RNA-SEQUENCING ANALYSIS

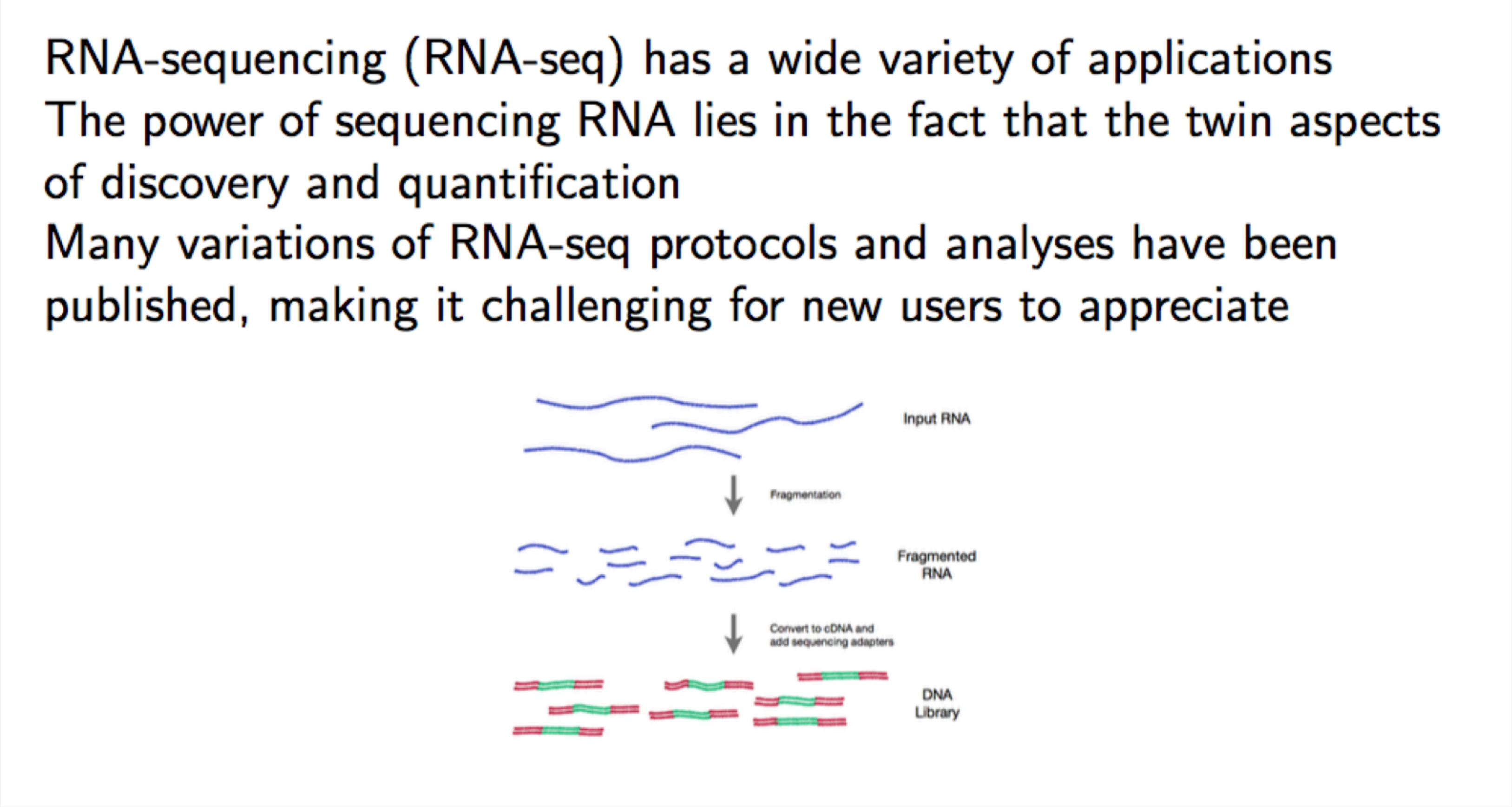
Joseph Powell

SISG- 2018

CONTENTS

- Introduction to RNA sequencing
- Data structure
- Analyses
  - Transcript counting
  - Alternative splicing
  - Allele specific expression
  - Discovery

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 EXTRACTIONS

What species of RNA are you interested in quantifying?

- Messenger RNA (mRNA) – accounts for just 5% of the total RNA in a cell
    - But is the most heterogeneous type in terms of both base sequence and size

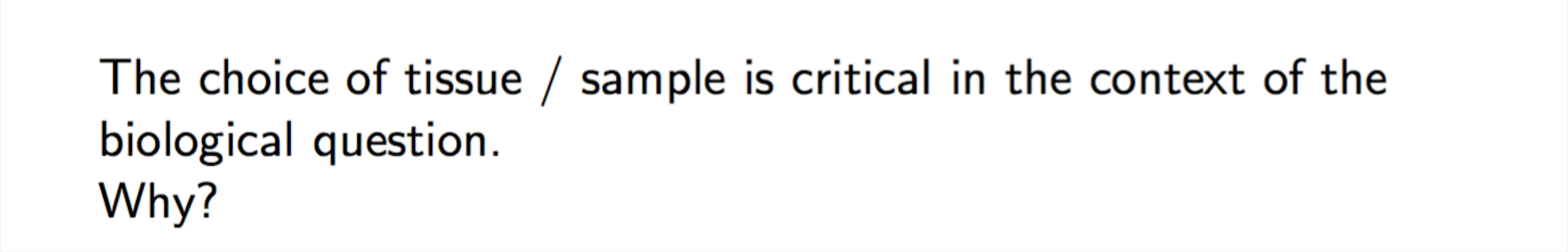
RNA EXTRACTIONS

What species of RNA are you interested in quantifying?

- Messenger RNA (mRNA) – accounts for just 5% of the total RNA in a cell
  - But is the most heterogeneous type in terms of both base sequence and size

- Non-coding (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.

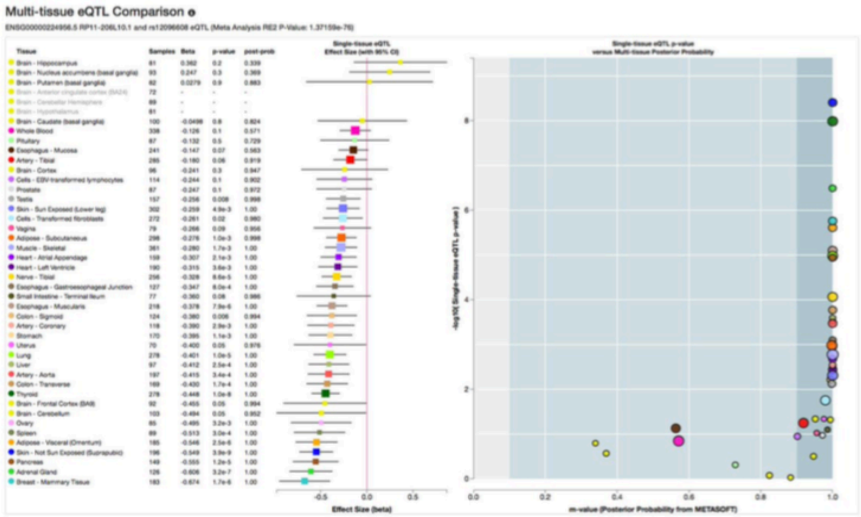
WHAT TISSUE AM I SAMPLING FROM?



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?



Multi-tissue eQTL Comparison

ENSG00000224956.5 RP11-206L10.1 and rs12096608 eQTL (Meta Analysis RE2 P-Value: 1.37159e-76)

Tissue	Samples	Beta	p-value	post-prob
● Brain - Hippocampus	81	0.362	0.2	0.339
● Brain - Nucleus accumbens (basal ganglia)	93	0.247	0.3	0.369
● Brain - Putamen (basal ganglia)	82	0.0279	0.9	0.883
● Brain - Anterior cingulate cortex (BA24)	72	-	-	-
● Brain - Cerebellar Hemisphere	89	-	-	-
● Brain - Hypothalamus	81	-	-	-
● Brain - Caudate (basal ganglia)	100	-0.0498	0.8	0.824
● Whole Blood	338	-0.126	0.1	0.571
● Pituitary	87	-0.132	0.5	0.729
● Esophagus - Mucosa	241	-0.147	0.07	0.583
● Artery - Tibial	285	-0.180	0.06	0.919
● Brain - Cortex	96	-0.241	0.3	0.947
● Cells - EBV-transformed lymphocytes	114	-0.244	0.1	0.902
● Prostate	87	-0.247	0.1	0.972
● Testis	157	-0.256	0.008	0.998
● Skin - Sun Exposed (Lower leg)	302	-0.259	4.9e-3	1.00
● Cells - Transformed fibroblasts	272	-0.261	0.02	0.980
● Vagina	79	-0.266	0.09	0.956
● Adipose - Subcutaneous	298	-0.276	1.0e-3	0.998
● Muscle - Skeletal	361	-0.280	1.7e-3	1.00
● Heart - Atrial Appendage	159	-0.307	2.1e-3	1.00
● Heart - Left Ventricle	190	-0.315	3.6e-3	1.00
● Nerve - Tibial	256	-0.328	8.6e-5	1.00
● Esophagus - Gastroesophageal Junction	127	-0.347	8.0e-4	1.00
● Small Intestine - Terminal Ileum	77	-0.360	0.08	0.986
● Esophagus - Muscularis	218	-0.378	7.9e-6	1.00
● Colon - Sigmoid	124	-0.380	0.006	0.994
● Artery - Coronary	118	-0.390	2.9e-3	1.00
● Stomach	170	-0.395	1.1e-3	1.00
● Uterus	70	-0.400	0.05	0.976
● Lung	278	-0.401	1.0e-5	1.00
● Liver	97	-0.412	2.5e-4	1.00
● Artery - Aorta	197	-0.415	3.4e-4	1.00
● Colon - Transverse	169	-0.430	1.7e-4	1.00
● Thyroid	278	-0.448	1.0e-8	1.00
● Brain - Frontal Cortex (BA9)	92	-0.455	0.05	0.994
● Brain - Cerebellum	103	-0.494	0.05	0.952
● Ovary	85	-0.495	3.2e-3	1.00
● Spleen	89	-0.513	3.0e-4	1.00
● Adipose - Visceral (Omentum)	185	-0.546	2.5e-6	1.00
● Skin - Not Sun Exposed (Suprapubic)	196	-0.549	3.9e-9	1.00
● Pancreas	149	-0.555	1.2e-5	1.00
● Adrenal Gland	126	-0.606	3.2e-7	1.00
● Breast - Mammary Tissue	183	-0.674	1.7e-6	1.00

Single-tissue eQTL
Effect Size (with 95% CI)
-0.5
0.0
0.5
Effect Size (beta)
Single-tissue eQTL p-value
versus Multi-tissue Posterior Probability
-log10(Single-tissue eQTL p-value)
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
m-value (Posterior Probability from METASOFT)

Resources: GTeX, TiGER, Encode, Fantom

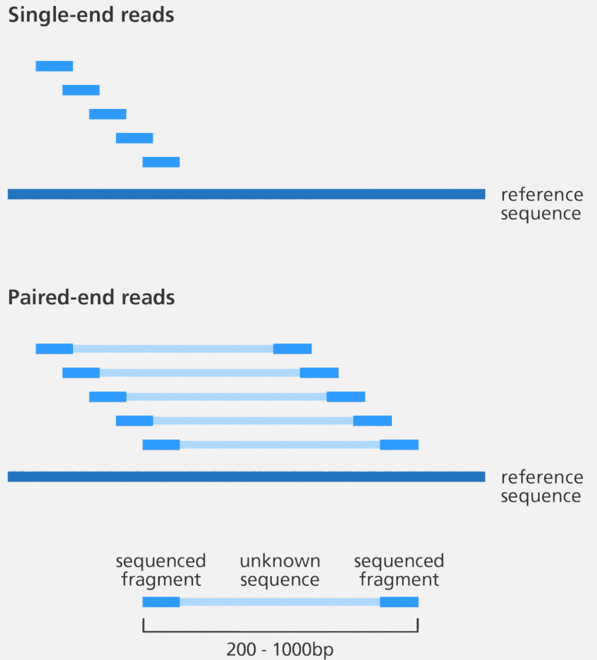
SINGLE VS. PAIRED-END

- 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
  - Isoform or splice variation
  - Allele specific expression

SINGLE VS. PAIRED-END

- 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
  - Isoform or splice variation
  - 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

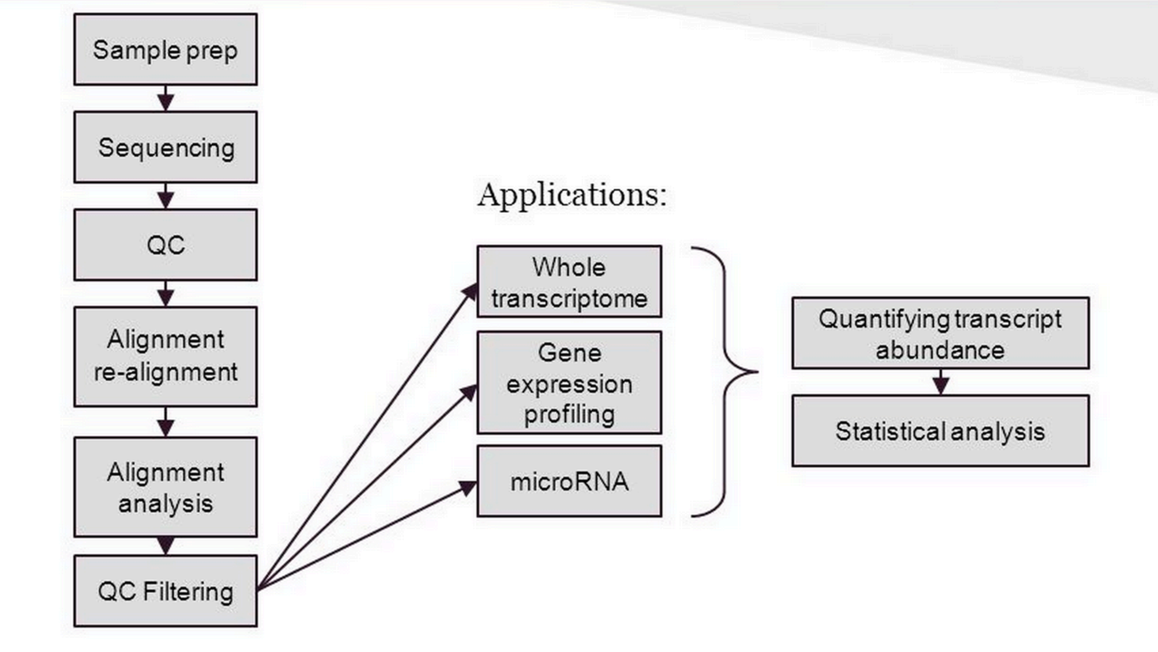
DATA

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

TRANSCRIPT COUNTING

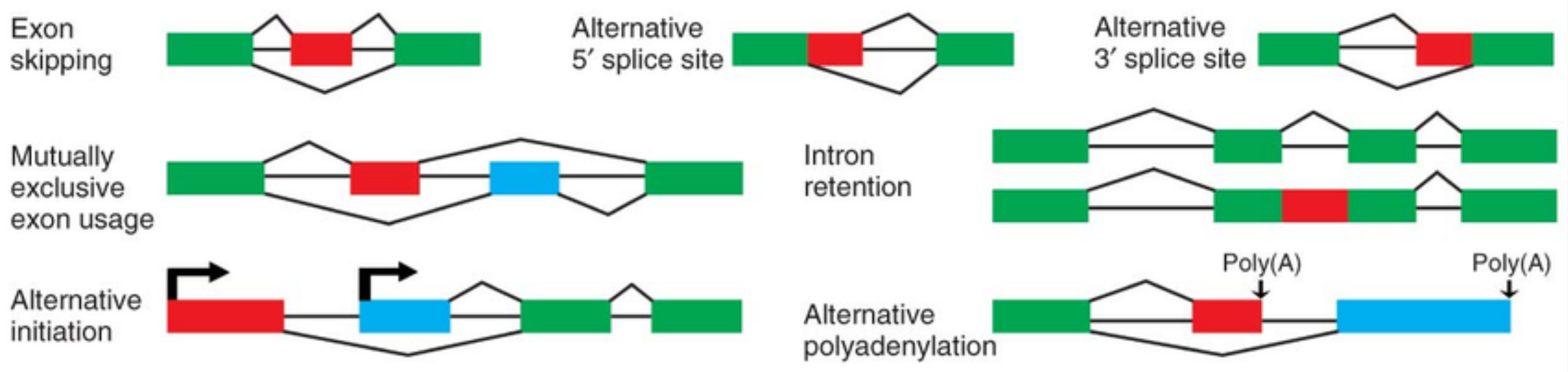
The most common application of RNA-seq is to estimate gene and transcript expression. This application is primarily based on the number of reads that map to each transcript sequence, although there are algorithms such as Sailfish that rely on k-mer counting in reads without the need for mapping



Wiki: list of RNA-Seq bioinformatics tools

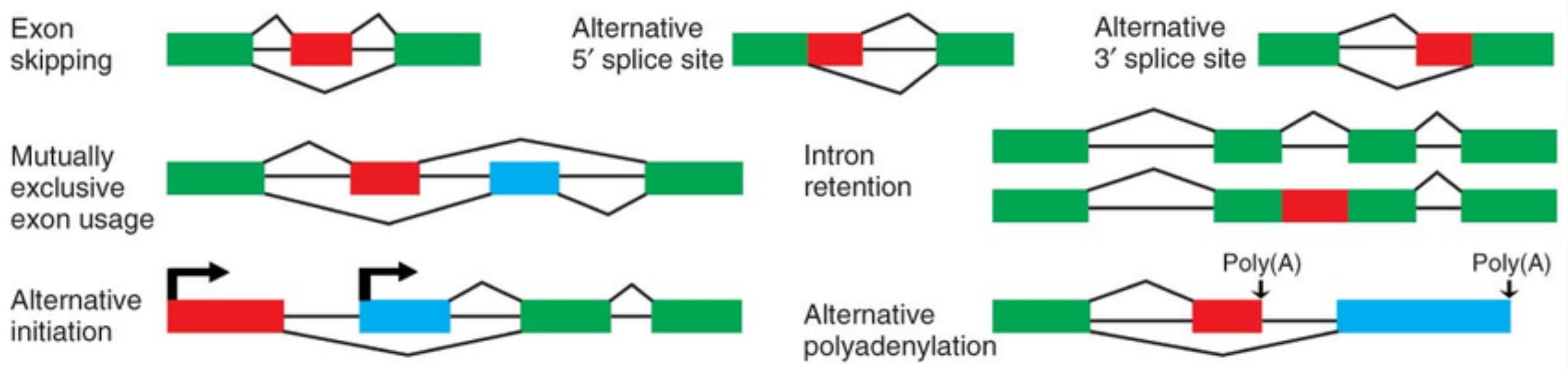
ALTERNATIVE SPLICING

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


Exon skipping
Alternative 5′ splice site
Alternative 3′ splice site
Mutually exclusive exon usage
Intron retention
Alternative initiation
Poly(A)
Poly(A)
Alternative polyadenylation

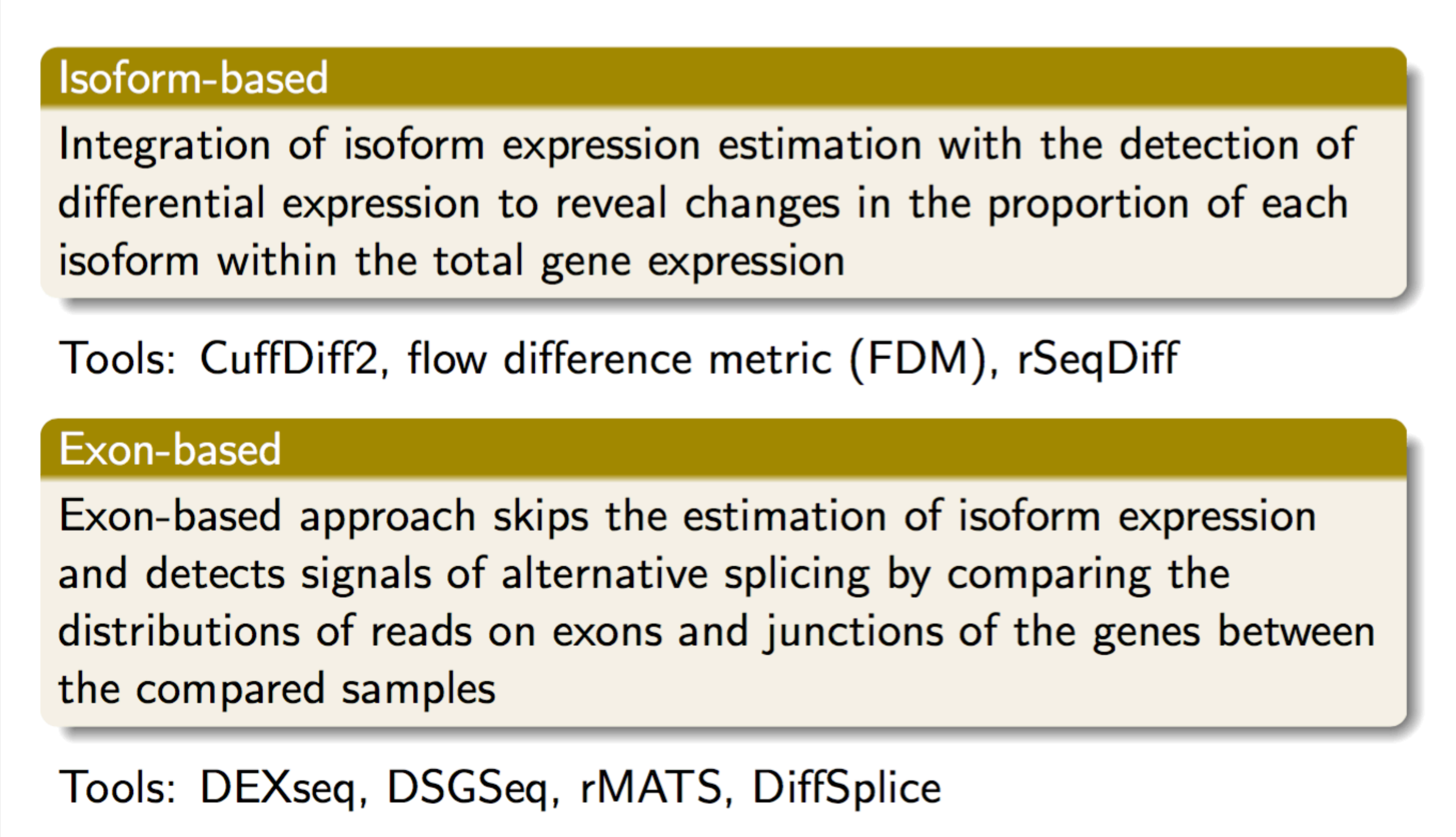
ALTERNATIVE SPLICING

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


Exon skipping
Alternative 5′ splice site
Alternative 3′ splice site
Mutually exclusive exon usage
Intron retention
Alternative initiation
Poly(A)
Poly(A)
Alternative polyadenylation

Detection methods fall into two major categories.

ALTERNATIVE SPLICING



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

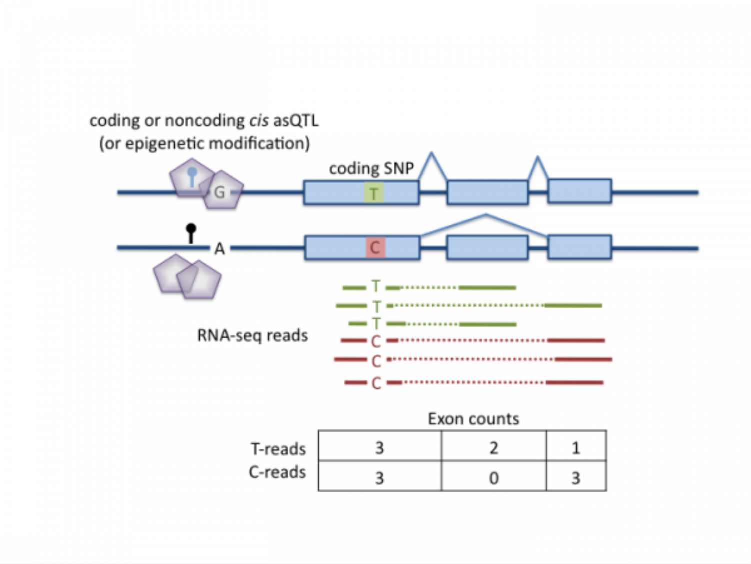
Exon-based

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

ALTERNATIVE SPLICING

Splice QTL - loci influencing splice variation

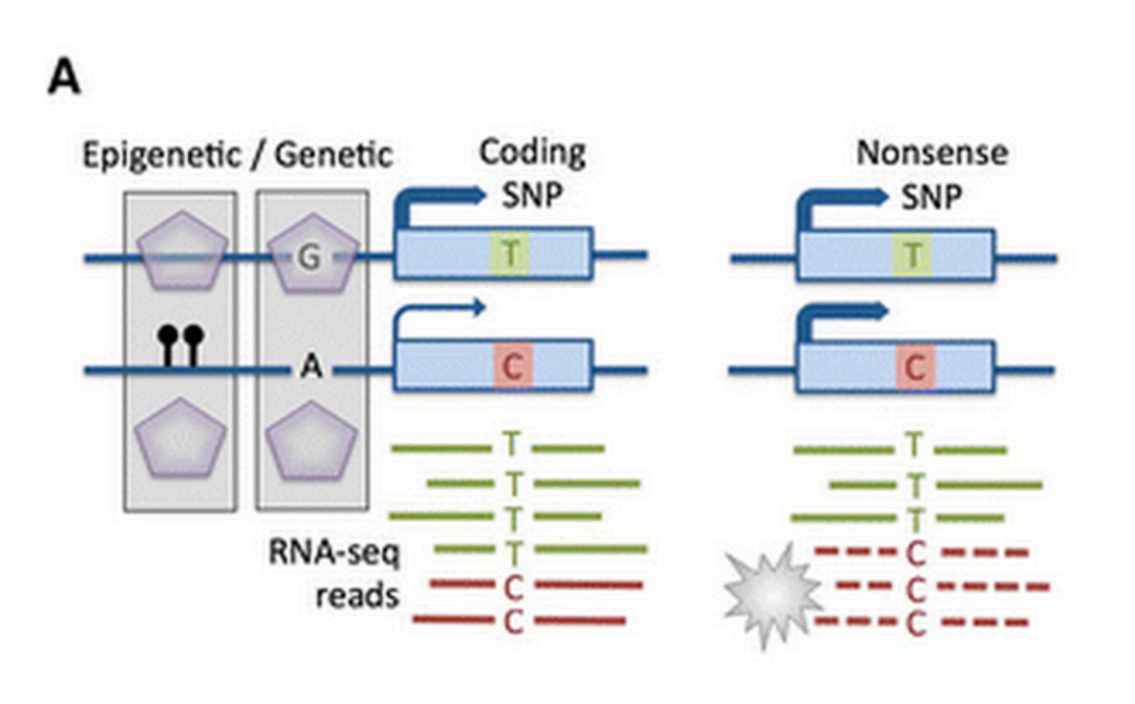


	Exon counts		
T-reads	3	2	1
C-reads	3	0	3

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

ALLELE SPECIFIC EXPRESSION

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



ALLELE SPECIFIC EXPRESSION

- Epigenetic phenomena, such as imprinting (when an inherited allele from one parent is consistently overexpressed)

- Allele-specific chromatin modifications.

- 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?

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.

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

THANK YOU

- Email me: j.powell@garvan.org.au
- Twitter: @JP_Garvan