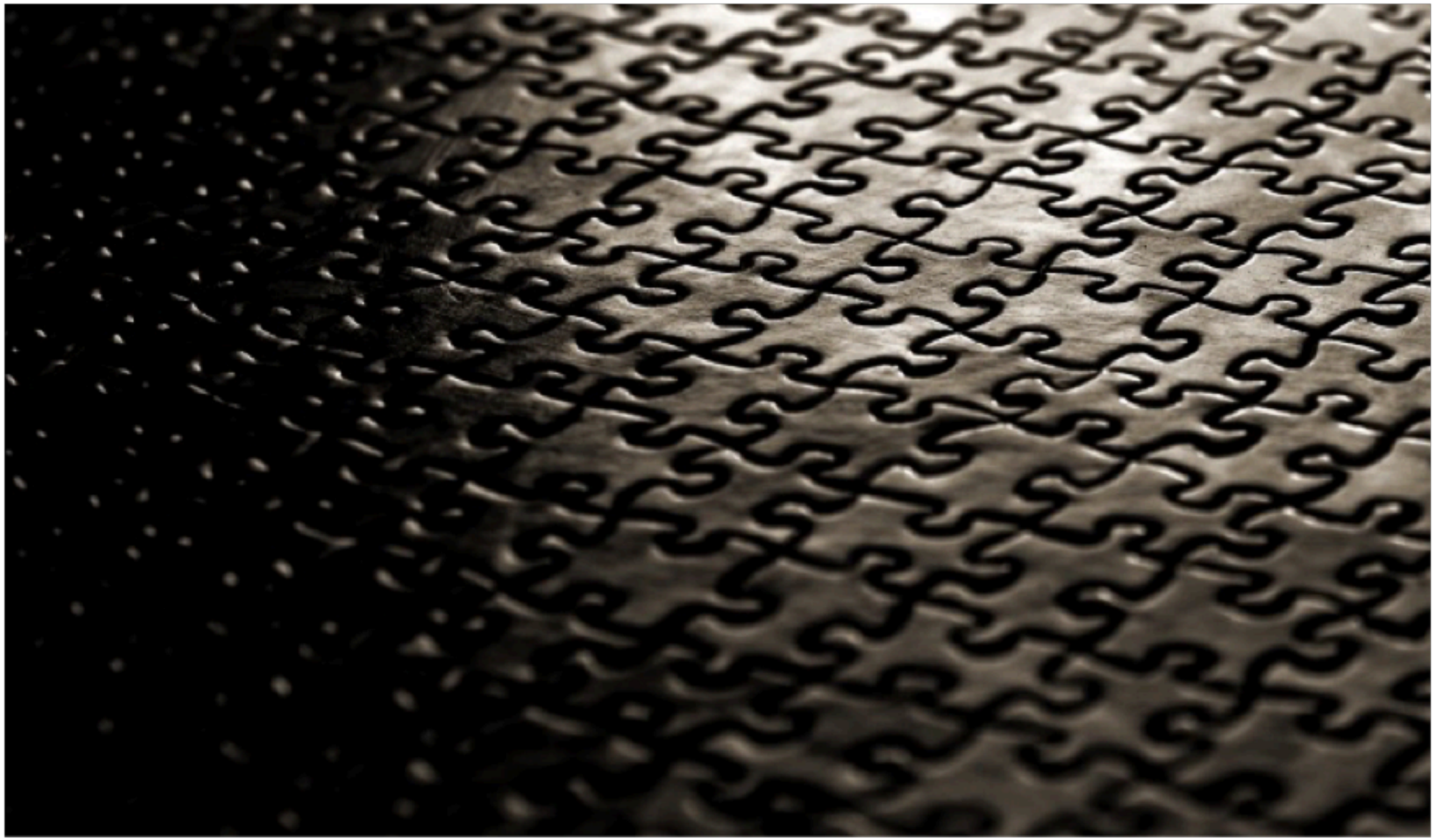


Aligning reads: tools and theory

Genome Mapping

Reads

STAR, HISAT2

Splice-aware Genome mapping

htseq-count, featureCounts

StringTie

Gene counting

Transcript discovery & counting

Homology-based BLAST2GO

Novel transcript annotation

Transcriptome Mapping

Reads

RSEM, Kallisto, Sailfish, Salmon

Transcript mapping and quantification

Assembly

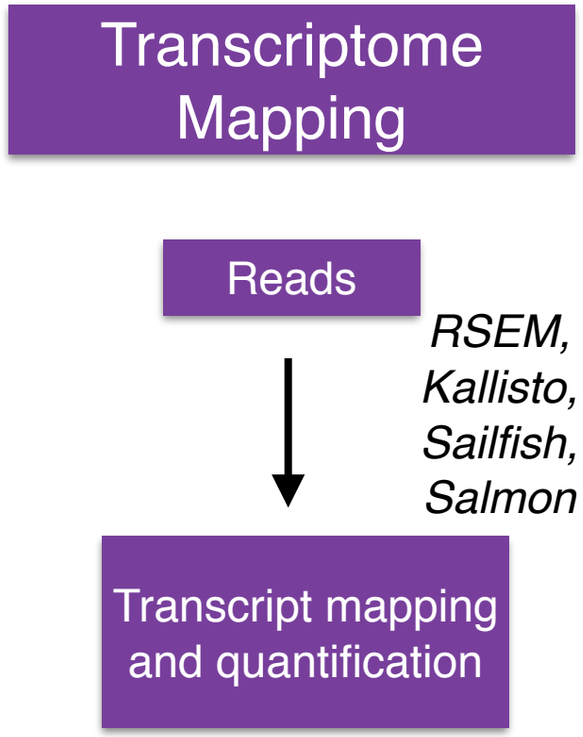
Reads

Trinity, Scripture, Stringtie

Assembly into transcripts

Trinotate

Novel transcript annotation


Transcriptome Mapping
Reads
RSEM, Kallisto, Sailfish, Salmon
Transcript mapping and quantification

Transcriptome Mapping

Reads

RSEM, Kallisto, Sailfish, Salmon

Transcript mapping and quantification

Biological samples/Library preparation

Sequence reads

FASTQ

Quality control: FASTQC

FASTQ

(+reference transcriptome index)

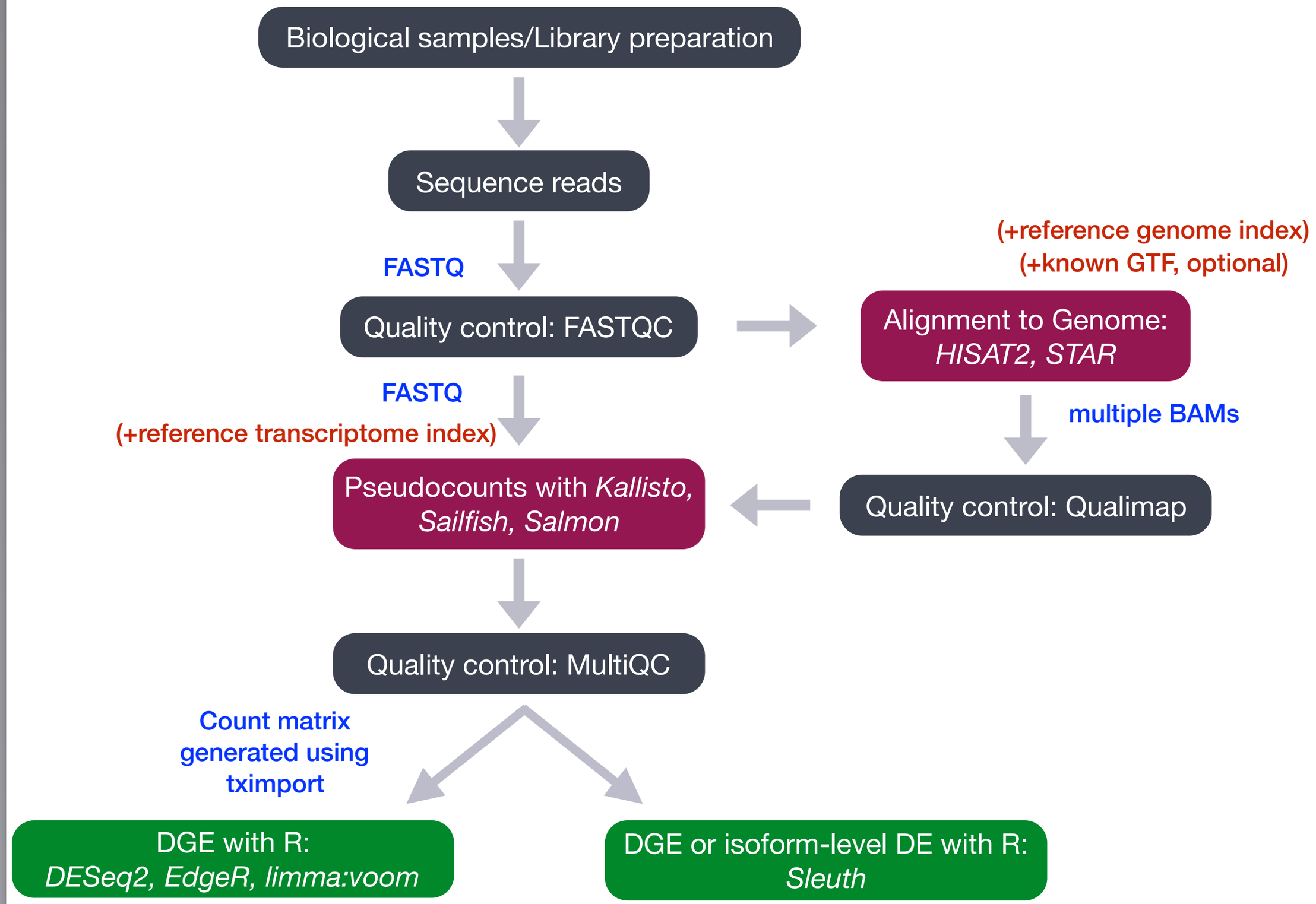
Pseudocounts with *Kallisto, Sailfish, Salmon*

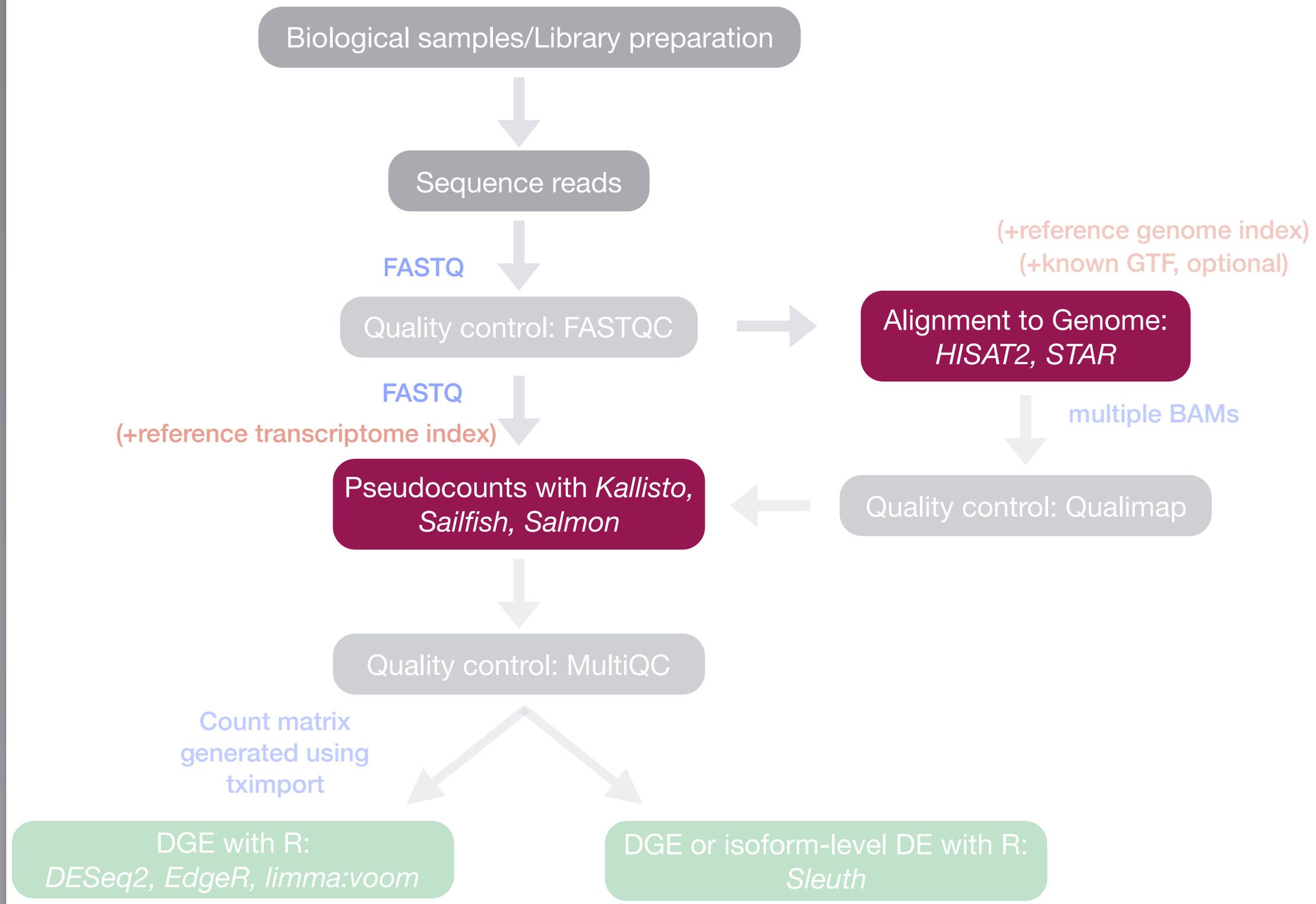
Quality control: MultiQC

Count matrix generated using tximport

DGE with R: *DESeq2, EdgeR, limma:voom*

DGE or isoform-level DE with R: *Sleuth*


Biological samples/Library preparation
Sequence reads
FASTQ
Quality control: FASTQC
FASTQ
(+reference transcriptome index)
Pseudocounts with Kallisto, Sailfish, Salmon
Quality control: MultiQC
Count matrix generated using tximport
DGE with R: DESeq2, EdgeR, limma:voom
DGE or isoform-level DE with R: Sleuth
(+reference genome index)
(+known GTF, optional)
Alignment to Genome: HISAT2, STAR
multiple BAMs
Quality control: Qualimap


Biological samples/Library preparation
Sequence reads
FASTQ
Quality control: FASTQC
FASTQ
(+reference transcriptome index)
Pseudocounts with Kallisto, Sailfish, Salmon
Quality control: MultiQC
(+reference genome index)
(+known GTF, optional)
Alignment to Genome: HISAT2, STAR
multiple BAMs
Quality control: Qualimap
Count matrix generated using tximport
DGE with R: DESeq2, EdgeR, limma:voom
DGE or isoform-level DE with R: Sleuth

Goal: Finding where in the genome these reads originated from

Genome

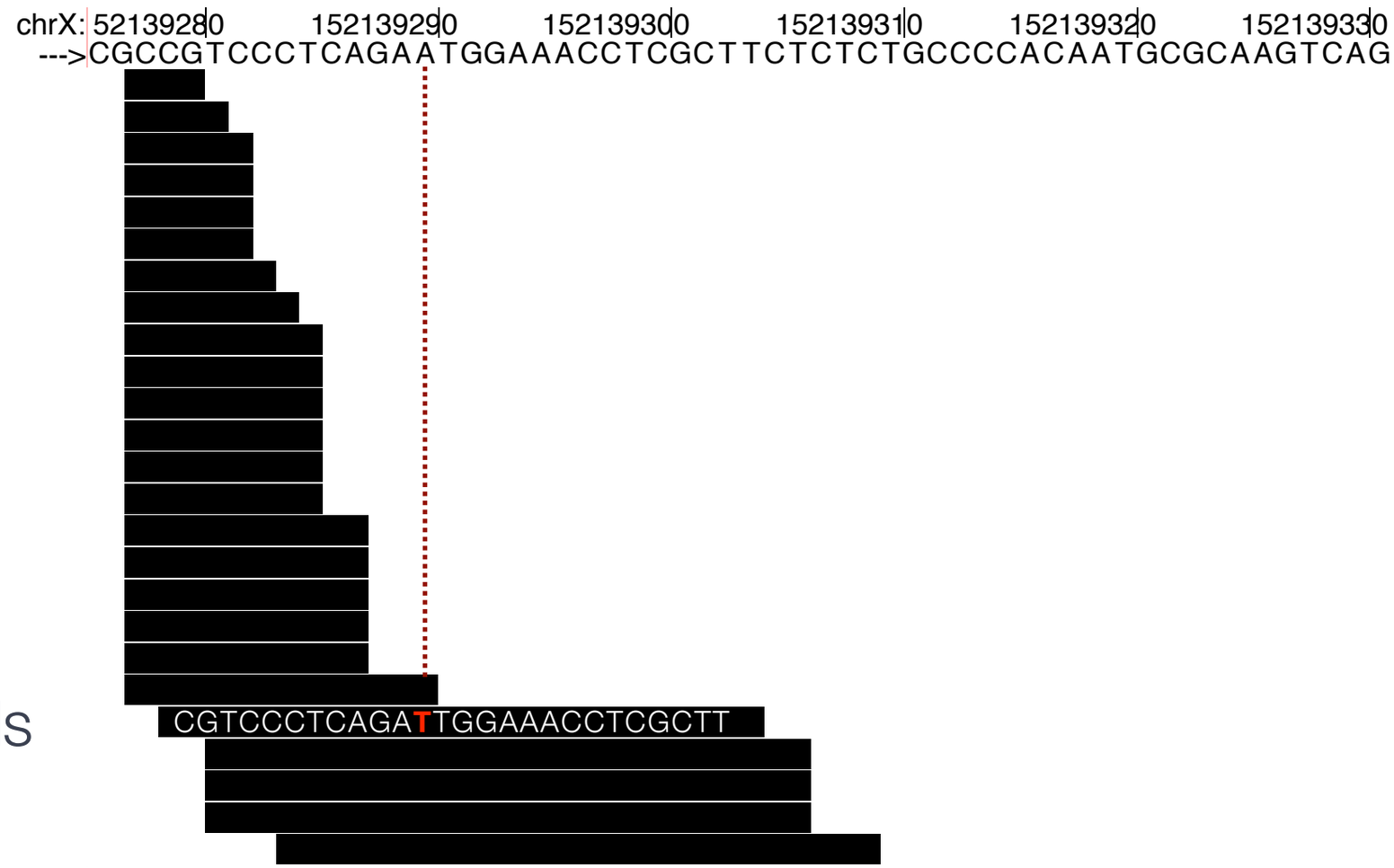
chrX: 52139280 52139290 52139300 52139310 52139320 52139330
--->CGCCGTCCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG

Sequence reads

CGTCCCTCAGAATGGAAACCTCGCTT

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

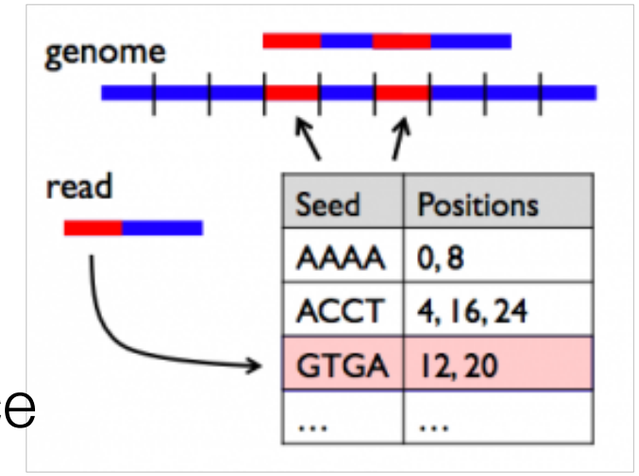
- Having an index of the reference sequence provides an efficient way to search
- Once index is built, it can be queried any number of times
- Every genome or transcriptome build requires a new index for the specific tool in question.

Commonly used indexing methods

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

Hash-based alignment (circa 1990)

- Pick k-mer size, build lookup of every k-mer in the reference mapped to its positions (the index)
- Break the query into k-mers
- Seed-and-extend strategy
- For BLAST, 100% match the query k-mer to reference then extend until score drops below 50%
- 0.1 - 1 sec per query; not feasible for NGS data



Hash-based alignment (present day)

- Need to make some concessions on sensitivity by making adaptations for use on NGS data:
  - allow for mismatches and/or gaps (ELAND, MAQ, SOAP)
  - using multiple seeds (BLAT, ELAND2)
- Memory intensive and slower (~16GB RAM required for hg19)
- Simpler in design but more sensitive

Suffix arrays

- A sorted table of all suffixes (substrings) of a given string
- A suffix array will contain integers that represent the starting indexes of the all the suffixes of a given string, after the aforementioned suffixes are sorted
- Requires large amount of memory to load the suffix array and genome sequence prior to alignment

Let the given string be "`mississippi`"

Suffixes	ID	Sorted Suffixes	Suffix Array
mississippi$	1	$	12
ississippi$	2	i$	11
ssissippi$	3	ippi$	8
sissippi$	4	issippi$	5
issippi$	5	ississippi$	2
ssippi$	6	mississippi$	1
sippi$	7	pi$	10
ippi$	8	ppi$	9
ppi$	9	sippi$	7
pi$	10	sissippi$	4
i$	11	ssippi$	6
$	12	ssissippi$	3

The suffix array will be:
{12, 11, 8, 5, 2, 1, 10, 9, 7, 4, 6, 3}

Burrows-Wheeler transform

- A compressed form of suffix arrays
- Tends to put runs of the same character together rather than alphabetically, which makes the compression work well

Suffixes ID	Sorted Suffixes	Suffix Array	Sorted Rotations (A_s matrix)	BWT Output (L)
mississippi$ 1	$	12	$mississippi	i
ississippi$ 2	i$	11	i$mississipp	p
ssissippi$ 3	ippi$	8	ippi$mississ	s
sissippi$ 4	issippi$	5	issippi$miss	s
issippi$ 5	ississippi$	2	ississippi$m	m
ssippi$ 6	mississippi$	1	mississippi$	$
sippi$ 7	pi$	10	pi$mississip	p
ippi$ 8	ppi$	9	ppi$mississi	i
ppi$ 9	sippi$	7	sippi$missis	s
pi$ 10	sissippi$	4	sissippi$mis	s
i$ 11	ssippi$	6	ssippi$missi	i
$ 12	ssissippi$	3	ssissippi$mi	i

Burrows-Wheeler transform

- Much less memory because of compression;

  ~1.5 GB of RAM required for hg19 index

- But compression results in diminished

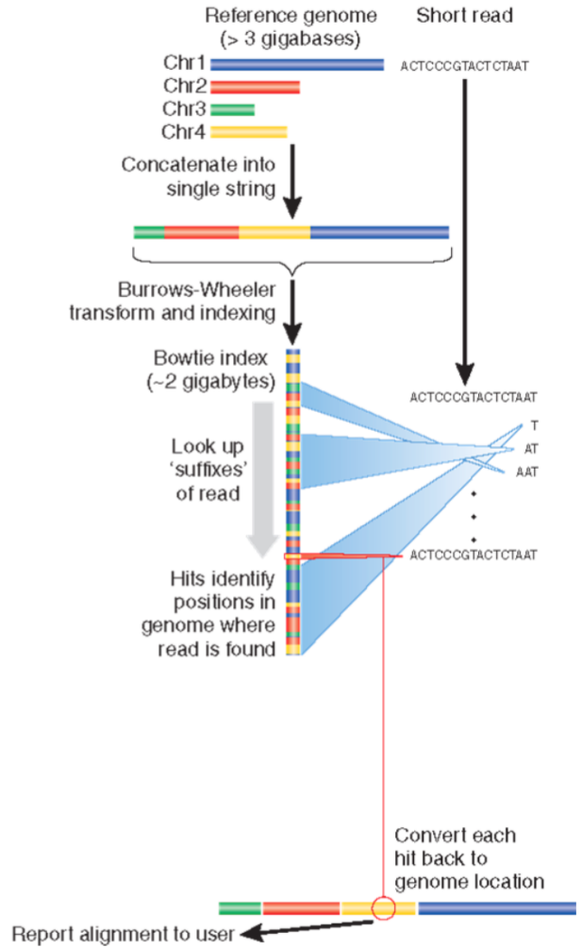
  efficiency of the string search operations

- Popular Tools:

  Bowtie2 (2012)

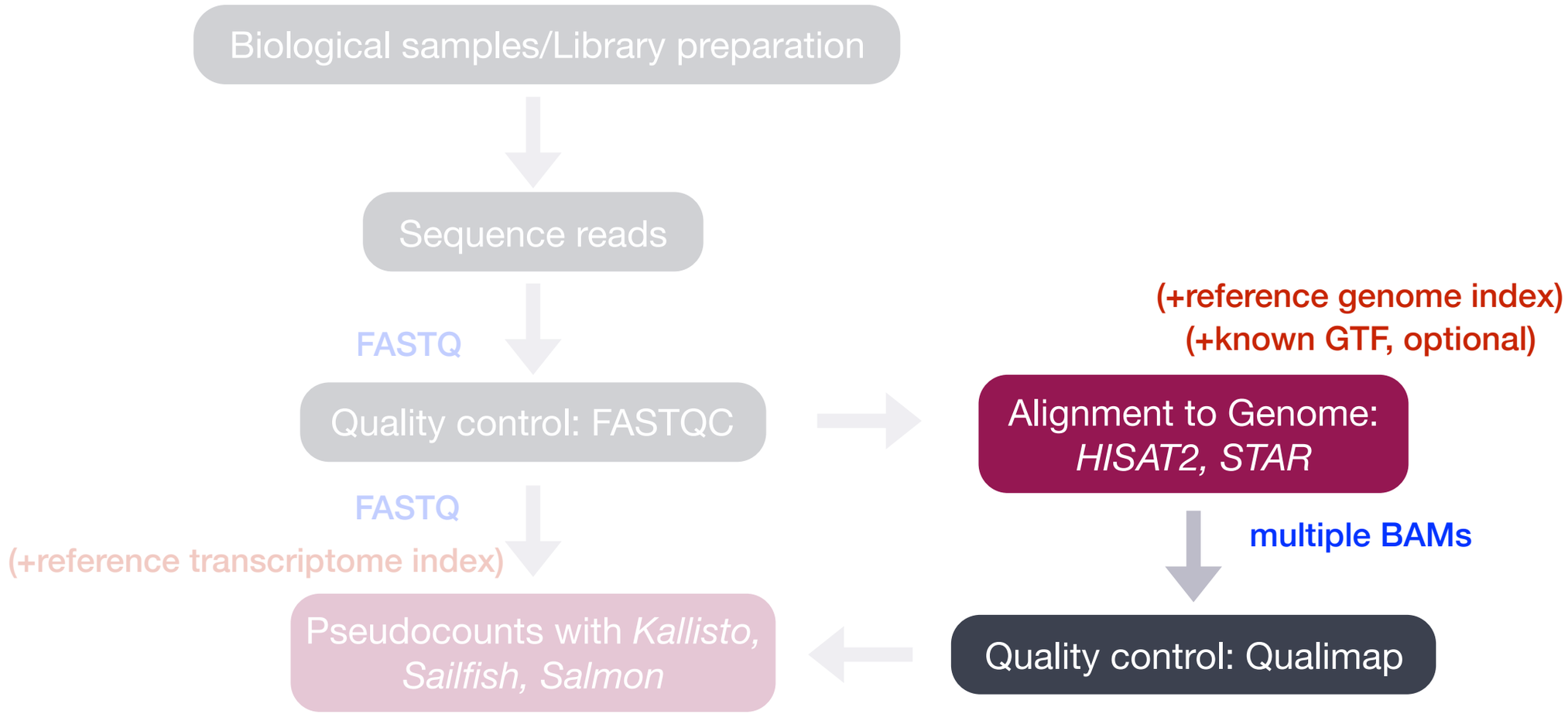
  SOAP2

  BWA-MEM (2013)



Reference data versions matter

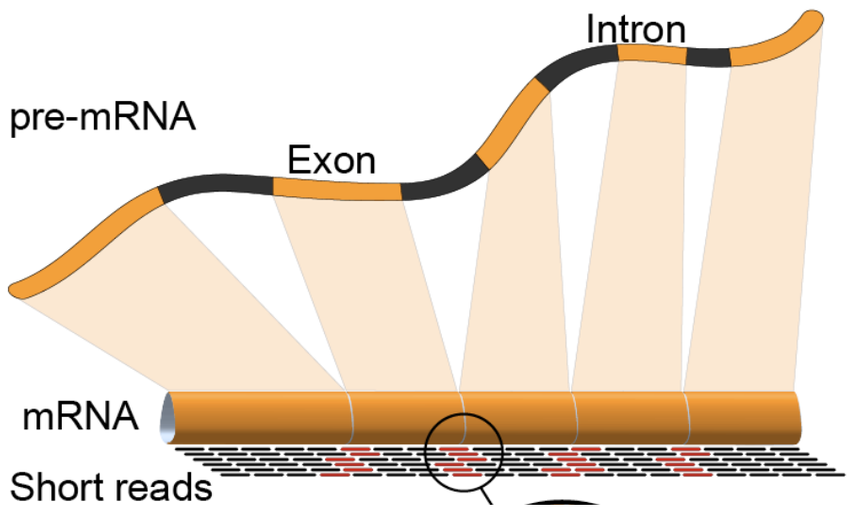
- Ensembl, UCSC and NCBI all often use the same genome assemblies or builds (e.g. GrCh38 == hg38)
- Make sure that the annotation file (GTF) is exactly matched with the genome file (fasta), or transcriptome file (fasta)
  - Same build version
  - Same source (e.g. both from FlyBase)

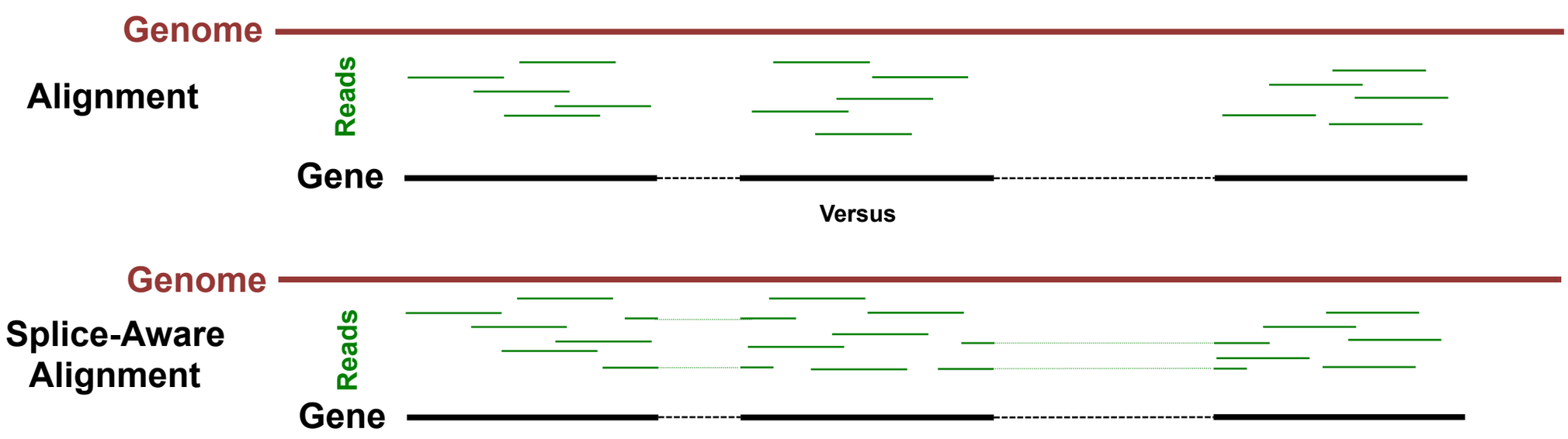


Genome alignment for QC

Alignment to genome

- Using an alignment tool that is aware of known splice junctions
- Don’t use default parameters; read the manual and ask questions
- Parameter sweeps may be needed if you are working on a non-model organism





Splice-aware alignment

Splice-aware alignment tools:

HISAT2, STAR, MapSplice, SOAPSplice, Passion, SpliceMap, RUM, ABMapper, CRAC, GSNAP, HMMSplicer, Olego, BLAT

There are excellent aligners available that are not splice-aware. These are useful for aligning directly to genes. However, you will lose isoform information.

Bowtie2, BWA, Novoalign (not free), SOAPaligner

Genome alignment output: SAM/BAM

- Sequence Alignment Map (SAM) format contains 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)
- BAM: Binary version of SAM alignment format files

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

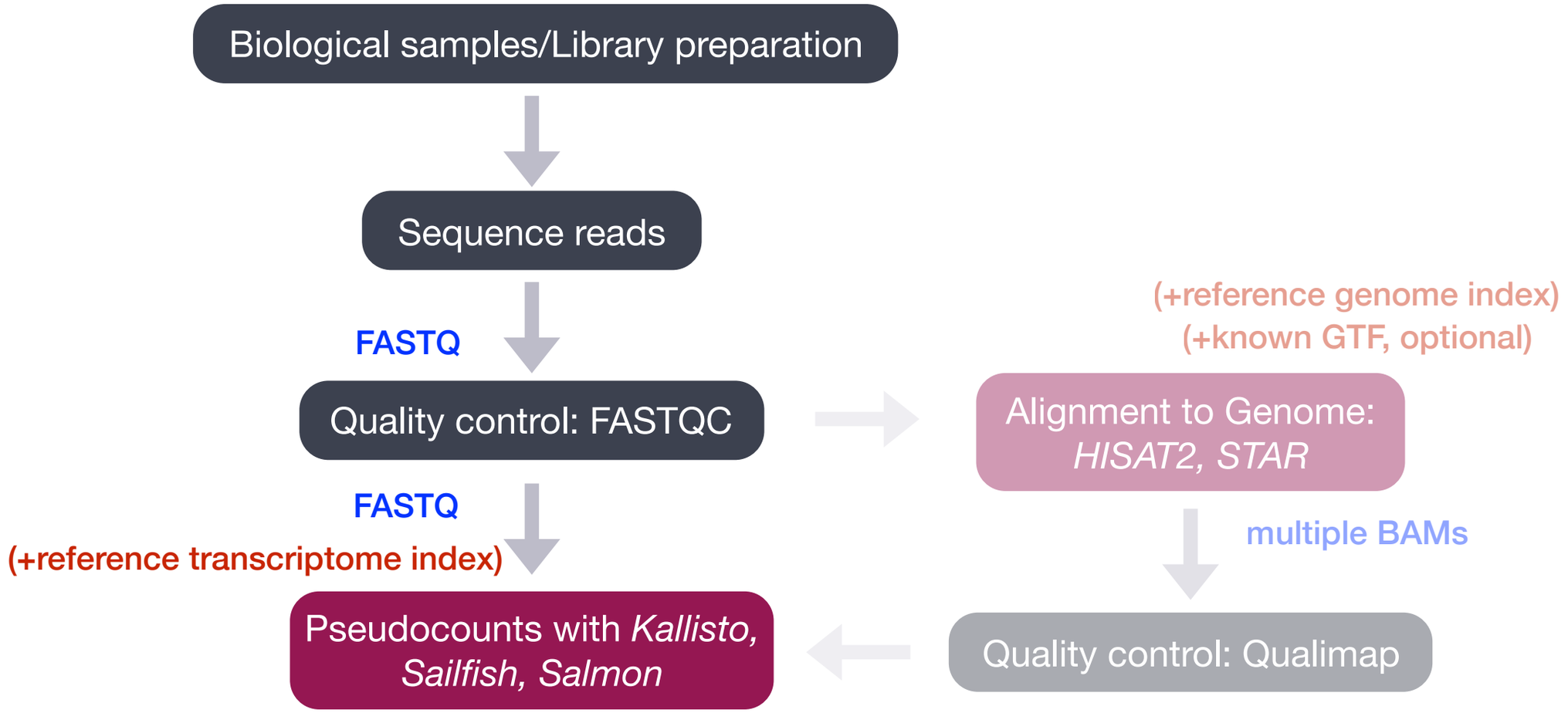
Samples should have fairly consistent percentages.

QC on BAM files

Gather QC metrics using:

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

More information about alignment QC



Transcriptome mapping for quantification

More efficient quantification approaches

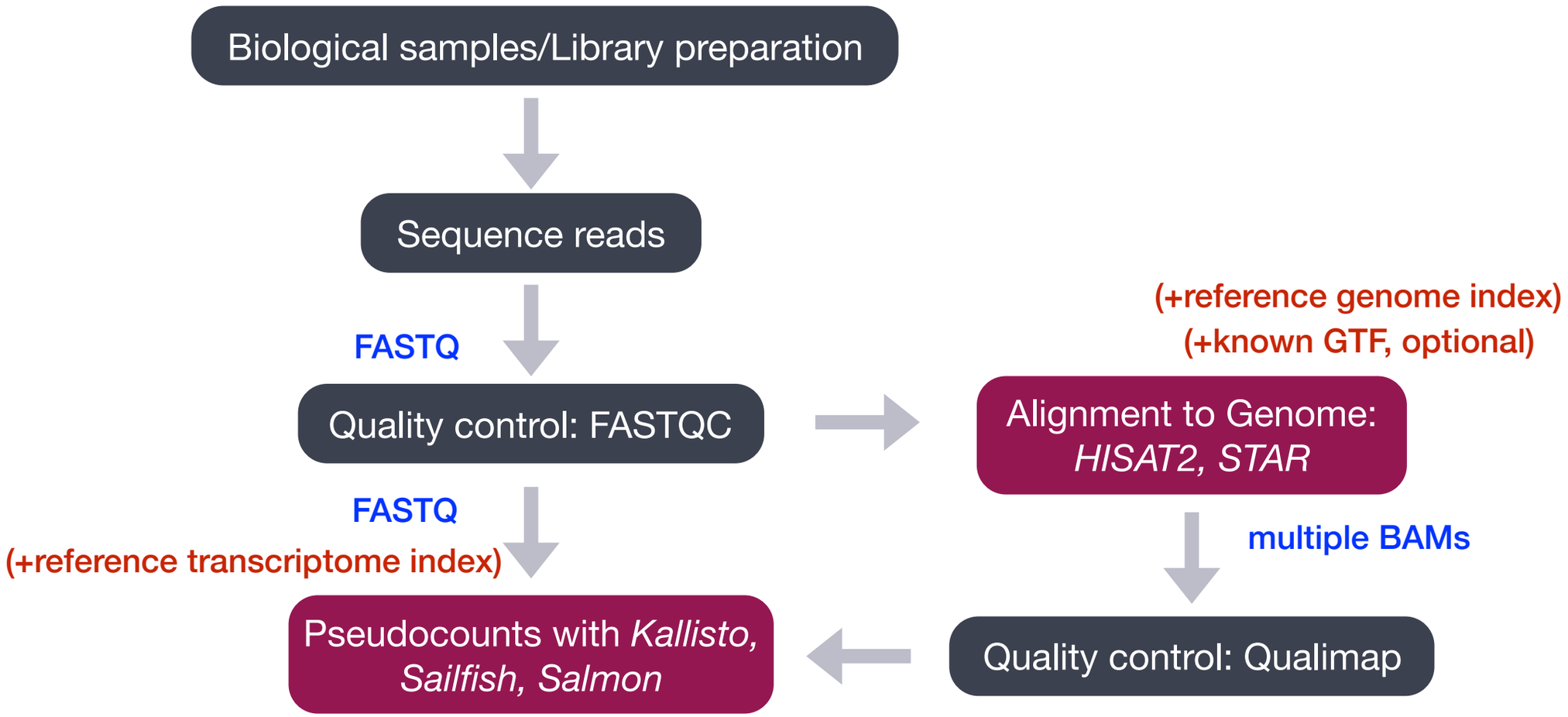
- Approaches that avoid base-to-base alignment
- Kallisto (quasi-aligner), Sailfish (kmer-based), Salmon (quasi-aligner), RSEM
- Faster, more efficient (~ >20x faster than alignment-based)
- Improved accuracy for transcript-level quantification
- Improvements in accuracy for gene-level quantification**

**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


Biological samples/Library preparation
Sequence reads
FASTQ
Quality control: FASTQC
FASTQ
(+reference transcriptome index)
Pseudocounts with Kallisto, Sailfish, Salmon
(+reference genome index)
(+known GTF, optional)
Alignment to Genome: HISAT2, STAR
multiple BAMs
Quality control: Qualimap