Alignment and Quantification of Gene Expression with Salmon

March 2023

Differential Gene Expression Analysis Workflow

Traditional Alignment

AIM: Given a reference sequence and a set of short reads, align each read to the reference sequence finding the most likely origin of the read sequence.

Alignment - Gap aware alignment

Aligners: STAR, HISAT2

Why Pseudo-alignment methods are fatster?

- Unlike alignment based methods, pseudo-alignment methods focus on transcriptome (~2% of genome)
- No base to base alignment required in pseudo-alignment methods

Quasi-mapping/Pseudo-alignment

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

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

Quasi-mapping/Pseudo-alignment

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

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*

What is read quantification?

- **Quantification**: How many reads have come from a genomic feature?
  - genomic feature can be gene or transcript or exon, but usually gene

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.

What is read quantification?

- **Quantification**: How many reads have come from a genomic feature?
  - genomic feature can be gene or transcript or exon, but usually gene

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.

Quantification tools

- Broadly classified into two types . . .
  - Alignment based:
    - Takes bam file as input,therefore reads must be mapped prior to quantification
    - quantifies using simple counting procedure
    - Pros: Intuitive
    - Cons: Slow and can not correct biases in RNAseq data
    - Tools: HTseq, SubRead etc.
  - Alignment-free:
    - Also called quasi-mapping or pseudoalignment
    - Starts from fastq files and base-to-base alignment of the reads is avoided
    - Pros: Very fast and removes biases
    - Cons: Not intuitive
    - Tools: Kallisto, Sailfish, **Salmon** etc

RNA-seq data biases

- **GC bias**: Higher GC content sequences are less likely to be observed as PCR is not efficient with high GC content sequences.
- **Positional bias**: for most sequencing methods, the 3 prime end of transcripts are more likely to be observed.
- **Complexity bias**: some sequences are easier to be bound and amplified than others.
- **Sequence-based bias**: Bias in read start positions arising from the differential binding efficiency of random hexamer primers
- **Fragment length bias**: Induced by size selection
- Above biases are sample specific
- Methods like Salmon attempt to mitigate the effect of technical biases by estimating sample-specific bias parameters.

Salmon workflow

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

Salmon workflow

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