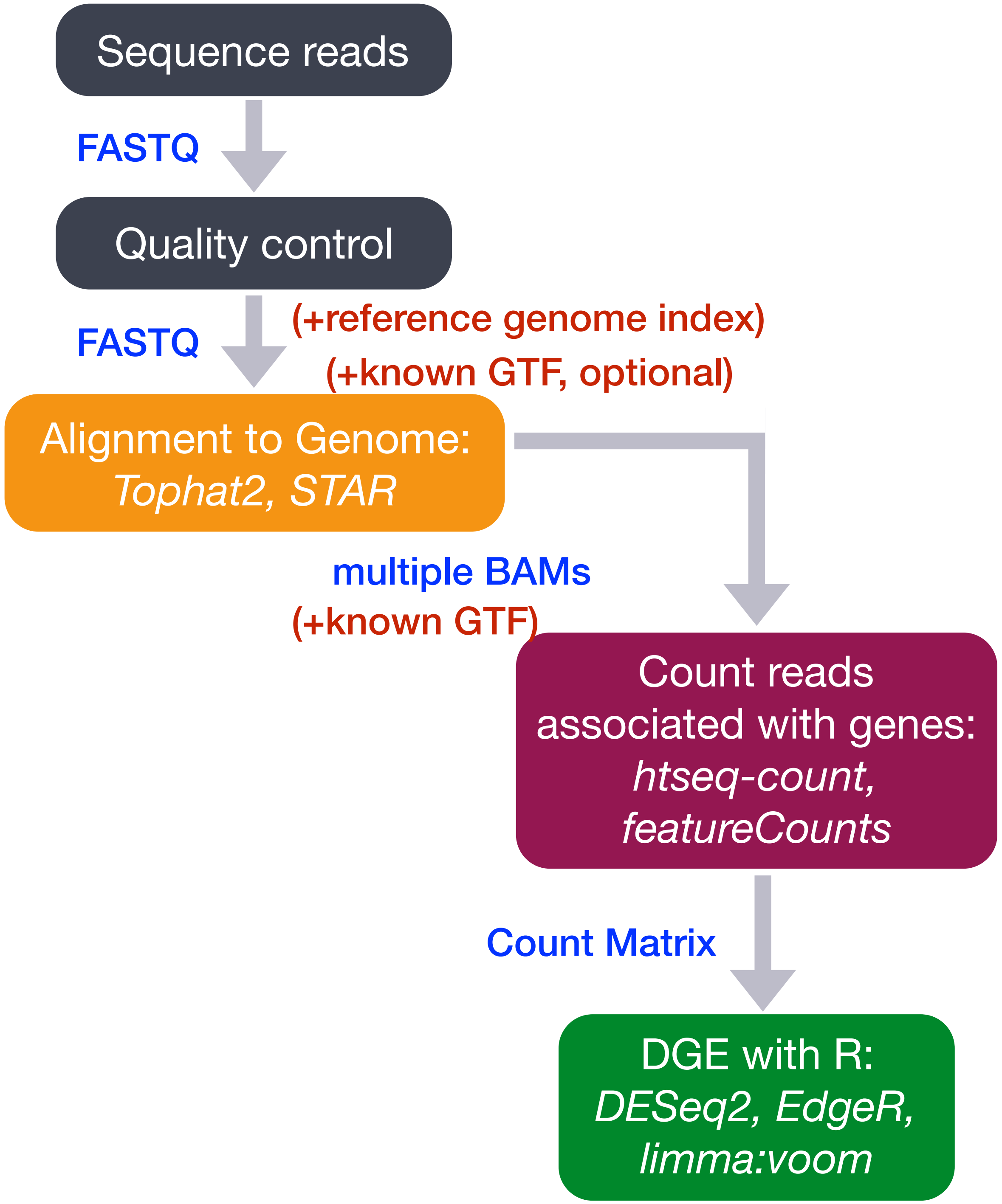
RNA-Seq: Analysis methods



Sequence reads

FASTQ

Quality control

FASTQ

(+reference genome index)

(+known GTF, optional)

Alignment to Genome: *Tophat2, STAR*

multiple BAMs

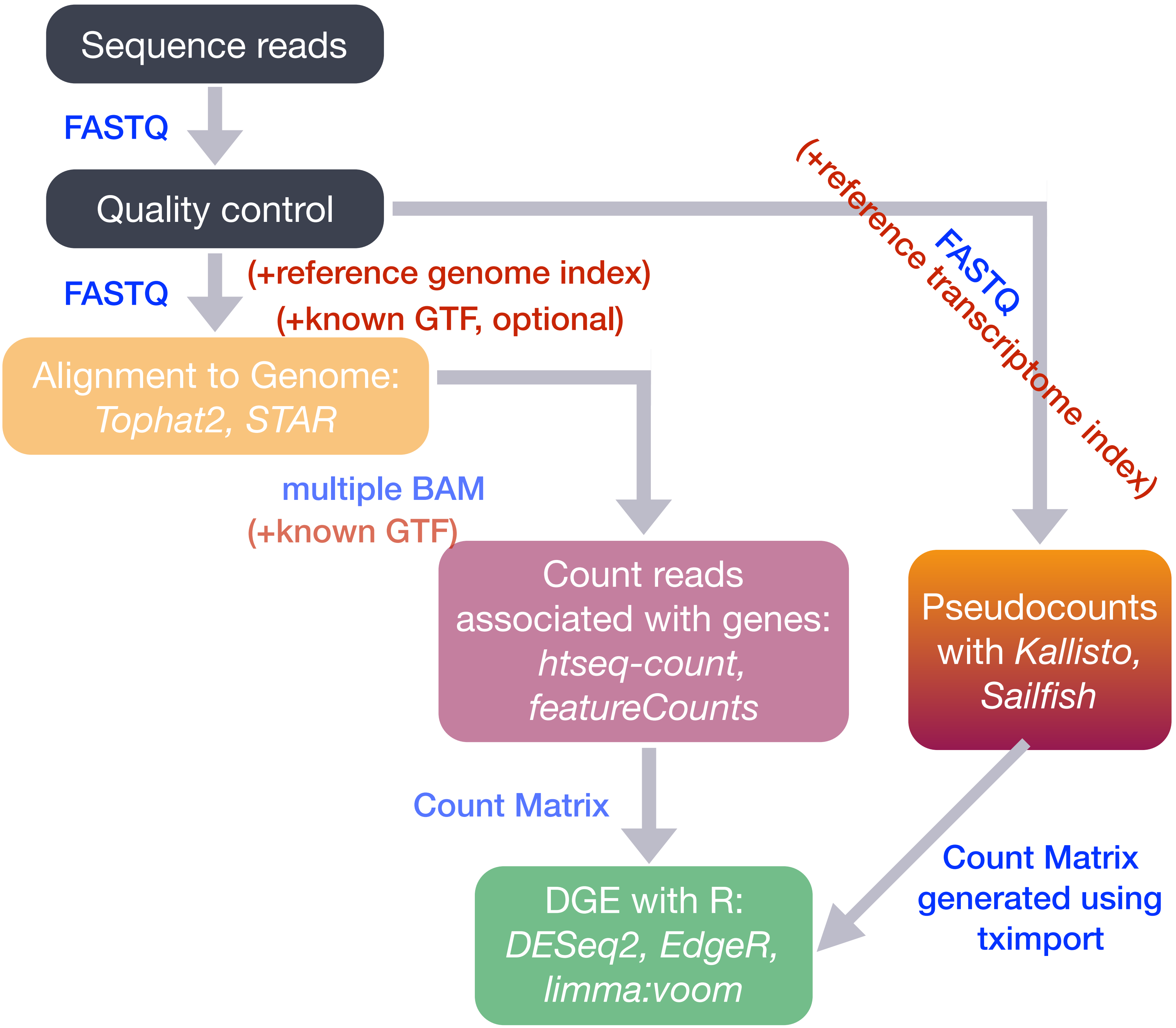
(+known GTF)

Count reads associated with genes: *htseq-count, featureCounts*

Count Matrix

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

Differential Gene Expression Analysis Workflow



Sequence reads

FASTQ

Quality control

FASTQ

(+reference genome index)

(+known GTF, optional)

Alignment to Genome: *Tophat2, STAR*

multiple BAM

(+known GTF)

Count reads associated with genes: *htseq-count, featureCounts*

Count Matrix

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

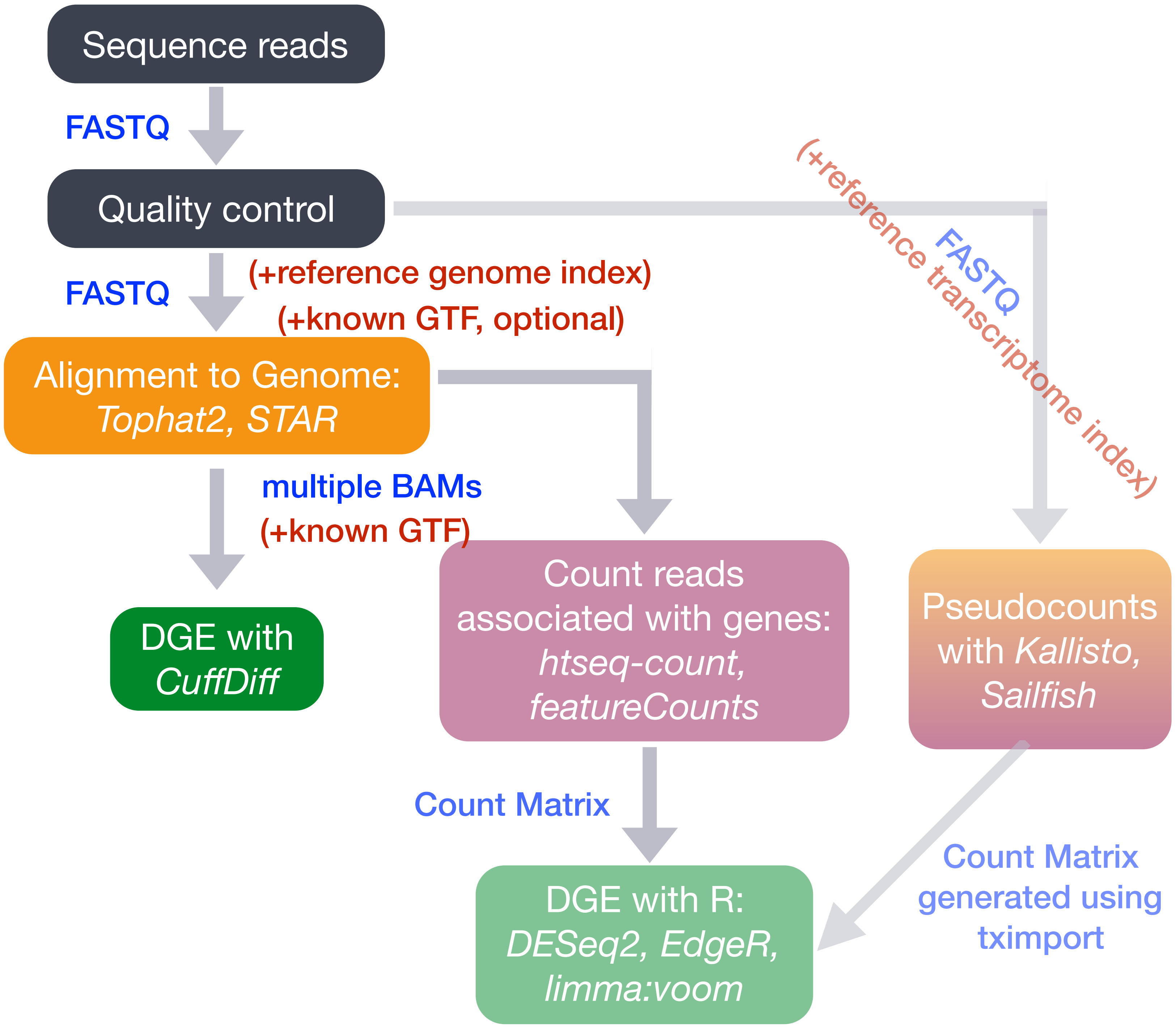
FASTQ

(+reference transcriptome index)

Pseudocounts with *Kallisto, Sailfish*

Count Matrix generated using tximport

Differential Gene Expression Analysis Workflow



Sequence reads

FASTQ

Quality control

FASTQ

(+reference genome index)

(+known GTF, optional)

Alignment to Genome: *Tophat2, STAR*

multiple BAMs

(+known GTF)

DGE with *CuffDiff*

Count reads associated with genes: *htseq-count, featureCounts*

Count Matrix

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

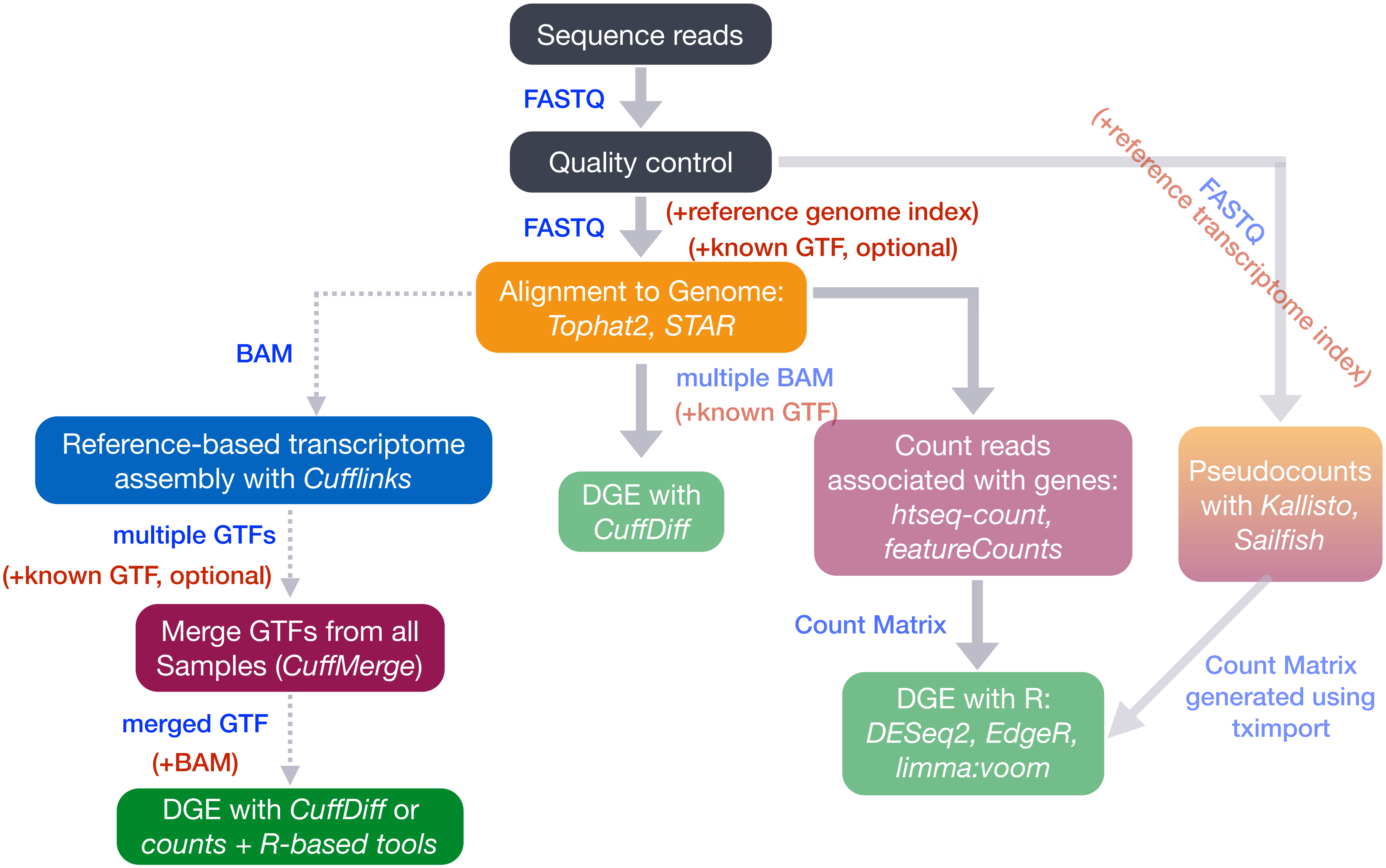
FASTQ

(+reference transcriptome index)

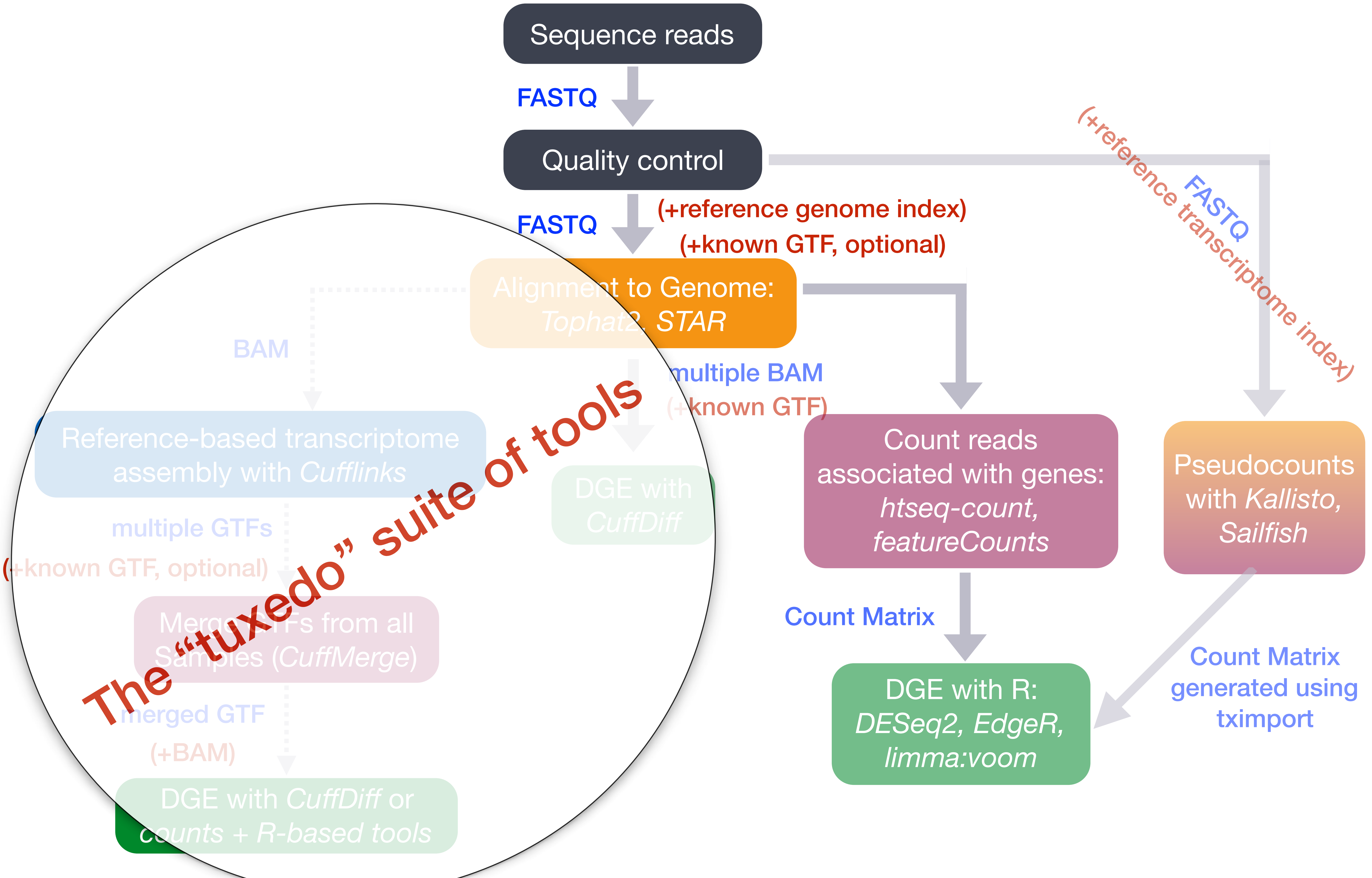
Pseudocounts with *Kallisto, Sailfish*

Count Matrix generated using tximport

Differential Gene Expression Analysis Workflow


Sequence reads
FASTQ
Quality control
FASTQ
(+reference genome index)
(+known GTF, optional)
Alignment to Genome: *Tophat2, STAR*
BAM
Reference-based transcriptome assembly with *Cufflinks*
multiple GTFs
(+known GTF, optional)
Merge GTFs from all Samples (*CuffMerge*)
merged GTF
(+BAM)
DGE with *CuffDiff* or *counts + R-based tools*
multiple BAM
(+known GTF)
DGE with *CuffDiff*
Count reads associated with genes: *htseq-count, featureCounts*
Count Matrix
DGE with R: *DESeq2, EdgeR, limma:voom*
FASTQ
(+reference transcriptome index)
Pseudocounts with *Kallisto, Sailfish*
Count Matrix generated using tximport

Differential Gene Expression Analysis Workflow



Sequence reads

FASTQ

Quality control

FASTQ

(+reference genome index)

(+known GTF, optional)

Alignment to Genome: *Tophat2, STAR*

BAM

Reference-based transcriptome assembly with *Cufflinks*

multiple GTFs

(+known GTF, optional)

Merge GTFs from all Samples (*CuffMerge*)

merged GTF

(+BAM)

DGE with *CuffDiff or counts + R-based tools*

multiple BAM

(+known GTF)

DGE with *CuffDiff*

Count reads associated with genes: *htseq-count, featureCounts*

Count Matrix

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

FASTQ

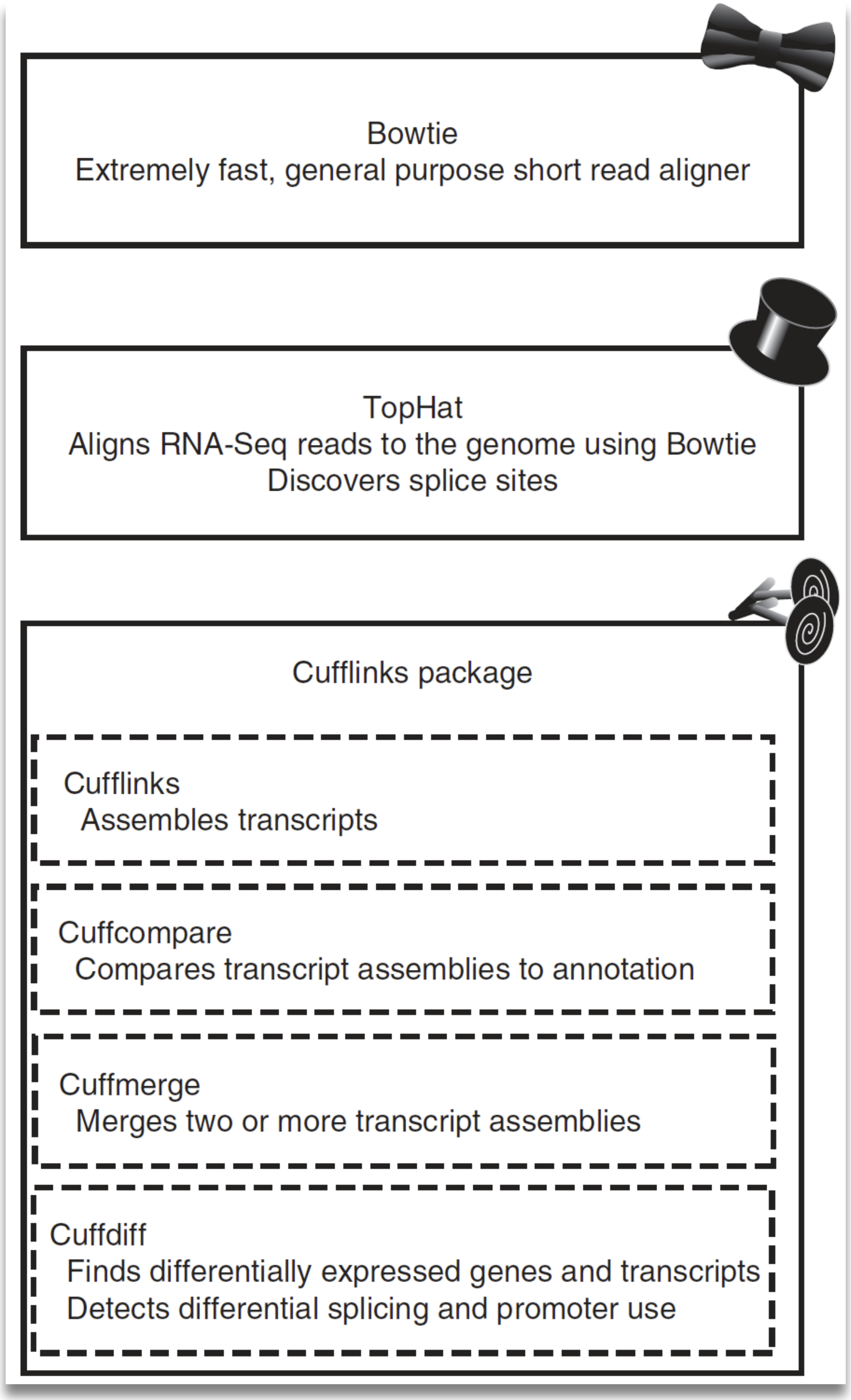
(+reference transcriptome index)

Pseudocounts with *Kallisto, Sailfish*

Count Matrix generated using tximport

The “tuxedo” suite of tools

Differential Gene Expression Analysis Workflow



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

Bowtie/Bowtie2 both use Burrows-Wheeler indexing for aligning reads.

Tophat uses either Bowtie or Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

The Cufflinks package has 4 components, the 2 major ones are listed below -

Cufflinks does **reference-based transcriptome assembly**

Cuffdiff does statistical analysis and identifies differentially expressed transcripts in a **simple pairwise comparison**, and a series of pairwise comparisons in a time-course experiment

Trapnell et al., Nature Protocols, March 2012

When is it appropriate to use Cufflinks/CuffDiff?

You are looking to identify as yet unknown genes/isoforms

Your experiment is a pairwise comparison

You are looking for isoform-level differential expression

Cons:

Complex experimental designs are not supported

False positives are rampant with novel discovery methods

When is it appropriate to use R-based analysis methods using raw counts?

For almost all experiments where you are looking to identify differentially expressed genes (known genes), irrespective of the complexity of the experimental design

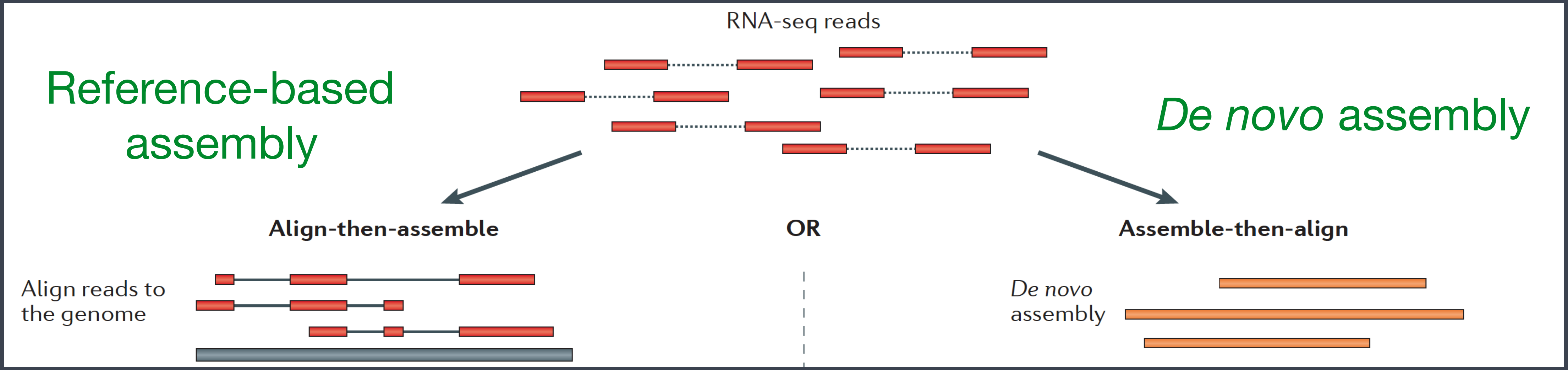
Cons:

A basic knowledge of R programming is required

Not useful for isoform-specific differential expression analysis

Undercounting due to discarded multi-mappers

Alternative methods: transcriptome assembly



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

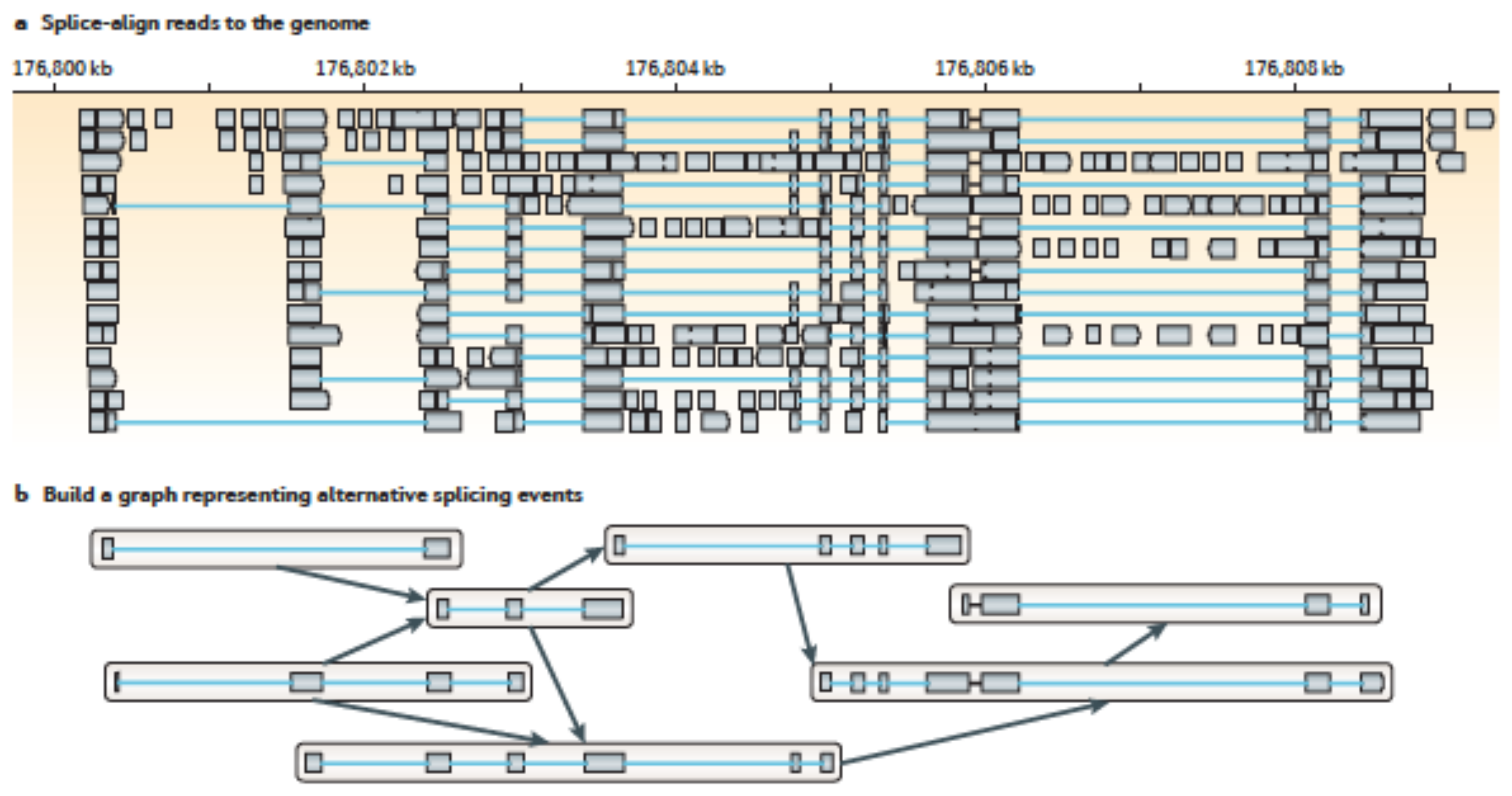
Reference-based assembly

This type of assembly is used when the genome sequence is known.

- ◇ Transcriptome data are not available
- ◇ Transcriptome information available is not good enough, i.e. missing isoforms of genes, or unknown non-coding regions
- ◇ The existing transcriptome information is for a different tissue type or state
- ◇ Cufflinks and Scripture are two reference-based transcriptome assemblers
- ◇ Annotation of any newly-discovered genes or isoforms will be performed downstream

Alternative methods: transcriptome assembly

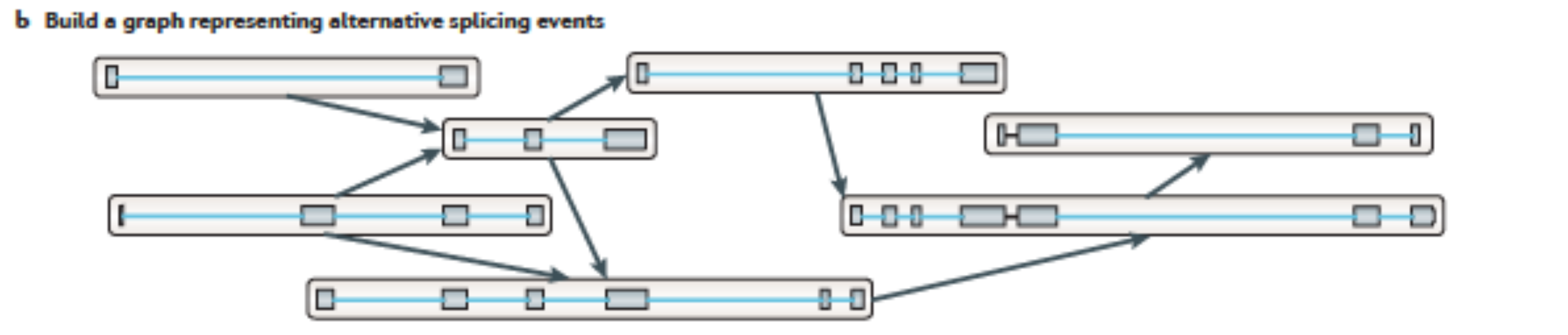
Reference-based assembly


a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

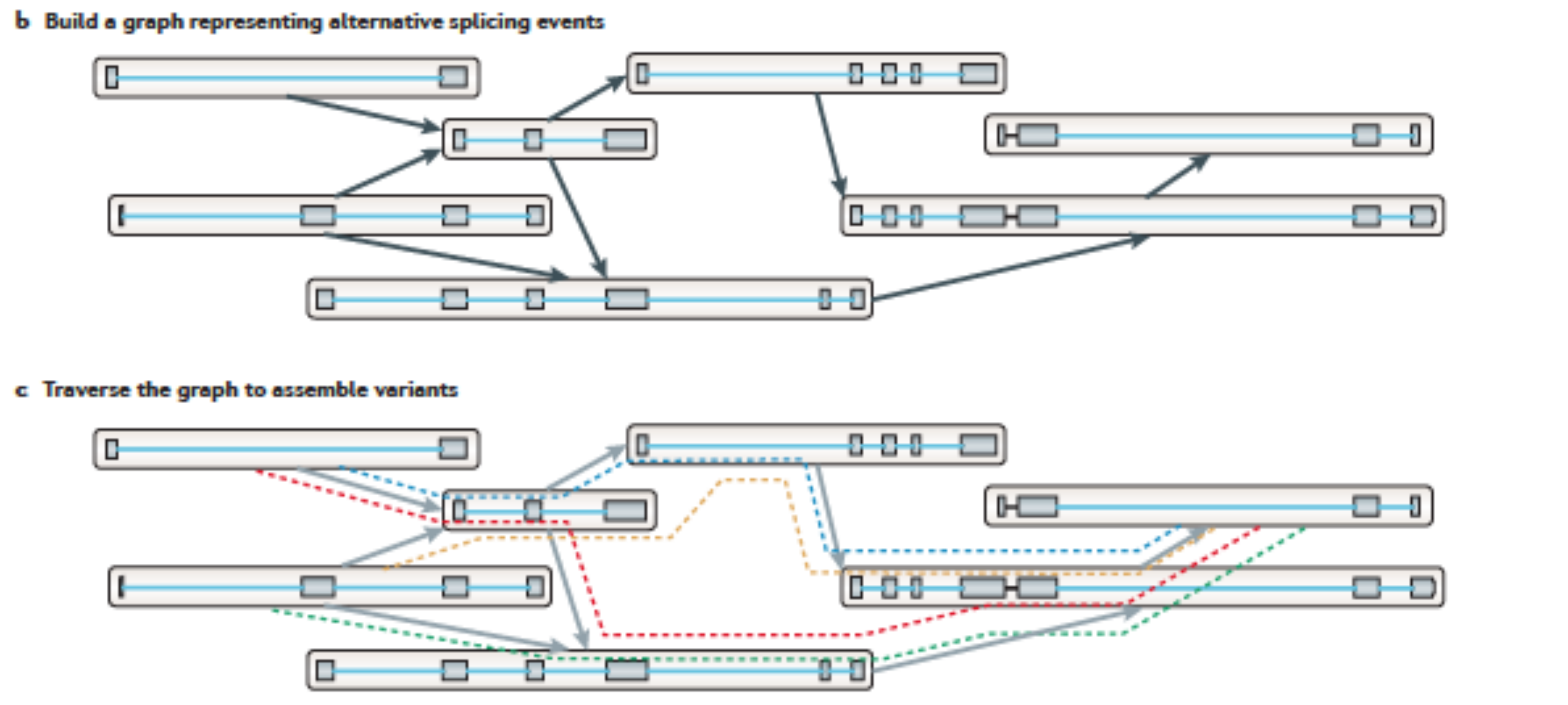
Reference-based assembly



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

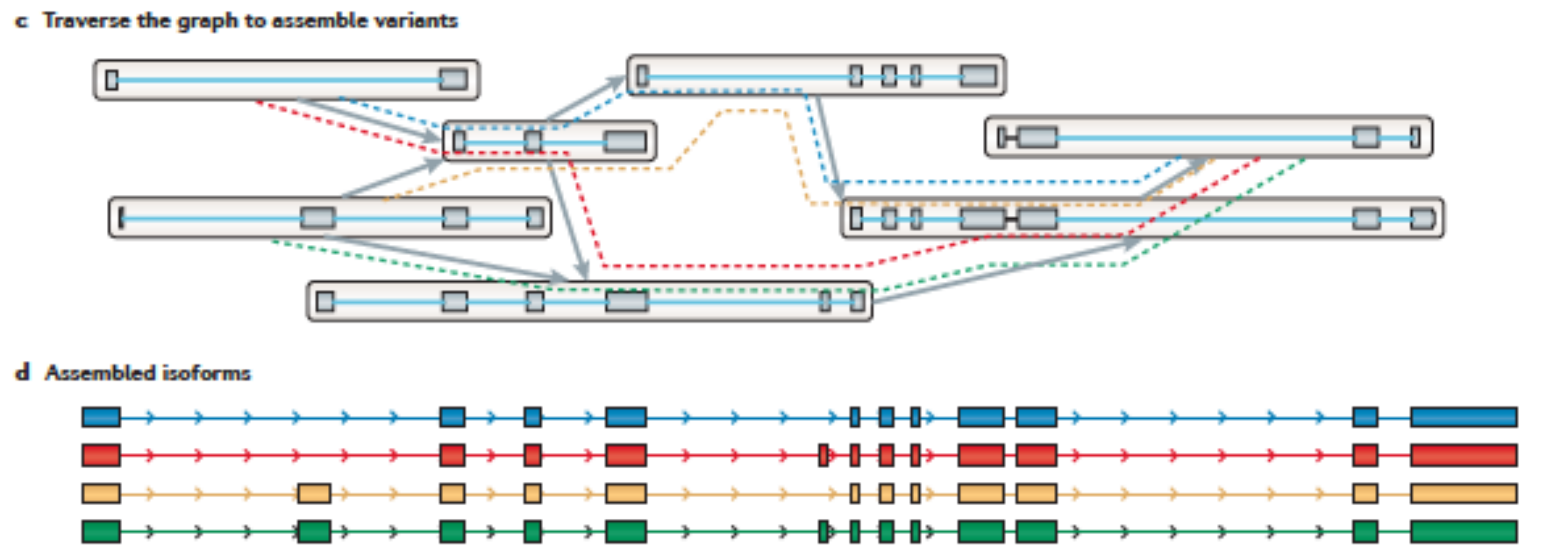
Reference-based assembly


b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants

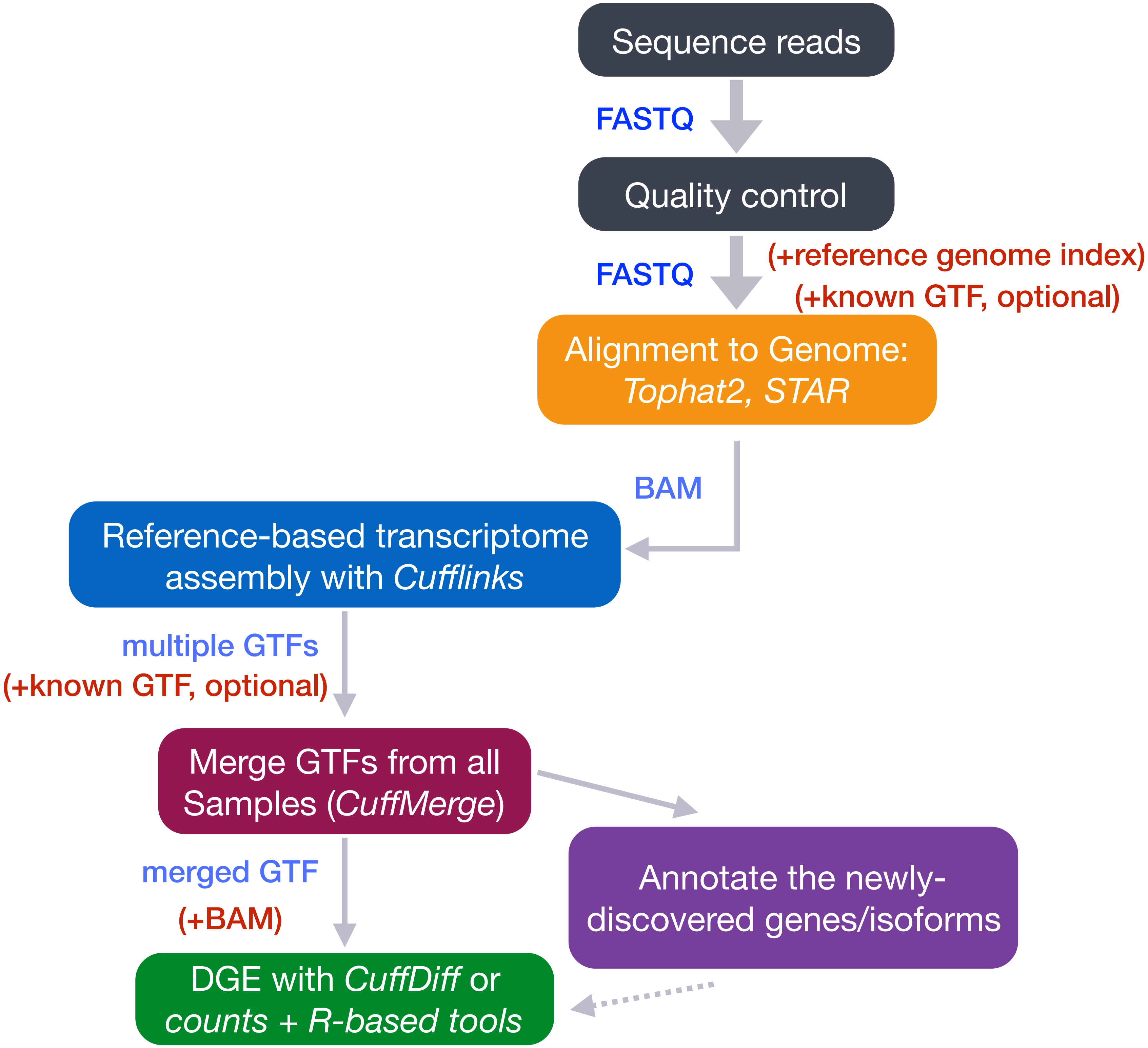
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

Reference-based assembly



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682


Sequence reads
FASTQ
Quality control
FASTQ
(+reference genome index)
(+known GTF, optional)
Alignment to Genome: *Tophat2, STAR*
BAM
Reference-based transcriptome assembly with *Cufflinks*
multiple GTFs
(+known GTF, optional)
Merge GTFs from all Samples (*CuffMerge*)
Annotate the newly-discovered genes/isoforms
merged GTF
(+BAM)
DGE with *CuffDiff* or *counts + R-based tools*

Differential Gene Expression Analysis Workflow

Alternative methods: transcriptome assembly

De novo assembly

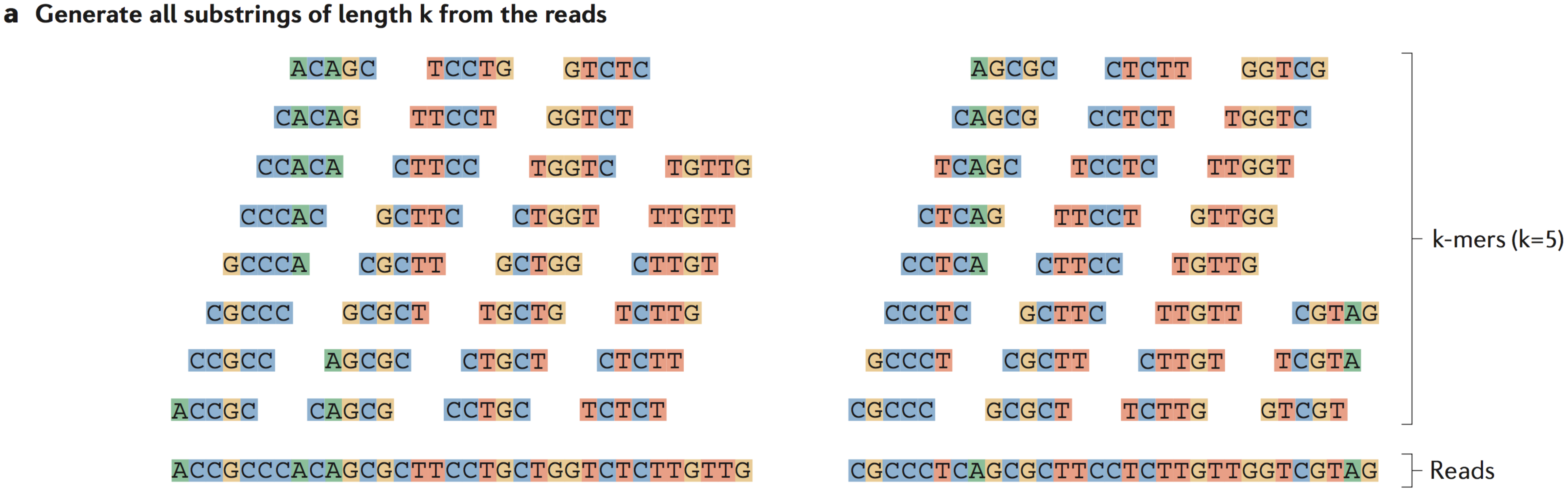
This type of assembly is used when very little information is available for the genome

- An assembly of this type is often the first step in putting together information about an unknown genome
- Amount of data needed for a good de novo assembly is higher than what is needed for a reference-based assembly
- Assemblies of this sort can be used for genome annotation, once the genome is assembled
- Oases, TransABySS, Trinity are examples of well-regarded transcriptome assemblers, especially Trinity
- Annotation of any newly-discovered genes or isoforms will be performed downstream

It is not uncommon to use both methods and compare and combine the assemblies, even when a genome sequence is known, especially for a new genome.

Alternative methods: transcriptome assembly

De novo assembly

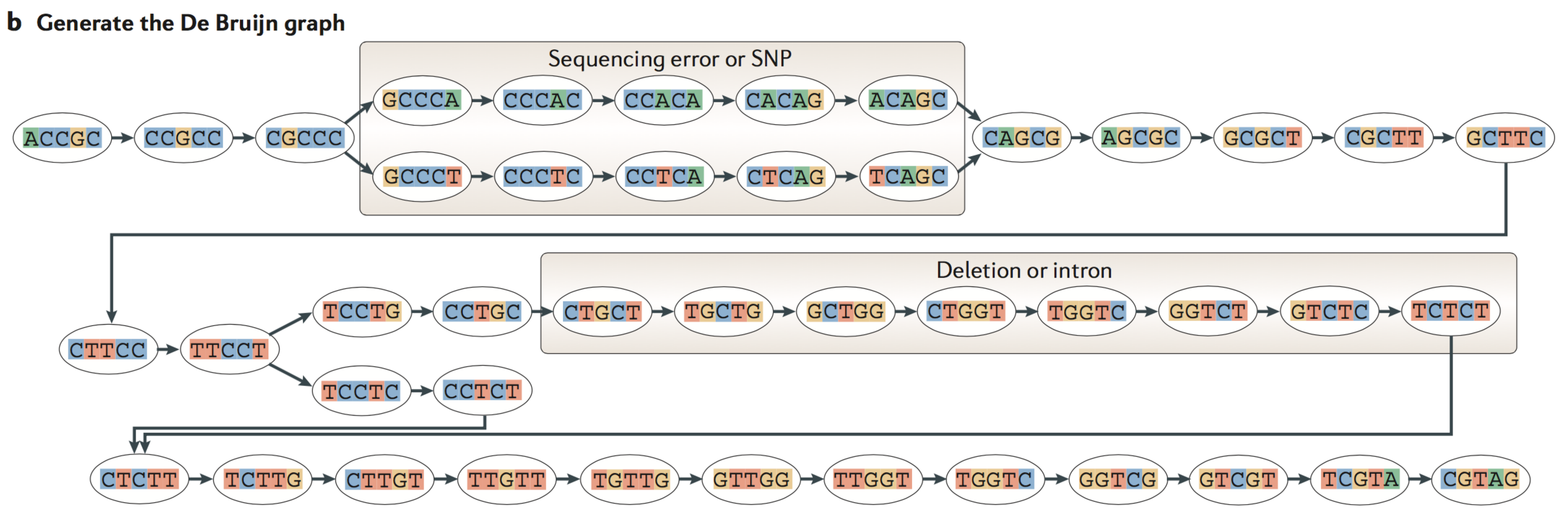


a Generate all substrings of length k from the reads

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

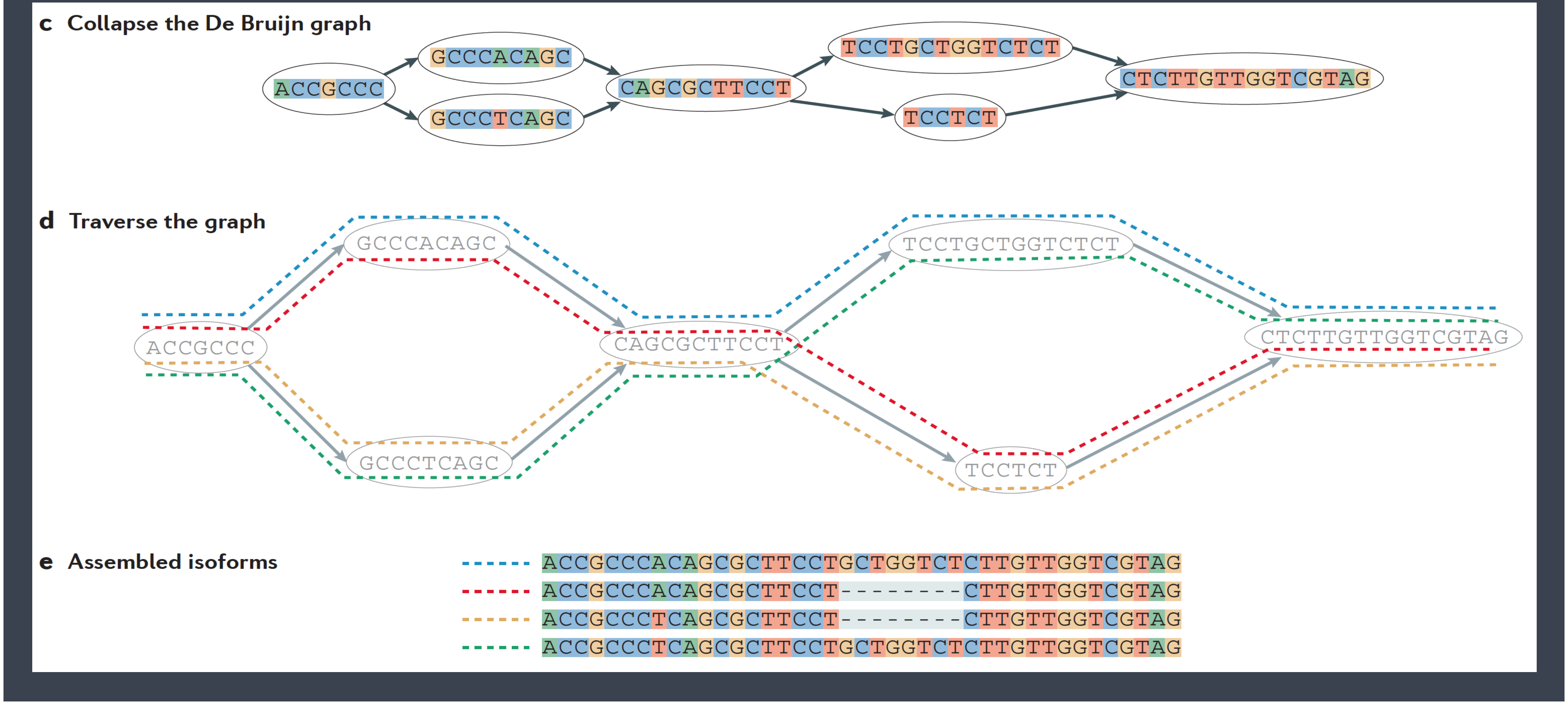
De novo assembly (De Bruijn graph construction)



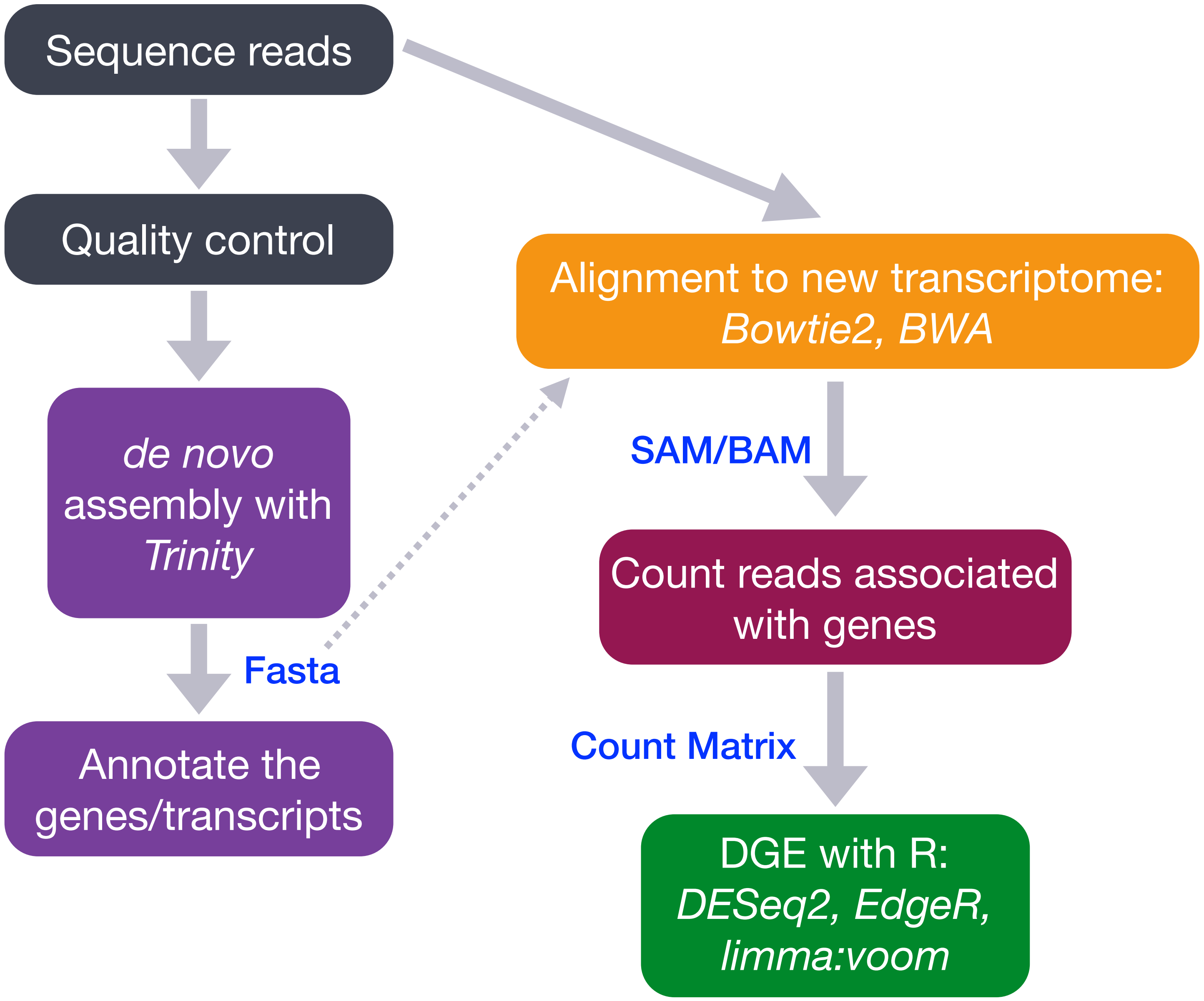
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Alternative methods: transcriptome assembly

De novo assembly (De Bruijn graph construction)


c Collapse the De Bruijn graph
ACCGCCC
GCCCACAGC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
d Traverse the graph
GCCCACAGC
ACCGCCC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
e Assembled isoforms
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682



Sequence reads

Quality control

de novo assembly with *Trinity*

Fasta

Annotate the genes/transcripts

Alignment to new transcriptome: *Bowtie2, BWA*

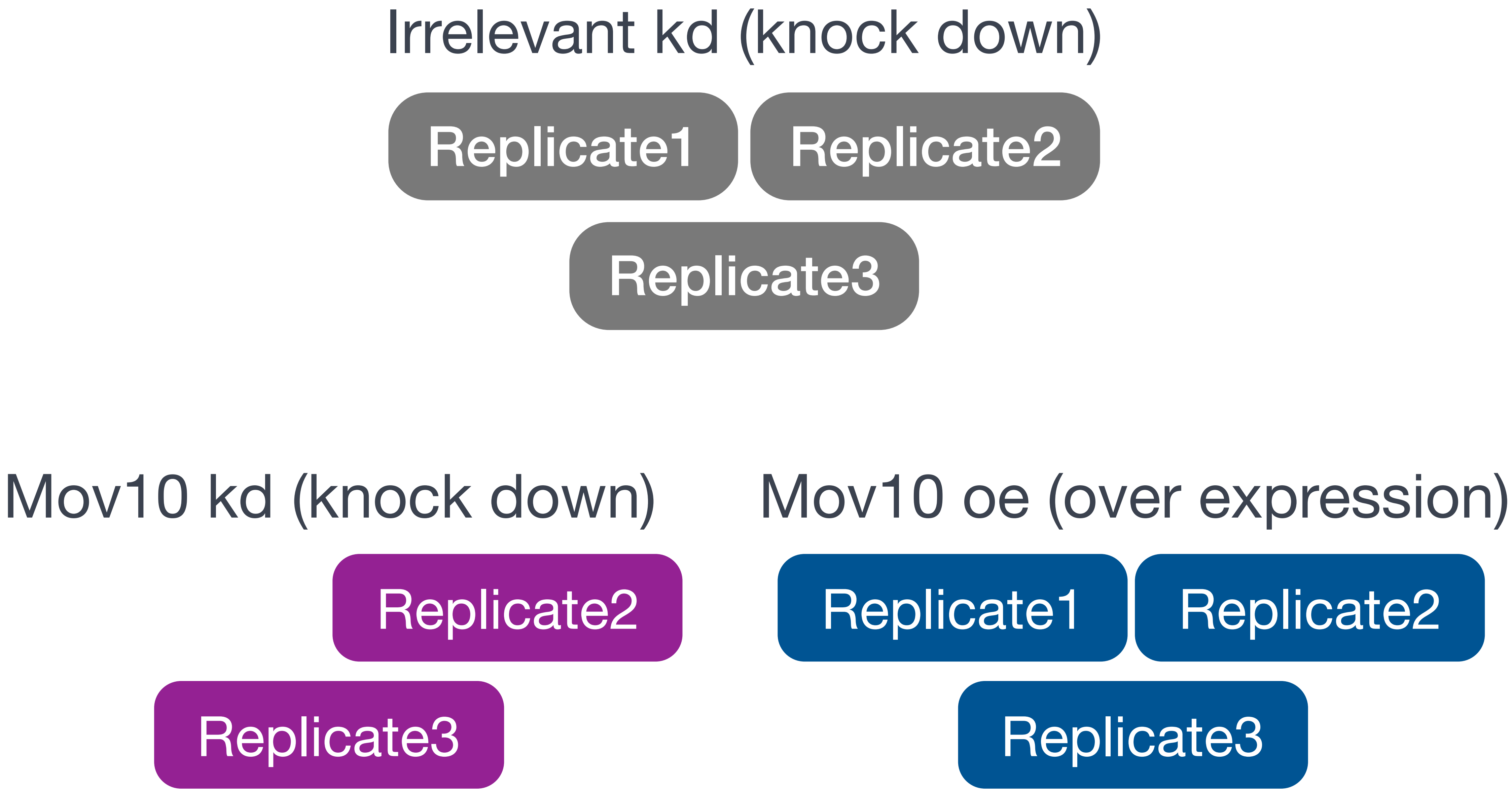
SAM/BAM

Count reads associated with genes

Count Matrix

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

Differential Gene Expression Analysis Workflow



Irrelevant kd (knock down)

Replicate1 Replicate2

Replicate3

Mov10 kd (knock down)

Replicate2

Replicate3

Mov10 oe (over expression)

Replicate1 Replicate2

Replicate3

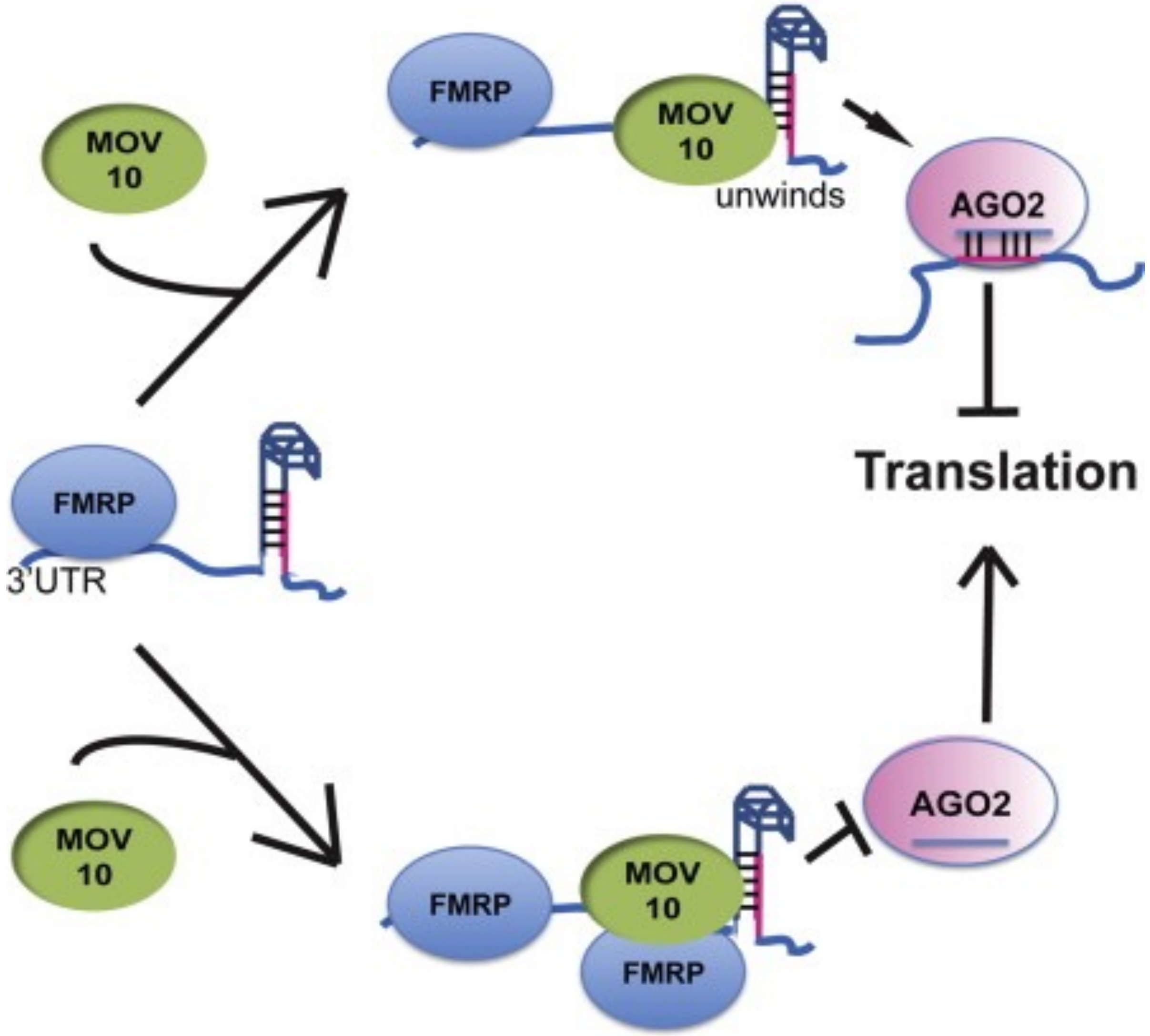
Dataset for Class

MOV10, is a putative RNA helicase that is also associated with FMRP in the context of the microRNA pathway. FMRP and **MOV10** associate and regulate the translation of a subset of RNAs

FMRP is “most commonly found in the brain, is essential for normal cognitive development and female reproductive function. Mutations of this gene can lead to fragile X syndrome, mental retardation, premature ovarian failure, autism, Parkinson's disease, developmental delays and other cognitive deficits.” - https://en.wikipedia.org/wiki/FMR1

Our questions:

- What patterns of expression can we identify with the loss/gain of MOV10?
- Are there any genes shared between the two conditions?



Biological Question

Kenny, P.J. et al, *Cell Reports,* 11 December 2014