Alignment and Quantification of Gene Expression with Salmon

October 2024

Differential Gene Expression Analysis Workflow

Alignment and Quantification overview

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 - Splicing aware alignment

Aligners: STAR, HISAT2

Alignment

- Traditional alignment perform base-by-base alignment
- It is (relatively) slow and computationally intensive

Alignment

- Traditional alignment perform base-by-base alignment
- Traditional alignment is (relatively) slow and computationally intensive

Alignment

- Traditional alignment perform base-by-base alignment
- Traditional alignment is (relatively) slow and computationally intensive

BAM/SAM file format

Sequence **A**lignment/**M**ap (SAM) format is the standard format for files containing aligned reads.

Definition of the format is available at https://samtools.github.io/hts-specs/SAMv1.pdf.

Two main parts:

- Header
  - contains meta data (source of the reads, reference genome, aligner, etc.)
  - header lines start with "`@`"
  - header fields have standardized two-letter codes
  - `@RG` for read group, used for merging BAMs together
- Alignment section
  - 1 line for each alignment
  - contains details of alignment position, mapping, base quality etc.
  - 11 required fields, but other content may vary depending on aligner and other tools used to create the file
- BAM is a binary version of SAM (not human readable)

BAM/SAM format - What does it look like?

Why are Pseudo-alignment methods faster?

Switch to *quasi-mapping* (Salmon) or *pseudo-alignment* (Kallisto)

- These tools avoids base-to-base alignment of the reads
- $\sim$ 20 times faster than the traditional alignment tools like STAR, HISAT2 etc
- Unlike alignment based methods, pseudo-alignment methods focus on transcriptome (∼2% of genome in human)
- Use exact kmer matching rather than aligning whole reads with mismatches and indels

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: RSEM (accounts for isoforms), 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, **Salmon** etc

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

If we had mapped our reads to the genome (rather than the transcript sequences), our mapping would look like this:

We also know the locations of exons of genes on the genome, from an annotation file (e.g. GFF or GTF)

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

What is read quantification?

However, Salmon does not work this way. We have mapped to the transcript sequences, not the genome. Quantification is performed as part of the quasi-mapping process.

Salmon also takes account of biases:

- **Multimapping**: Reads which map equally well to multiple locations
- **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
- [illegible]

Salmon

Two essential steps

1. Create transcriptome index
    - This makes downstream quasi-mapping and quantification step efficient and faster
    - Once you create an index, you can use it again and again
    - Consider the Kmer size (default 31 for >75bp)
    - Adding decoy sequences to filter contaminants
2. Quasi-mapping and quantification

Practical

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