RNA sequencing (RNA-Seq) and differential expression analysis using RNA-Seq

Vincent Plagnol

UCL Genetics Institute

Outline

1 Introduction

2 Alignment of RNA-Seq data to a reference genome

Outline

1 Introduction

2 Alignment of RNA-Seq data to a reference genome

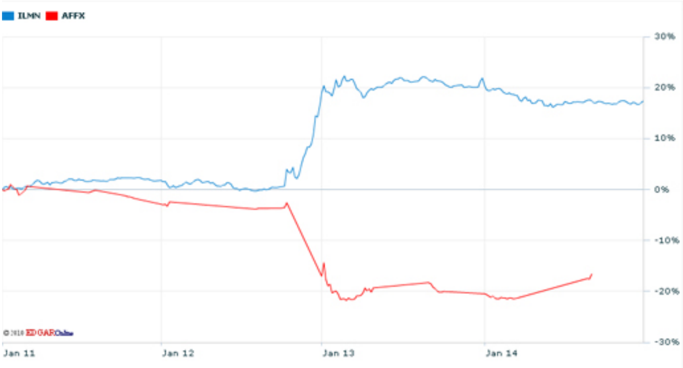
What is RNA-Seq?

- RNA-Seq refers to the possibility of sequencing the mRNA rather than the DNA of an individual.
- There are many ways to achieve this:
  - poly(A) library preparation is common and enriches for mature spliced RNA.
  - A nuclear RNA prep will focus instead on pre-splicing species.
  - Total RNA sequencing is an alternative.
- It is even possible to sequence the RNA of a single cell.
- Protocols are more fiddly and subtle than DNA sequencing.
  - Data quality is often lower, resulting in lower complexity and sometimes shorter reads.
  - The bioinformatics are also much more challenging.

What are we trying to achieve?

- Differential expression analysis.
    - This is your “stock analysis” of microarray data.
    - Compare two conditions, perhaps two genotypes, or two drug treatments, and see what genes are up or down-regulated.
    - This is very similar to a read depth analysis.
    - This analysis typically needs replicates (more on this later).
- Another more quantitative aim is to discover new isoforms and splice variants.
    - This is really taking advantage of the sequencing, and not something microarrays can do.
    - This would be closer to a split read analysis to identify the junctions.

Expression level estimation: the death of arrays?



(from Daniel MacArthur, Genetic Inference)

Tools for alignment of raw reads to the reference

- Tophat: alignment of RNAseq data to a reference genome with the identification of novel splice sites.
    - `tophat` web page
    - TopHat: discovering splice junctions with RNA-Seq
- RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome
    - RSEM web page.
    - RSEM paper
- STAR:
    - The STAR webpage is here.
    - The reference paper.

Additional tools/papers

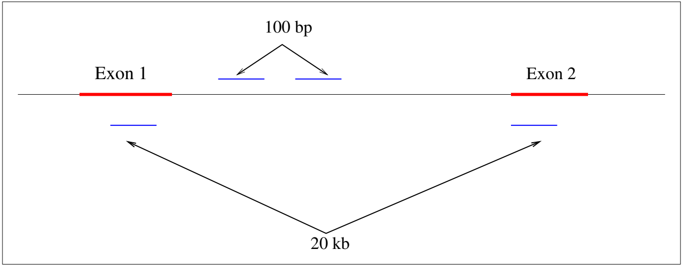
- RSeqQC: set of tools for RNAseq quality control.
  - http://code.google.com/p/rseqc/
  - Formerly known as EverSeq.
- A reference paper: Mapping and quantifying mammalian transcriptomes by RNA-Seq, Mortazavi et al, 2008

Outline

1 Introduction

2 Alignment of RNA-Seq data to a reference genome

The difficulty of mapping PE reads to RNA data

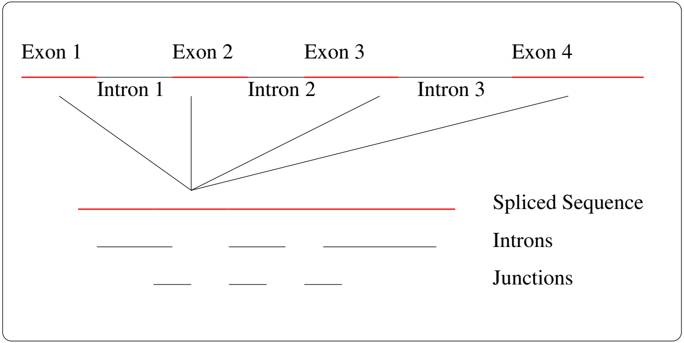


The situation is even worse if reads overlap the exon-exon junctions.

Where it gets worse: isoform specific estimation

- Each gene has multiple isoforms, sometimes quite similar to each other.
- An obvious question is the relative expression level of these isoforms: which is dominant for example?
- This isoform estimation problem can be built into the alignment or done a posteriori.

An (old-fashioned?) option: design a transcriptome reference sequence

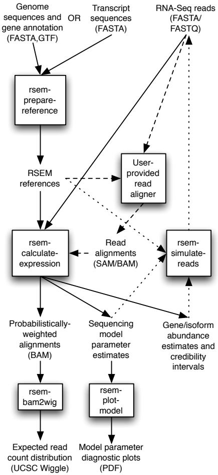


This was done for each annotated transcript and we added a full reference genome with masked genes.
See for example Heap et al, Human Molecular Genetics 2010.

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

- RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data.
- It essentially generates transcripts and maps reads to these multiple isoforms.
    - A side effect is that it will not be useful to discover novel isoforms.
- A procedure is built in to assign reads in a probabilistic manner (EM algorithm).
- It provides posterior mean and 95% credibility interval estimates for expression levels which is a key feature.

The strategy taken by RSEM



Genome sequences and gene annotation (FASTA,GTF)
OR
Transcript sequences (FASTA)
RNA-Seq reads (FASTA/ FASTQ)
rsem-prepare-reference
RSEM references
User-provided read aligner
rsem-calculate-expression
Read alignments (SAM/BAM)
rsem-simulate-reads
Probabilistically-weighted alignments (BAM)
Sequencing model parameter estimates
Gene/isoform abundance estimates and credibility intervals
rsem-bam2wig
rsem-plot-model
Expected read count distribution (UCSC Wiggle)
Model parameter diagnostic plots (PDF)

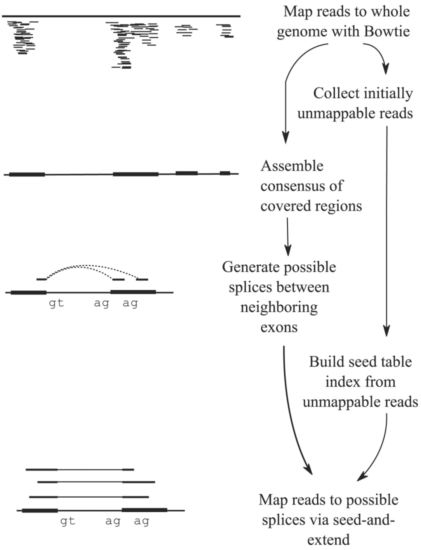
Exercise: What gtf file should I use?

- There are vast differences between the GTF files available to you.
- The NCBI GTF files are enormous and contain almost any exon ever annotated.
- This can be counter-effective for many steps.
- I recommend using the Illumina iGenomes which are useful for RNA analysis.
- Note that I had to process these files quite a bit to use ensembl gene IDs.
    - Feel free to ask me if you need to do the same.

A second option: align to the reference genome but be "transcriptome aware"

- A more straightforward approach is to align to the standard reference genome, but allowing for gaps that are generated by introns.
- Clearly the knowledge of where the introns are is useful and should be factored in.
- Some allowance for novel discoveries is also key, to not miss novel and interesting transcripts.

The tophat workflow



Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

A standard call to tophat

- Note the Bowtie input indexes as well as the gene structure information in gtf format (this contains all the junctions).
- Segment-length tries to split the reads into smaller chunks that we attempt to align around junctions.
- -r specifies the expected distance between mate-paired reads (220 seems high).
- -p 1 specifies a single processor for Bowtie.

```
tophat --no-coverage-search -o Ctl1 -p 1
--segment-length 20 -r 220 --library-type
fr-unstranded -G
Mus_musculus/NCBI/build37.2/Annotation/Genes/genes.gtf
Mus_musculus/NCBI/build37.2/Sequence/Bowtie2Index/genome
Ctl1_p1.fastQ Ctl1_p2.fastQ
```

How does `tophat` find junctions?

- `Tophat` generates its database of possible splice junctions from two sources of evidence.
- The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons.
  - With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio.
- The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping.
- Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron.
- We only suggest users use this second option (–coverage-search) for short reads (< 45bp) and with a small number of reads (< 10 million).