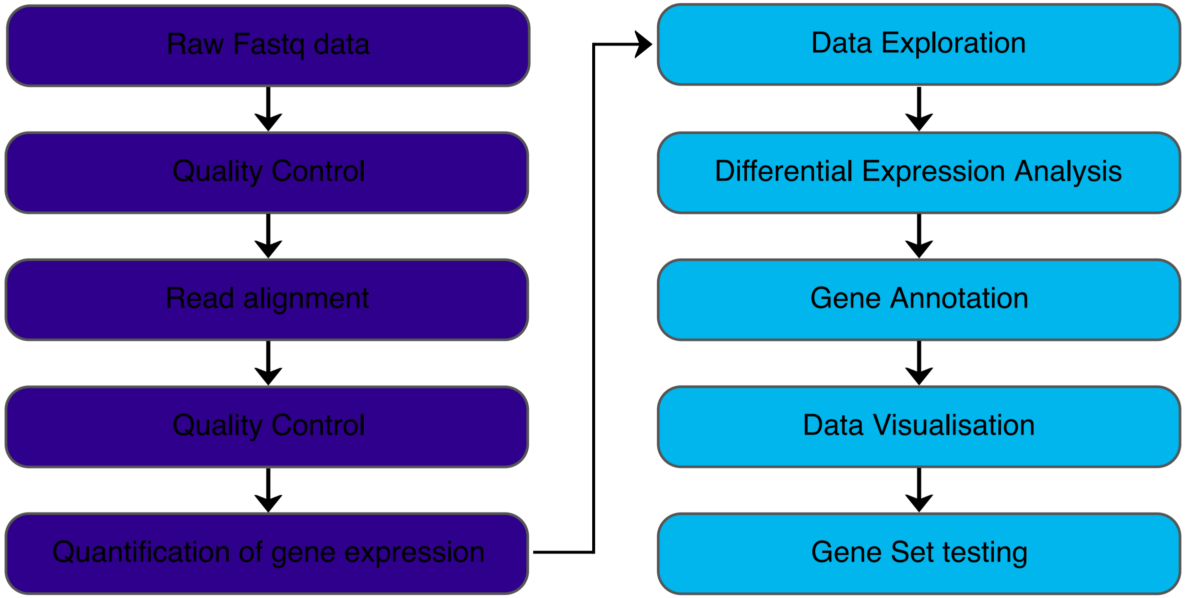
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Quantification of Gene Expression with Salmon

March 2021

Differential Gene Expression Analysis Workflow

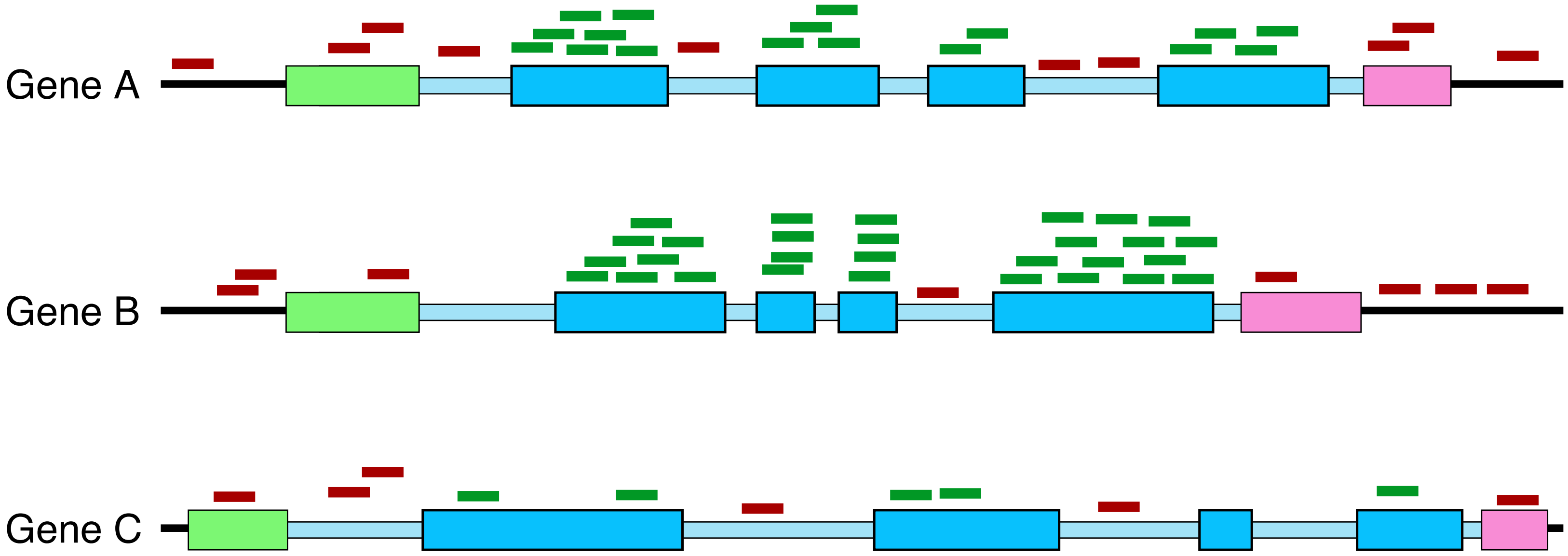


A Simple Counting Approach

We now have the locations of our reads on the genome.

We also know the locations of exons of genes on the genome.

So the simplest approach is to count how many reads overlap each gene.

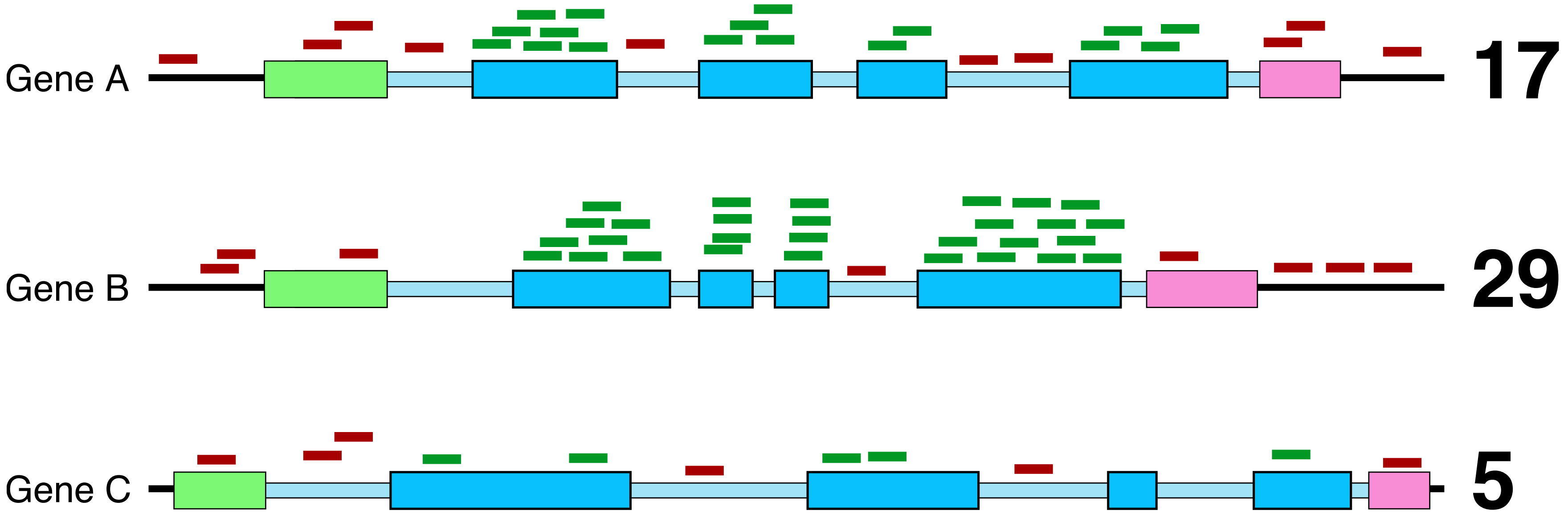


A Simple Counting Approach

We now have the locations of our reads on the genome.

We also know the locations of exons of genes on the genome.

So the simplest approach is to count how many reads overlap each gene.



e.g. featureCounts or HTSeq

Problems with the Simple Counting Approach

- Genes have multiple transcripts, alternative splicing introduces ambiguity
- Traditional alignment is (relatively) slow and computationally intensive
- Read sampling is not uniform, there are biases

Problems with the Simple Counting Approach

- Genes have multiple transcripts, alternative splicing introduces ambiguity
- Traditional alignment is (relatively) slow and computationally intensive
- Read sampling is not uniform, there are biases

More sophisticated approaches:

- CuffLinks - Trapnell *et al.* (2010) Nature Biotechnology doi:10.1038/nbt.1621
- RSEM - Li and Dewey (2011) BMC Bioinformatics doi:10.1186/1471-2105-12-323
- Sailfish - Patro *et al.* (2014) Nature Biotechnology doi:10.1038/nbt.2862
- Kallisto - Bary *et al.* (2016) Nature Biotechnology doi:10.1038/nbt.3519
- **Salmon** - Patro *et al.* (2017) Nature Methods doi:10.1038/nmeth.4197

Problems with the Simple Counting Approach

- Genes have multiple transcripts, alternative splicing introduces ambiguity

Count against the transcriptome instead.

Summarise to gene level for differential gene expression analysis.

Quasi-mapping/Pseudo-alignment

- Traditional alignment is (relatively) slow and computationally intensive

Switch to *quasi-mapping* or *pseudo-alignment* to transcriptome

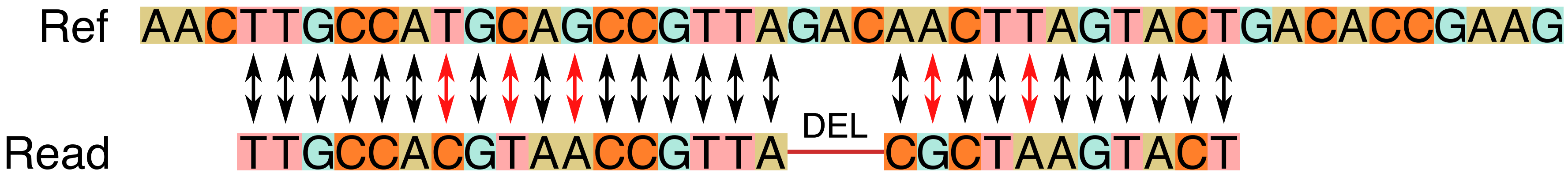
Ref AACTTGCCATGCAGCCGTTAGACAACTTAGTACTGACACCGAAG

Read TTGCCACGTAACCGTTACGCTAAGTACT

Quasi-mapping/Pseudo-alignment

- Traditional alignment is (relatively) slow and computationally intensive

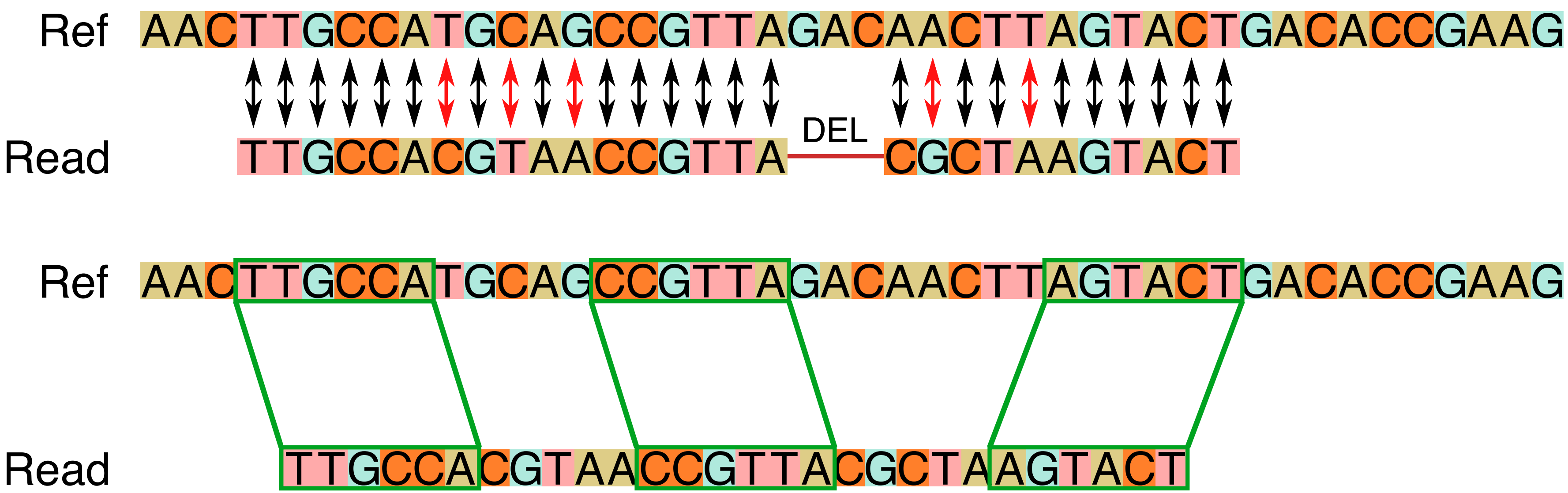
Switch to *quasi-mapping* or *pseudo-alignment*



Quasi-mapping/Pseudo-alignment

- Traditional alignment is (relatively) slow and computationally intensive

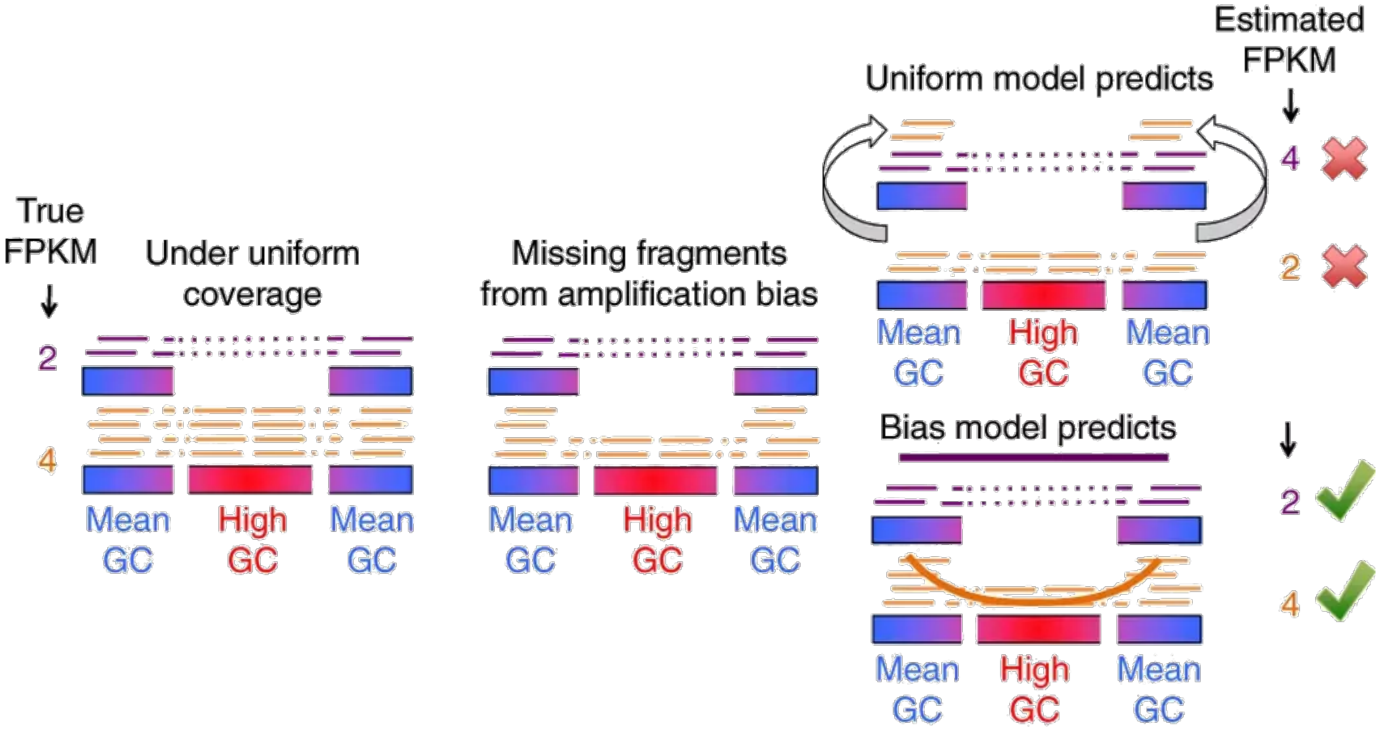
Switch to *quasi-mapping* or *pseudo-alignment*


Ref
AACTTGCCATGCAGCCGTTAGACAACTTAGTACTGACACCGAAG
DEL
Read
TTGCCACGTAACCGTTA
CGCTAAGTACT
Ref
AACTTGCCATGCAGCCGTTAGACAACTTAGTACTGACACCGAAG
Read
TTGCCACGTAACCGTTACGCTAAGTACT

Bias models

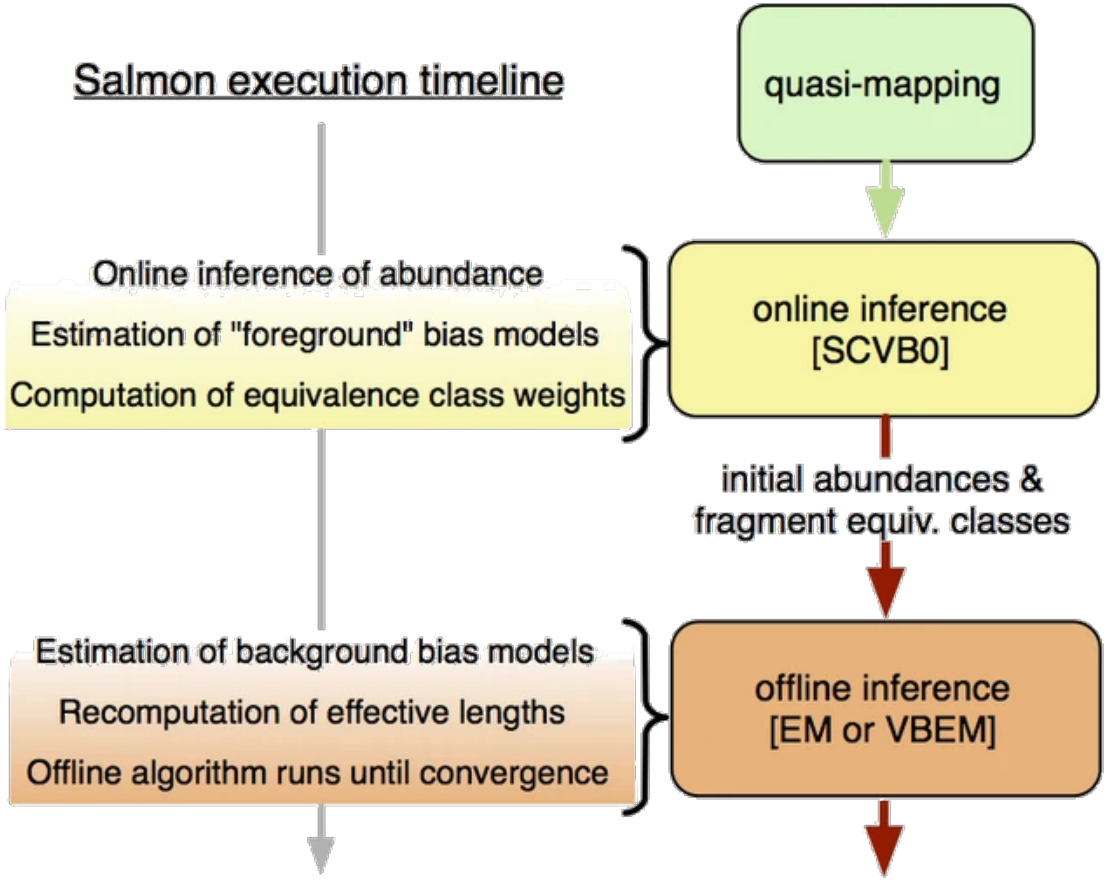
- Read sampling is not uniform, there are biases

Include modelling for GC bias, positional bias and sequence bias in the quantification algorithm



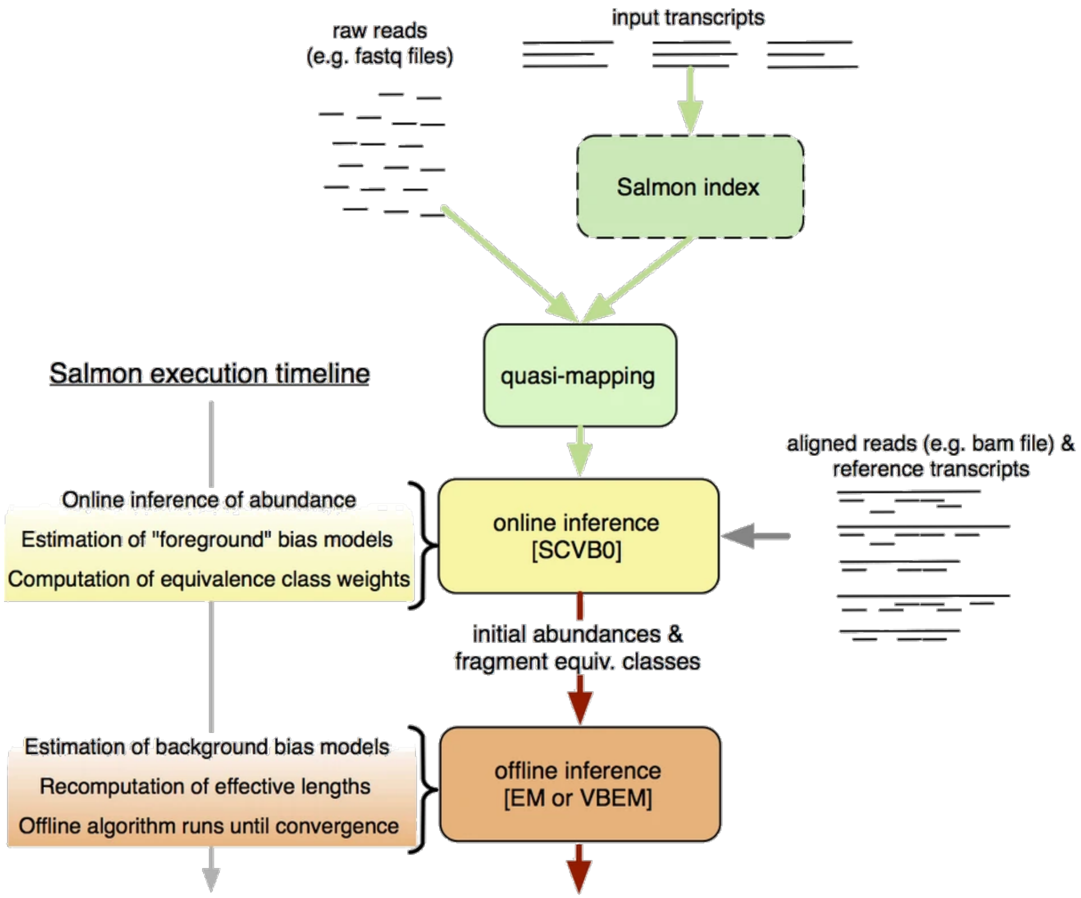
Love *et al.* (2016) Nature Biotechnology doi:10.1038/nbt.3682

Salmon workflow


Salmon execution timeline
quasi-mapping
Online inference of abundance
Estimation of "foreground" bias models
Computation of equivalence class weights
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
Estimation of background bias models
Recomputation of effective lengths
Offline algorithm runs until convergence
offline inference
[EM or VBEM]

Patro *et al.* (2017) Nature Methods doi:10.1038/nmeth.4197

Salmon workflow


raw reads
(e.g. fastq files)
input transcripts
Salmon index
quasi-mapping
Salmon execution timeline
aligned reads (e.g. bam file) &
reference transcripts
Online inference of abundance
Estimation of "foreground" bias models
Computation of equivalence class weights
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
Estimation of background bias models
Recomputation of effective lengths
Offline algorithm runs until convergence
offline inference
[EM or VBEM]

Patro *et al.* (2017) Nature Methods doi:10.1038/nmeth.4197

Practical

1. Create and index to the transcriptome with Salmon
2. Quantify transcript expression using Salmon