (N)} \right) \times 1{,}000{,}000$$

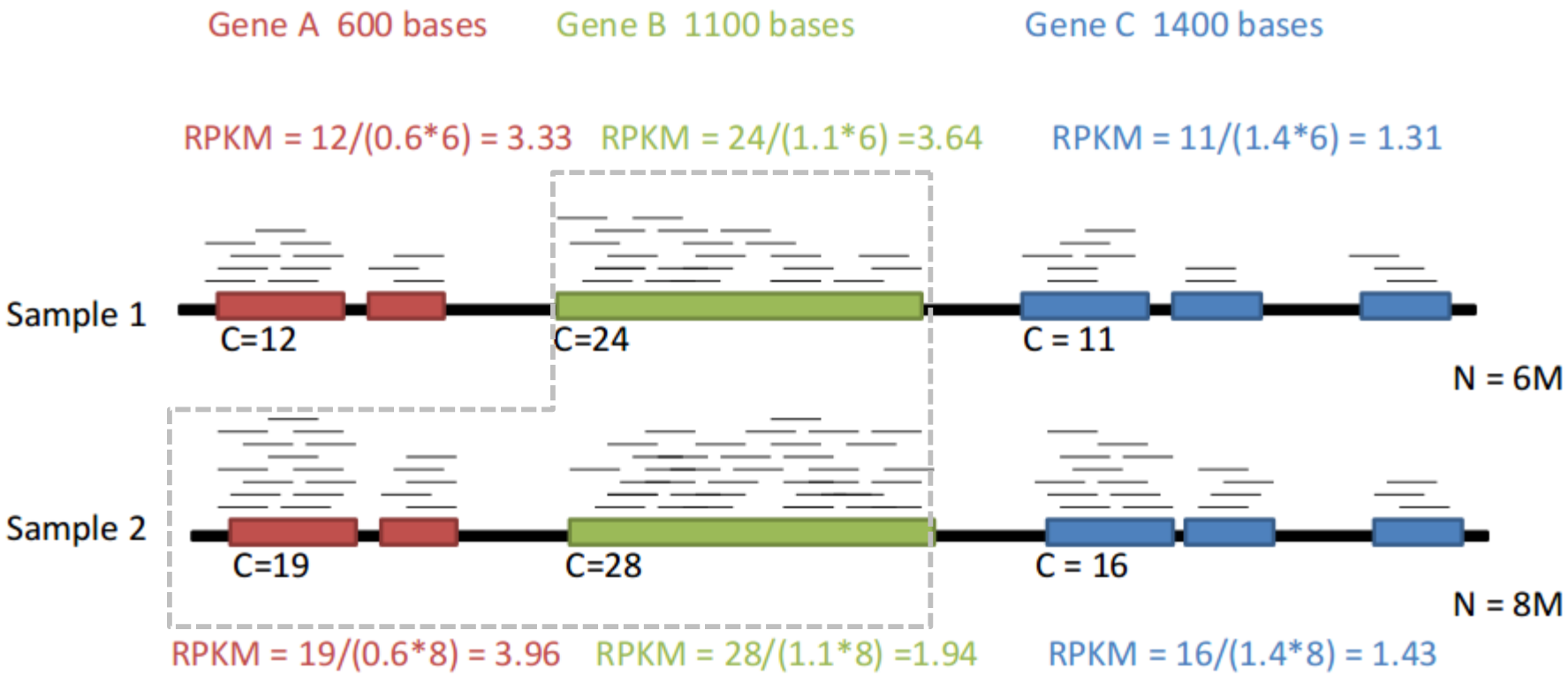
- **Transcripts per million**

TPM = (RPK / sum of all RPKs on sample) x 1,000,000

→ where **RPK** = #aligned reads onto feature (C) / feature length in kb

More details at https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/ (reliable source for consultation purposes only)

RPKM Example



https://izabelcavassim.wordpress.com/2015/03/09/rpkm-and-fpkm-normalization-units-of-expression/

Bring your issues on!