Analysis of transcriptional variation from RNA Sequence Data

Joseph Powell

Computational and Single Cell Genomics, Institute for Molecular Bioscience

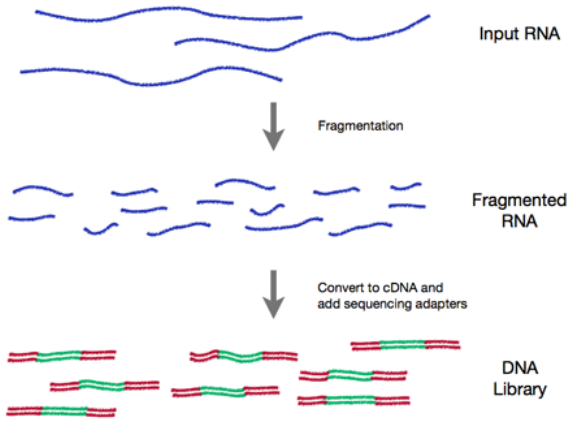
---

Summer Institute in Statistical Genetics

Brisbane, 13th Feb 2016

Applications and information content

RNA-sequencing (RNA-seq) has a wide variety of applications
The power of sequencing RNA lies in the fact that the twin aspects of discovery and quantification
Many variations of RNA-seq protocols and analyses have been published, making it challenging for new users to appreciate



RNA-extraction protocol

What species of RNA are you interested in quantifying?

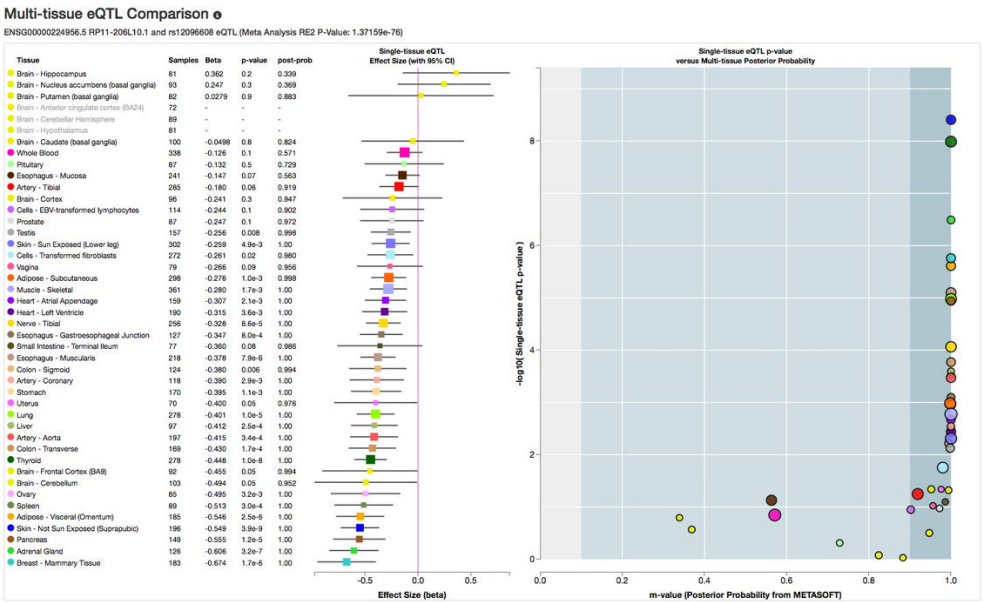
- Messenger RNA (mRNA) - accounts for just 5% of the total RNA in the cell. mRNA is the most heterogeneous types of RNA regarding both base sequence and size.
- A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein.
  - Ribosomal RNA (rRNA) are found in the ribosomes and account for 80% of the total RNA present in the cell
  - Transfer RNA (tRNA) are an essential component of translation, where their main function is the transfer of amino acids during protein synthesis
  - Long non-coding RNAs (long ncRNAs, lncRNA) are defined as non-protein coding transcripts longer than 200 nucleotides.

What tissue am I sampling from?

The choice of tissue/sample is critical in the context of the biological question.
Why?

What tissue am I sampling from?

The choice of tissue / sample is critical in the context of the biological question.
Why?

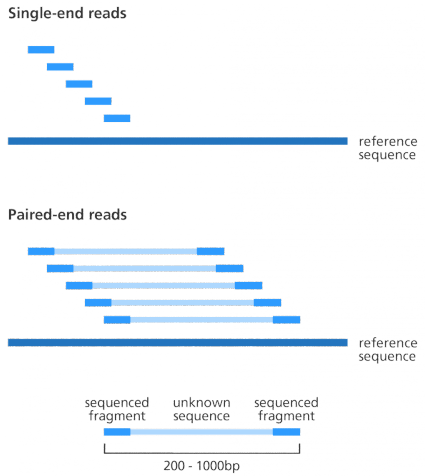


Resources: GTeX, TiGER, Encode, Fantom

Single vs Paired end reads

Sequencing can involve single-end (SE) or paired-end (PE) reads. PE is preferable for;

- de novo transcript discovery or
- isoform or splice expression analysis
- allele specific expression



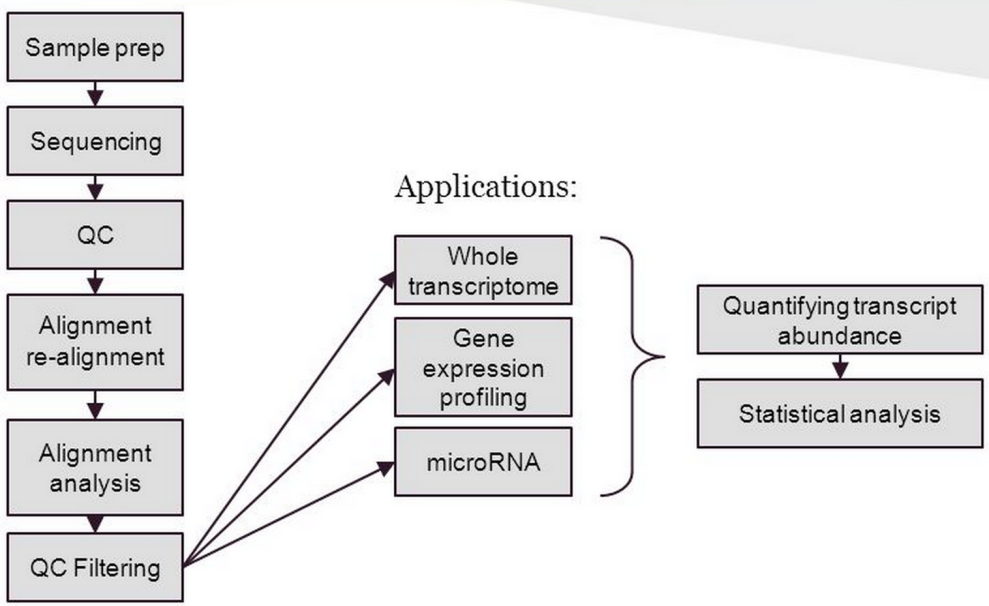
It comes at a price - literally!

Data

Consideration: An important factor is sequencing depth or library size, which is the number of sequenced reads for a given sample.
More transcripts will be detected, and their quantification will be more precise as the sample is sequenced to a deeper level. Nevertheless, optimal sequencing depth again depends on the aims of the experiment.

What Transcriptional phenotypes can we measure?

- Transcript counts (at various levels)
- Alternative splicing events isoform variation
- Novel transcripts



Counting

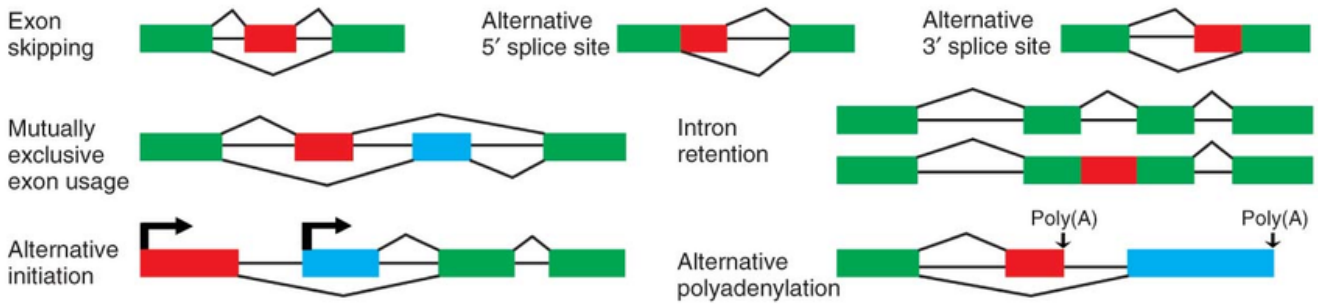
The most common application of RNA-seq is to estimate gene and transcript expression. This application is primarily based on the number of reads that map to each transcript sequence, although there are algorithms such as Sailfish that rely on k-mer counting in reads without the need for mapping

Wiki: list of RNA-Seq bioinformatics tools

Alternative splicing analysis

Alternative splicing

Transcript-level differential expression analysis can detect changes in the expression of transcript isoforms from the same gene.



Detection methods fall into two major categories.

plice variation

Isoform-based

Integration of isoform expression estimation with the detection of differential expression to reveal changes in the proportion of each isoform within the total gene expression

Tools: CuffDiff2, flow difference metric (FDM), rSeqDiff

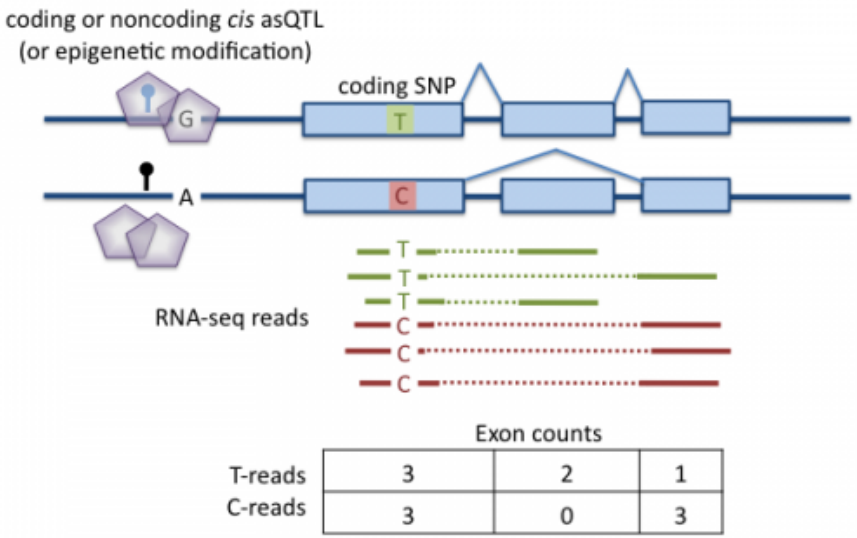
Exon-based

Exon-based approach skips the estimation of isoform expression and detects signals of alternative splicing by comparing the distributions of reads on exons and junctions of the genes between the compared samples

Tools: DEXseq, DSGSeq, rMATS, DiffSplice

Genetic control of these events

Splice QTL - loci influencing splice variation



Exon counts

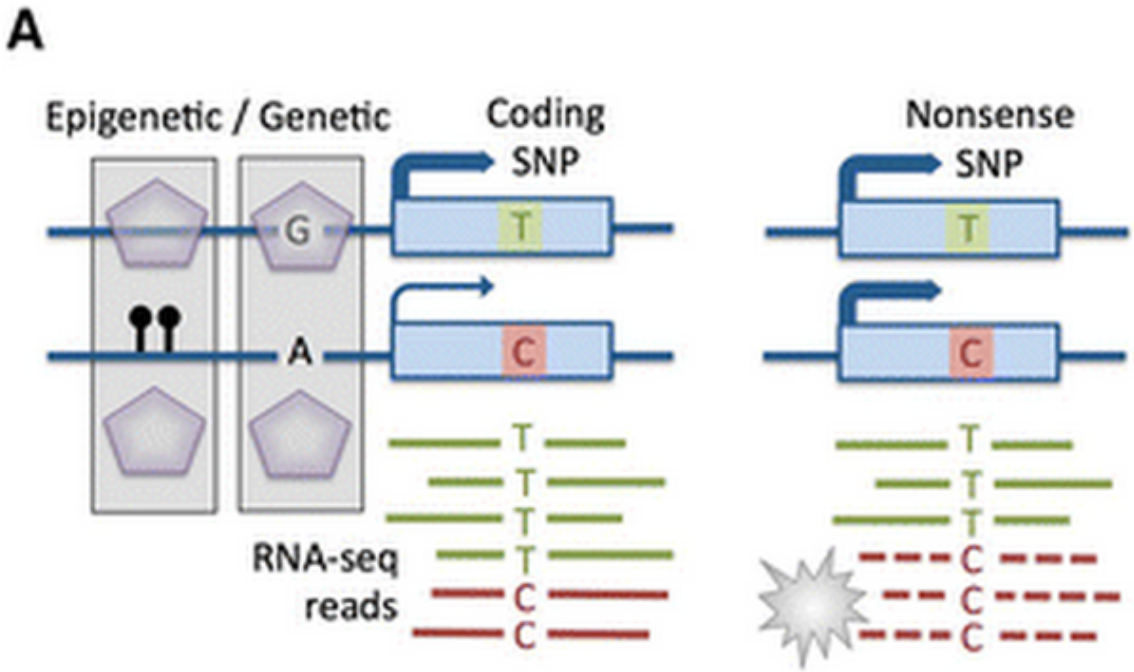
T-reads	3	2	1
C-reads	3	0	3

Alternative splicing and splice QTLs (sQTL) will be mapped using sQTLseekeR, which identifies SNPs that are associated with changes in the relative abundance of gene transcript isoforms. sQTLseekeR fits a multivariate linear model that can be extended to include additional fixed or random effects

Unit of allelic expression

ASE

Allele Specific Expression (ASE) is variation in the transcript abundance of two haplotypes of a individual distinguished by heterozygous sites



Causes of allele-specific expression

1. Epigenetic phenomena, such as imprinting (when an inherited allele from one parent is consistently overexpressed)
2. Allele-specific chromatin modifications.
3. Alternatively, DNA sequence variants in the promoter or within the transcribed region of a gene can affect the rate of transcription or the rate of decay of the transcript, respectively. How do we test

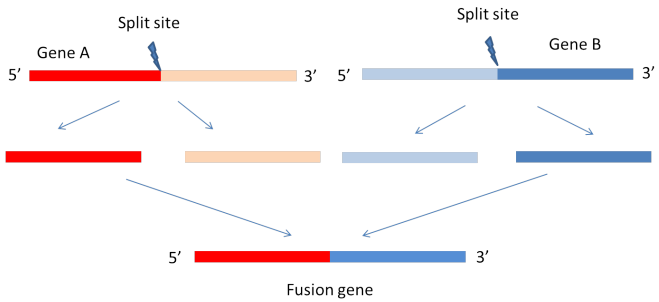
for these mechanisms?

Gene Fusion

Gene Fusion

Fused genes that can arise from chromosomal rearrangements and are analogous to novel isoform discovery. Although one challange is a much larger search space as we can no longer assume that the transcript segments are co-linear on a single chromosome.

Artifacts are common even using state-of-the-art tools, which primarily result from misalignment of read sequences due to polymorphisms, homology, and sequencing errors.



Discovery

Novel transcripts

Identifying novel transcripts using the short reads is a challenging tasks in RNA-seq. Short reads rarely span across several splice junctions and thus make it difficult to directly infer all full-length transcripts.

Data: PE reads and higher coverage help to reconstruct lowly expressed transcripts.
Study design: Replicates are essential to resolve false-positive calls at the low end of signal detection.
Methods: that incorporate existing annotations by adding them to the possible list of isoforms: Cufflinks, iReckon, SLIDE and StringTie