Quantification of Gene Expression with Salmon

March 2021
Differential Gene Expression Analysis Workflow

- Raw Fastq data
- Quality Control
- Read alignment
- Quality Control
- Quantification of gene expression
- Data Exploration
- Differential Expression Analysis
- Gene Annotation
- Data Visualisation
- Gene Set testing
A Simple Counting Approach

Align to the genome to get the locations of our reads.

Gene annotation references give us 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

Gene A

Gene B

Gene C
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:

- Sailfish - Patro et al. (2014) Nature Biotechnology doi:10.1038/nbt.2862
- Kallisto - Bary et al. (2016) Nature Biotechnology doi:10.1038/nbt.3519
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: AACTTGCCATGCAAGCCGTAGACAACTTAGTACTGACACCCGAAG

Read: TTGCCACGTAAACCGTTACGCTAAGTACT
Quasi-mapping/Pseudo-alignment

- Traditional alignment is (relatively) slow and computationally intensive

Switch to quasi-mapping or pseudo-alignment

Ref  AACTTTGCCATGCAAGCCGTTAGAACAACTTTAGTACTGACACCCGAAG

Read TTGCCACCGTGAAACCCTTADELCGCTAAGTACT
Quasi-mapping/Pseudo-alignment

- Traditional alignment is (relatively) slow and computationally intensive

Switch to quasi-mapping or pseudo-alignment
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

Online inference of abundance
- Estimation of "foreground" bias models
- Computation of equivalence class weights

 quasi-mapping

online inference [SCVB0]
- Initial abundances & fragment equiv. classes

offline inference [EM or VBEM]
- Estimation of background bias models
- Recomputation of effective lengths
- Offline algorithm runs until convergence

Salmon workflow

Practical

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