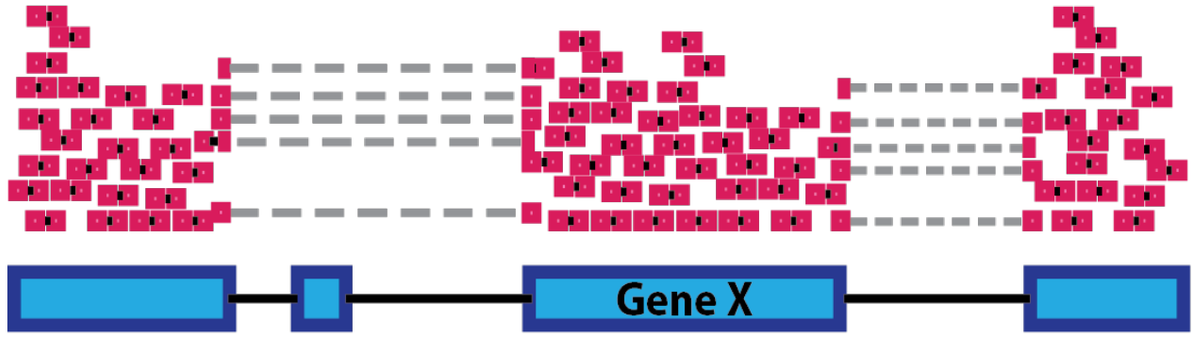
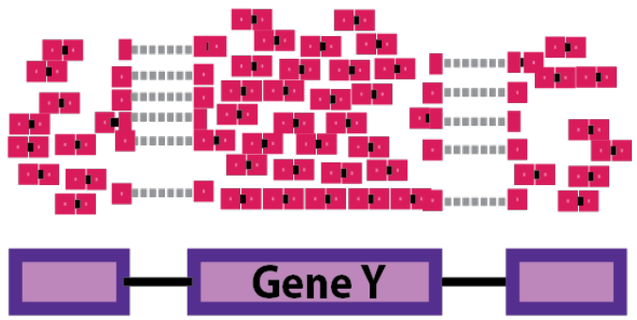
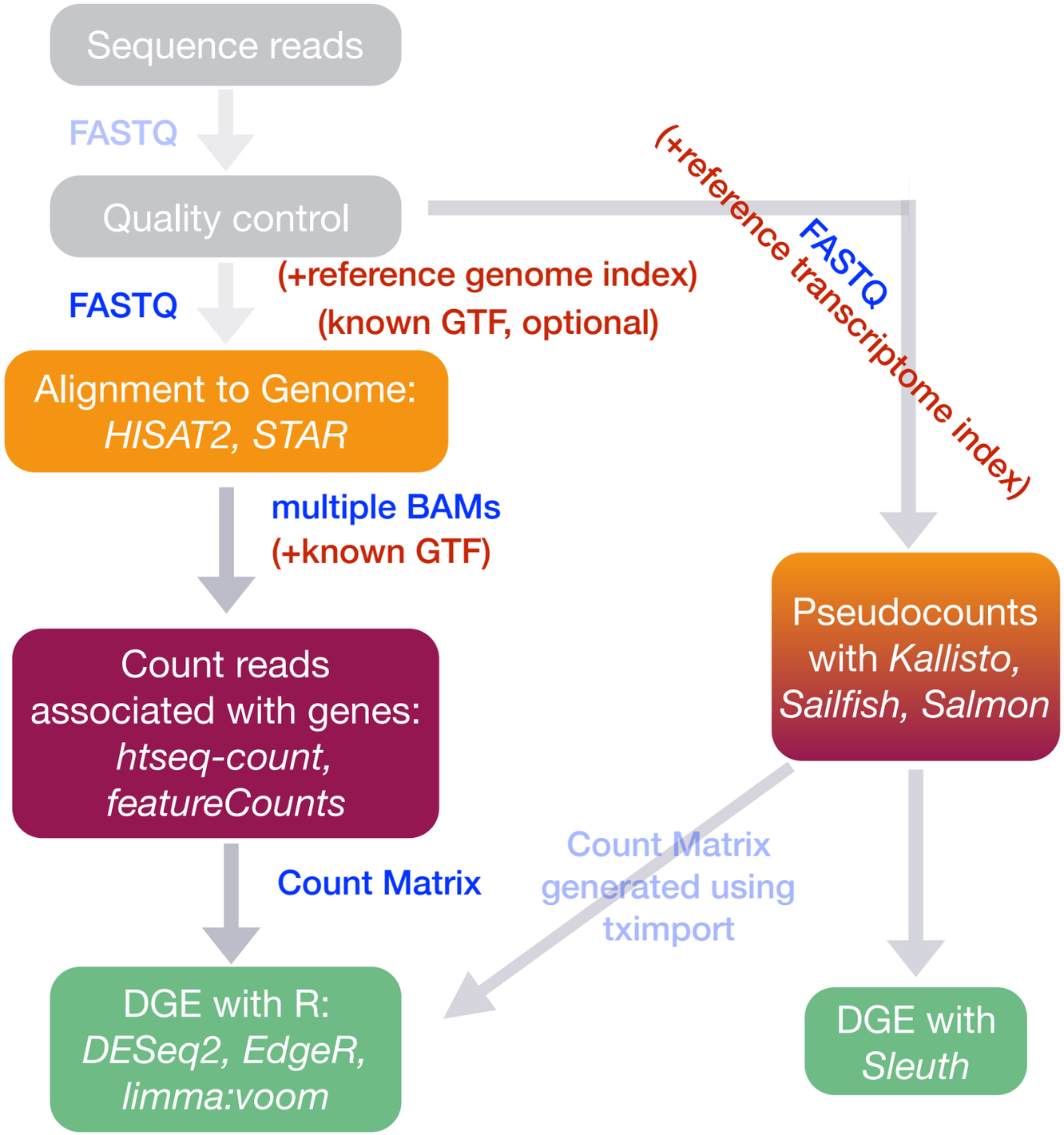
Quantifying gene expression

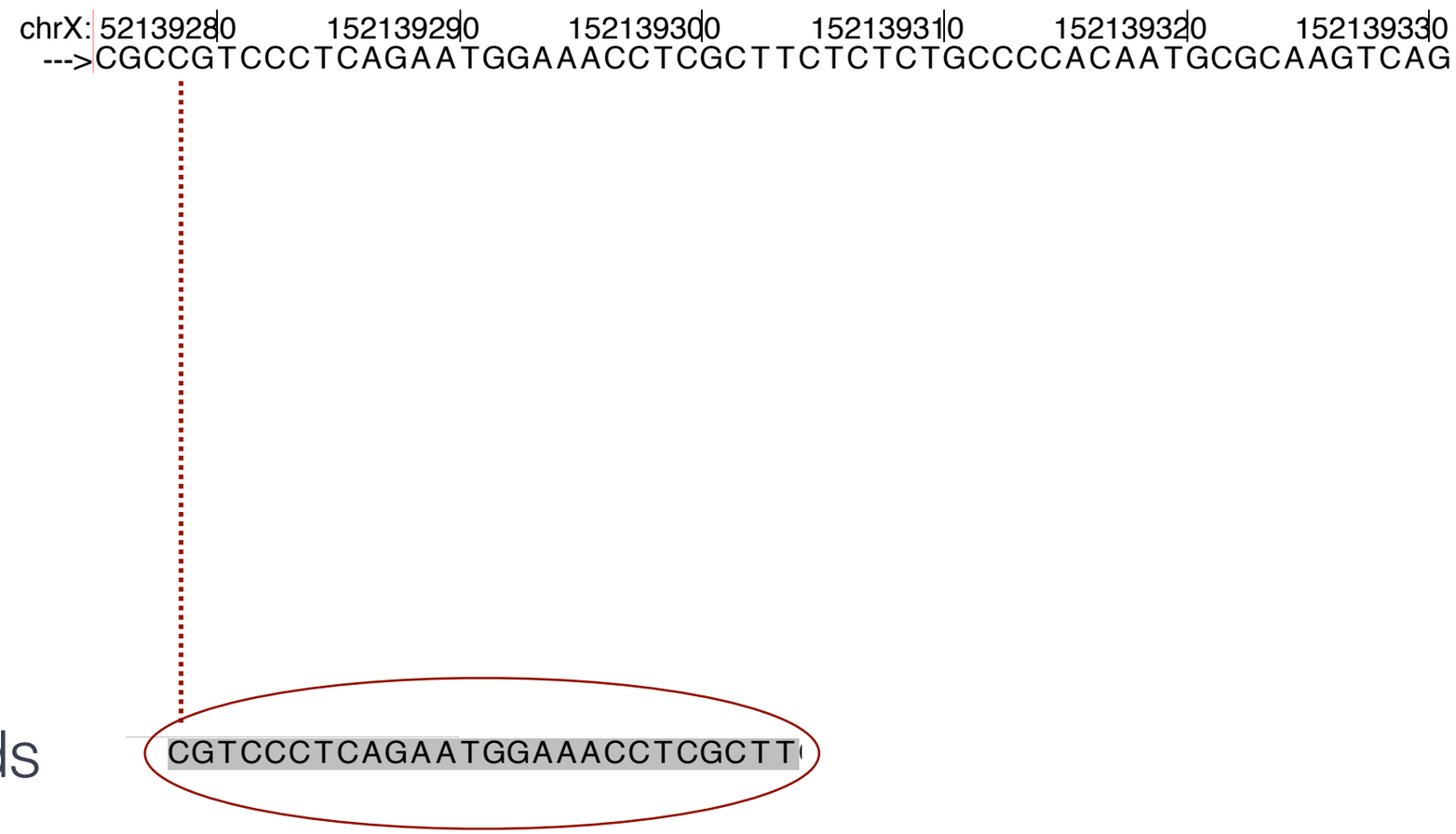




- ✓ Genome
- ✓ *GTF (annotation)?*



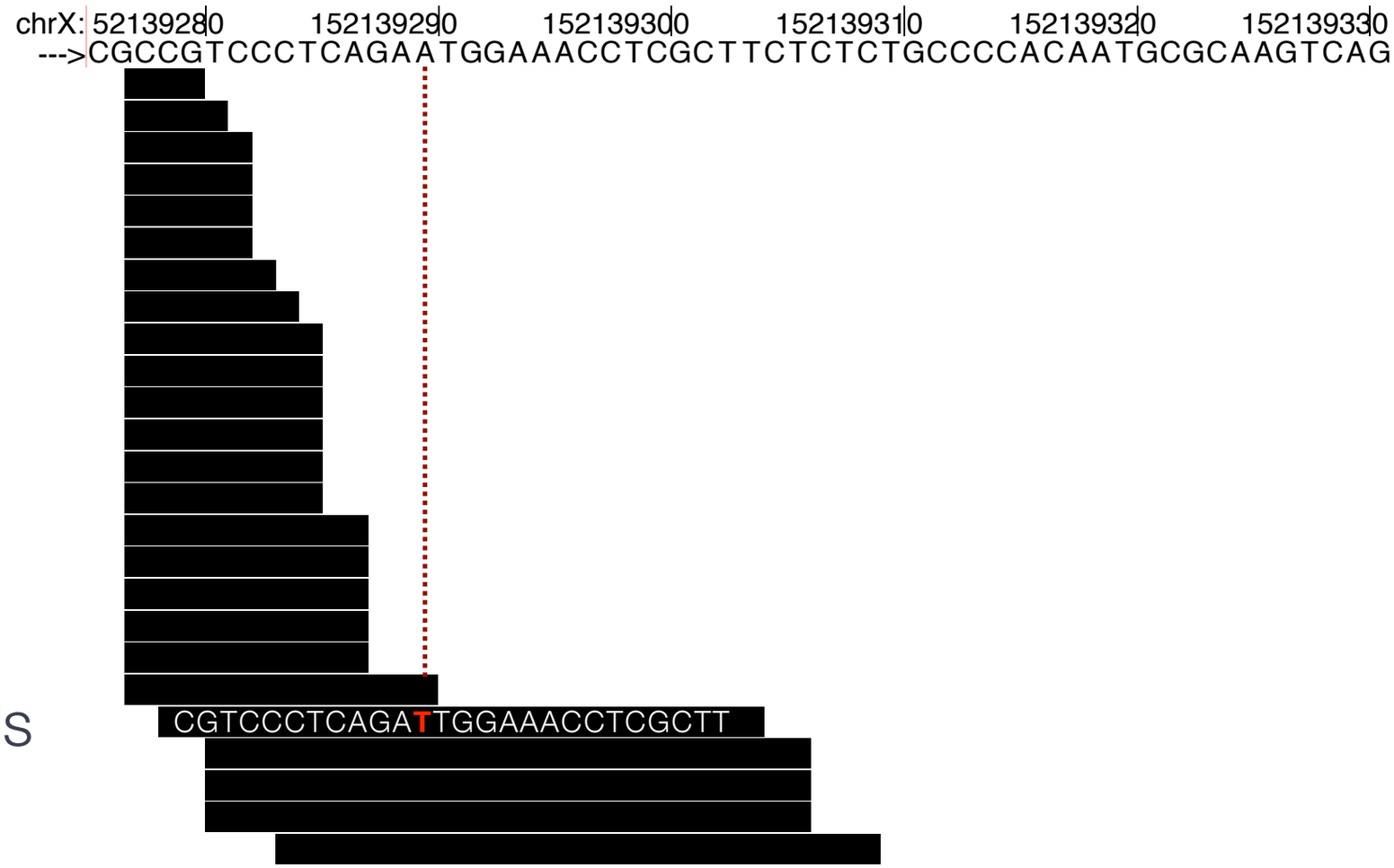
Genome



Sequence reads

A simple case of string matching

Genome



Sequence reads

A simple case of string matching?

Non-comprehensive list of challenges

- Large, incomplete and repetitive genomes OR transcriptomes with overlapping transcripts (isoforms)
- Short reads: 50-150 bp
  - Non-unique alignment
  - Sensitive to non-exact matching (variants, sequencing errors)
- Massive number of short reads
- Small insert size: 200-500 bp libraries
- Compute capacity for efficient mapping

Building an index

- Having an index of the reference sequence provides an efficient way to search
- Once index is built, it can be queried any number of times
- Every genome or transcriptome build requires a new index for the specific tool in question.

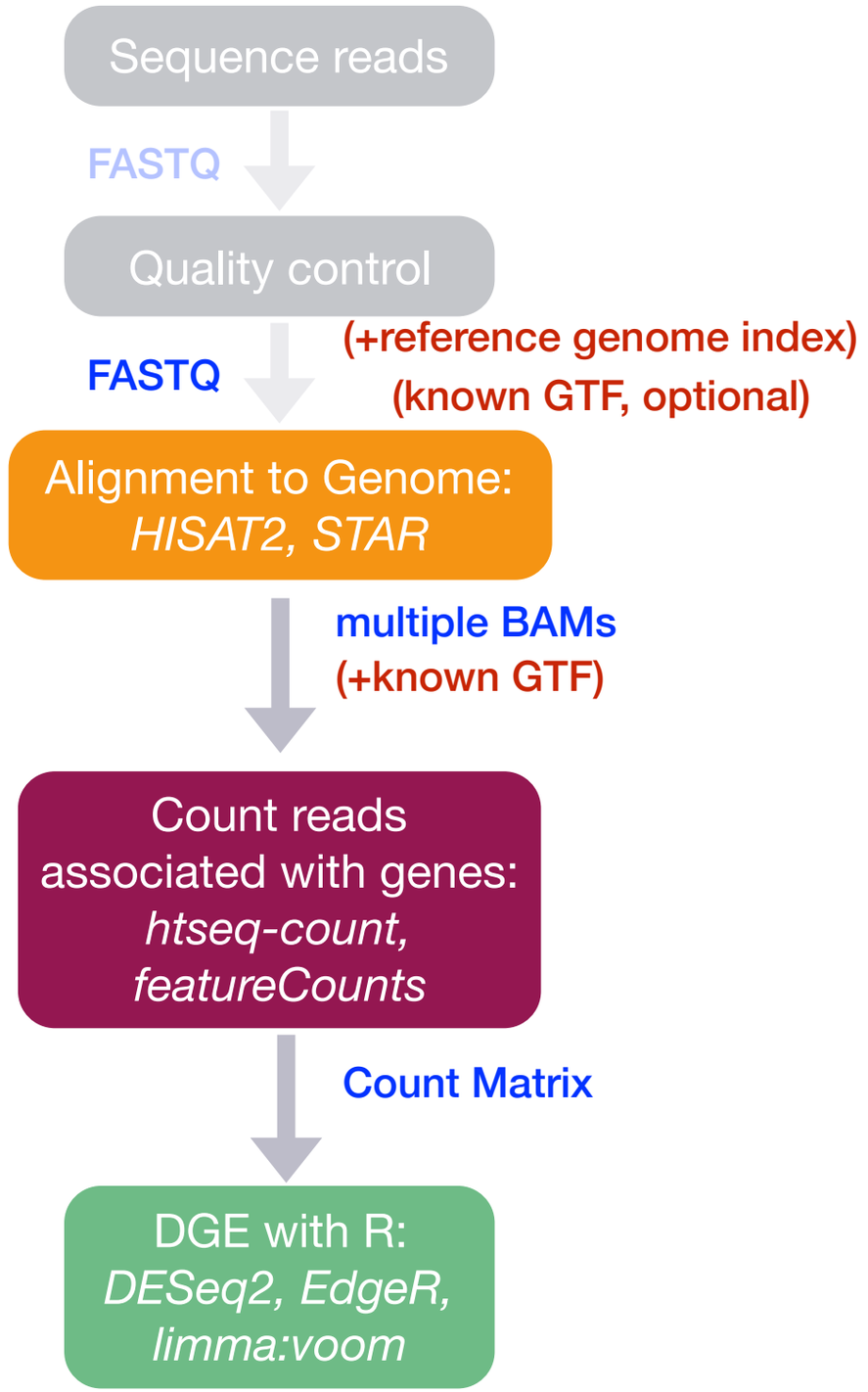
Commonly used indexing methods

- Hash-based (Salmon, Kallisto)
- Suffix arrays (Salmon, STAR)
- Burrows-Wheeler Transform (BWA, Bowtie2)

Genome versions matter

- Ensembl, UCSC and NCBI all often use the same genome assemblies or builds (e.g. GrCh38 == hg38)
- Make sure that the annotation file (GTF) is exactly matched with the genome file (fasta)
  - Same genome version
  - Same source (e.g. both from FlyBase)

- ✓ Genome FASTA
- ✓ GTF (annotation)



Alignment to genome

- Is it important that the genome index is created with awareness of known splice junctions?
- Don’t use default parameters; read the manual and ask questions about parameters
- Parameter sweeps may be needed if you are working on a non-model organism

BAM alignment files

- Binary version of SAM alignment format files
- Recommended over SAM files for saving alignments
- Contain information on a per-read basis:

  -- Coordinates of alignment, including strand

  -- Mismatches

  -- Mapping information (unique?, properly paired?, etc.)

  -- Quality of mapping (tool-specific scoring systems)

More information about SAM/BAM

QC on BAM files

Evaluating the quality of the aligned data can give important information about the quality of the library:

-- Total % of reads aligning to the genome? % of uniquely mapping reads? % of properly paired PE reads?

-- Genomic origin of reads (exonic, intronic, intergenic)

-- Quantity of rRNA

-- Transcript coverage and 5'-3' bias

Samples should have fairly consistent percentages.

QC on BAM files

Gather QC metrics using:

- *Log files from alignment run*
- *Qualimap*
- *RNASeQC (paper)*

More information about alignment QC

Quantification from BAM files

- htseq-count
- featureCounts

aligned read:

```
start: 113217600  end: 113217650
```

GTF

```
chr1 unknown  exon        113217048  113217252   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon        113217048  113217351   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon        113217470  113217671   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS         113217535  113217671   .   +   0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon 113217535  113217537   .   +       gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

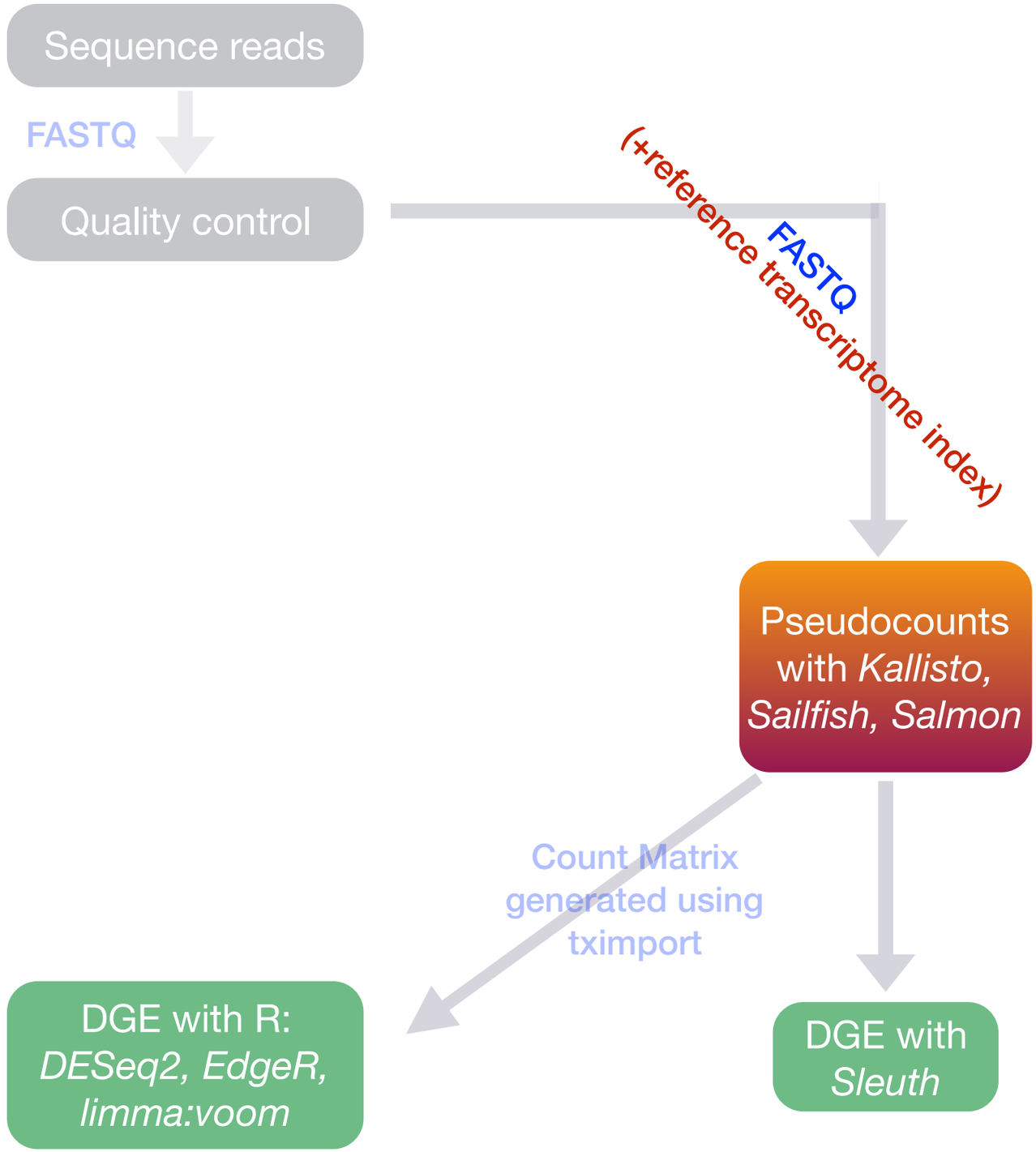
feature type

feature

Quantification from BAM files

- htseq-count and featureCounts
  - -- Strandedness
  - -- Stringency
- Results in a gene-level counts matrix (raw)
- Output ready for DGE analysis using tools like DESeq2 or EdgeR

✓ Transcriptome FASTA



More efficient quantification approaches

- Approaches that avoid base-to-base alignment
- Kallisto (quasi-aligner), Sailfish (kmer-based), Salmon (quasi-aligner), RSEM
- Faster, more efficient (~ >20x faster than alignment-based)
- Improved accuracy for transcript-level quantification
- Improvements in accuracy for gene-level quantification**

**doi: 10.12688/f1000research.7563.2

More efficient quantification approaches

- Results in a matrix of abundance estimates (not raw) at the isoform-level
- Abundance estimates can be used for differential isoform expression using sleuth (designed for Kallisto output)
- Gene-level counts can be calculated using tximport

  -- ready for DGE analysis using tools like DESeq2 or EdgeR