Introduction To Transcriptomics

Dr Stevie Pederson (They/Them) stevie.pederson@thekids.org.au

Black Ochre Data Labs, The Kids Research Institute Australia

Acknowledgement Of Country

I'd like to acknowledge the Kaurna people as the traditional owners and custodians of the land we know today as the Adelaide Plains, where I live & work.

I also acknowledge the deep feelings of attachment and relationship of the Kaurna people to their place.

I pay my respects to the cultural authority of Aboriginal and Torres Strait Islander peoples from other areas of Australia, and pay my respects to Elders past, present and emerging, and acknowledge any Aboriginal Australians who may be with us today

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

What Is Transcriptomics

What Is Transcriptomics?

Transcription is the process of making an RNA copy of a gene sequence

- DNA can be described as being like a giant book of instructions
- Some regions are defined as *genes*
  - Originally considered to be the basic unit of inheritance
  - Now used to describe a region of DNA transcribed into RNA

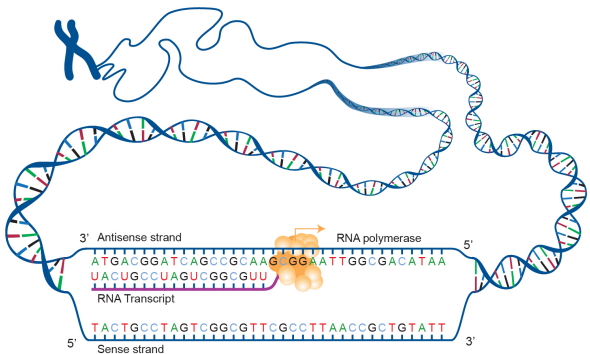


Figure 1: Figure taken from Anderson and Barbee (2016).

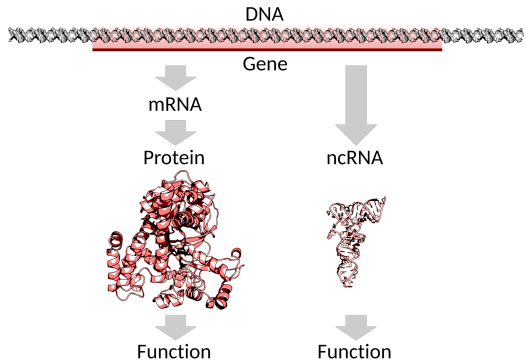
What Is Transcriptomics?

Transcriptomics is the study of transcribed RNA

Transcribed RNA falls into 2 broad classes

1. Messenger RNA (mRNA)
   - Codes for protein sequences
2. Non-coding RNA (ncRNA)
   - Multiple types of *functional* RNA
   - rRNA, tRNA $\implies$ protein translation
   - lncRNA, miRNA, snRNA, piRNA etc

Prokaryotes $\implies$ mRNA, rRNA & tRNA



By Thomas Shafee - Own work, CC BY 4.0, Wikimedia

The RNA Population Of a Eukaryotic Cell

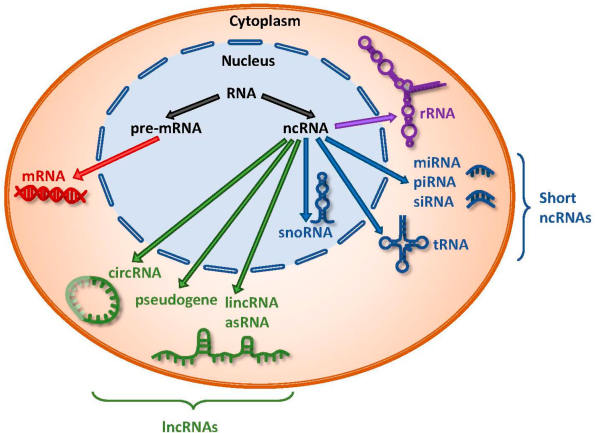


Figure 2: Image taken from Chan and Tay (2018)

- rRNA $\approx$ 80%[1]
- tRNA $\approx$ 15%
- All other RNA $\approx$ 5%

[1] https://bionumbers.hms.harvard.edu/bionumber.aspx

Functional RNA

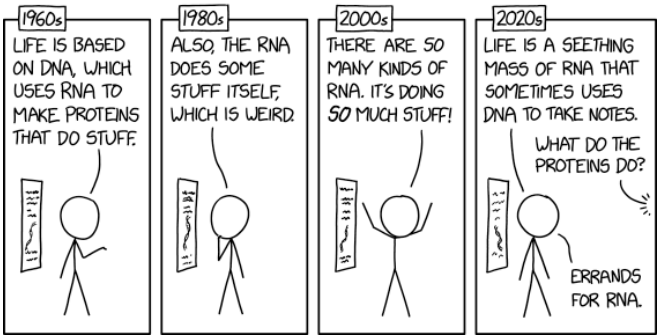


Figure 3: https://m.xkcd.com/3056/

(An incomplete list)

- pre-mRNA + mRNA
- lncRNA + lincRNA
- miRNA, siRNA, shRNA, piRNA
- rRNA + tRNA
- snRNA + snoRNA
- SRP RNA
- eRNA
- circRNA

Eukaryotic mRNA Processing

- Nuclear mRNA have 5' cap added
  - Protects single-stranded mRNA from degradation
  - Regulates nuclear export
  - Promotes translation into protein
- mRNAs are polyadenylated at the 3' end (-AAAAAAAAAAAAA)
  - Also protects from degradation
  - Aids in transcription termination, export and translation
- Introns are spliced out as required

Eukaryotic mRNA Processing

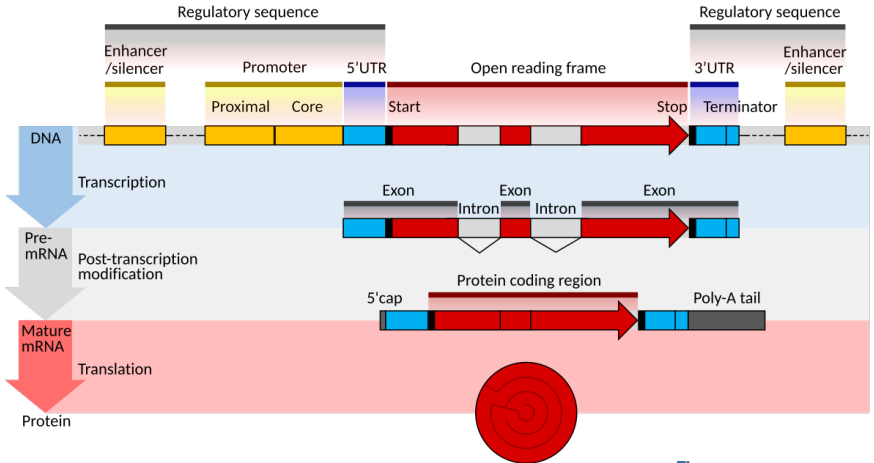


Figure 4: Taken from Shafee and Lowe (2017)

Eukaryotic mRNA Processing

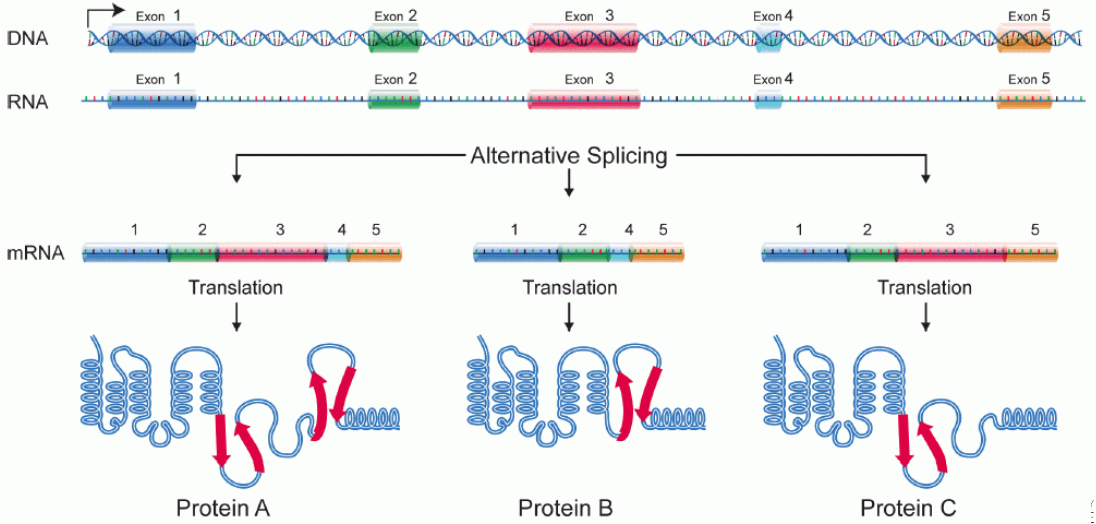


Figure 5: Image by the National Human Genome Research Institute

Why Study Transcriptomics?

Why Study Transcriptomics?

- Is a snapshot of highly **dynamic biological processes**
  - Captures response to stimulus and steady-state dynamics
- Assumed to be low-level
  - DNA $\rightarrow$ RNA $\rightarrow$ Protein $\rightarrow$ Metabolites, Signalling molecules, etc ...
- Use to make inference about these biological processes of interest
  - Can infer specific cell-cell communication methods
  - Identify therapeutic targets for Cardiovascular Disease, biomarkers for CAR-T cells etc

Quantitative Approaches

- RNA expression is a rapid, early response to stimuli
  - Could be immune signalling, drug treatment etc
- Also changes in steady-state over time
  - First trimester placenta is hypoxic $\implies$ later is normoxic
- Change in a gene's transcriptional activity $\implies$ change in RNA abundance
  - Capturing changes in abundance $\implies$ measure RNA quantities
- Expression patterns $\implies$ identify cell-types in a heterogeneous sample
- Changes in splicing patterns
  - Require methods for quantifying isoforms *within* a gene
  - May be changing proportions within gene-level abundances

Sequence Based Approaches

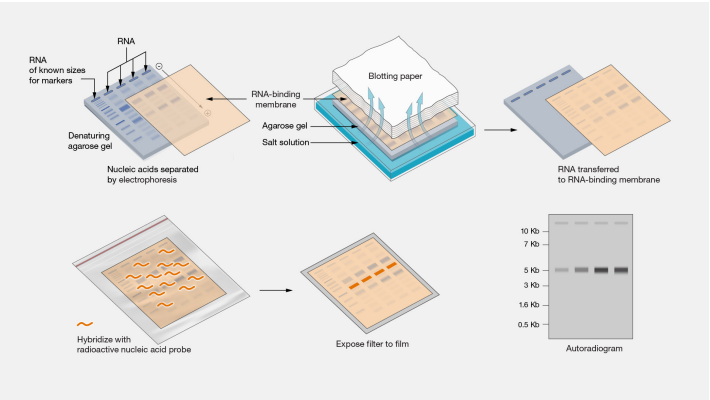
- Identify novel transcript sequences
- No reference genome/transcriptome
  - Compare novel sequences against known transcriptomes $\implies$ infer function
- Sequences may diverge from reference
  - Do SNPs + InDels impact splicing/expression patterns in individual organisms
- Unexpected splicing patterns and chimeric RNA
  - Real genomes are less "neat" than reference genomes
  - Can play a key role in leukaemias & other cancers $\implies$ clinical diagnostic

The Development of Transcriptomics

Early Transcriptomics

- The field developed with few reference sequences
  - Human Genome Project (1990-2003)
- Single sequence methods
  - Quantitative: Northern Blot (1977) + qPCR (1996)
  - Sequence Identification: Sanger Sequencing (1977)
- High-Throughput Era
  - Quantitation: SAGE (1995) $\rightarrow$ Microarrays (1996)
  - Sequence Identification: ESTs (1991)

Northern Blots



- Probes require sequence knowledge
- Clear Presence/Absence calls
- Crude quantitation: Densitometric Analysis

Figure 6: Figure taken from https://www.genome.gov/genetics-glossary/Northern-Blot

RT-qPCR

- "Gold-standard" for measurement of transcription levels
  - Single gene $\implies$ not a high-throughput technique
- Targets a single transcript region with specific primers to produce cDNA $\rightarrow$ Polymerase Chain Reaction (PCR)
- Each PCR cycle approximately doubles the target region
- cDNA produced is identified using fluorophores
  - Fluorescence doubles with each cycle
- Once fluorescence passes a detection threshold, the cycle number is recorded
  - Known as the Cycle Threshold (C_T) value

RT-qPCR

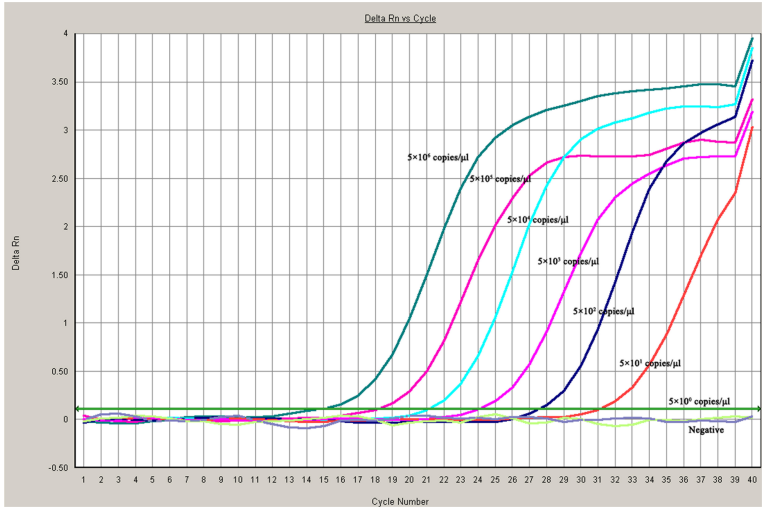


Figure 7: A 10-fold dilution series

RT-qPCR

- Higher C_T values $\implies$ lower numbers of target molecule at the beginning
- These can be used to estimate and compare abundance levels (i.e. gene expression)

- Is vulnerable to technical artefacts (e.g. pipetting & sample variability)
- Often includes one or more "housekeeper" genes thought to be stably expressed

- C_T values are *normalised* to the housekeeper genes $\implies C_{T_{hk}}$

RT-qPCR

- Higher C_T values $\implies$ lower numbers of target molecule at the beginning
- These can be used to estimate and compare abundance levels (i.e. gene expression)

- Is vulnerable to technical artefacts (e.g. pipetting & sample variability)
- Often includes one or more "housekeeper" genes thought to be stably expressed

- C_T values are *normalised* to the housekeeper genes $\implies C_{T_{hk}}$
  - $\log_2$ transformed values are used: $\Delta C_T = \log_2 C_{T_g} - \log_2 C_{T_{hk}}$

RT-qPCR

- Higher C_T values $\implies$ lower numbers of target molecule at the beginning
- These can be used to estimate and compare abundance levels (i.e. gene expression)

- Is vulnerable to technical artefacts (e.g. pipetting & sample variability)
- Often includes one or more "housekeeper" genes thought to be stably expressed

- C_T values are *normalised* to the housekeeper genes $\implies C_{T_{hk}}$
  - $\log_2$ transformed values are used: $\Delta C_T = \log_2 C_{T_g} - \log_2 C_{T_{hk}}$
- Change between conditions is the *change in* $\Delta C_T \implies \Delta\Delta C_T$

RT-qPCR

- Higher C_T values $\implies$ lower numbers of target molecule at the beginning
- These can be used to estimate and compare abundance levels (i.e. gene expression)

- Is vulnerable to technical artefacts (e.g. pipetting & sample variability)
- Often includes one or more "housekeeper" genes thought to be stably expressed

- C_T values are *normalised* to the housekeeper genes $\implies C_{T_{hk}}$
  - $\log_2$ transformed values are used: $\Delta C_T = \log_2 C_{T_g} - \log_2 C_{T_{hk}}$
- Change between conditions is the *change in* $\Delta C_T \implies \Delta\Delta C_T$
- Represents change on the $\log_2$ scale, i.e. *log fold-change*

Expressed Sequence Tags (ESTs)

- The first attempt at capturing the larger transcriptome was ESTs (Adams et al. 1991)
- Identified 609 human brain mRNA sequences
  - Selected for polyA-mRNA then reverse transcribed
  - Used random primers $\rightarrow$ Sanger Sequencing
- 10 years before the Human Genome Project
  - Gene discovery was a hot topic

Serial Analysis of Gene Expression (SAGE)

- First high-throughput quantification method was *Serial Analysis of Gene Expression* (SAGE) (Velculescu et al. 1995)

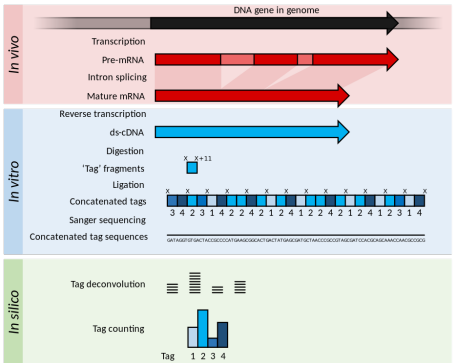


Figure 8: Thomas Shafee, CC BY 4.0, via Wikimedia Commons

- mRNA → cDNA using biotinylated primers
- cDNA bound to beads (using biotin) & cleaved
- 11mer "tags" were ligated into long sequences using linker sequences
- Sequenced using Sanger Sequencing
- Deconvolution & counting
- First count-based transcriptomic methods developed

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Microarray Technology

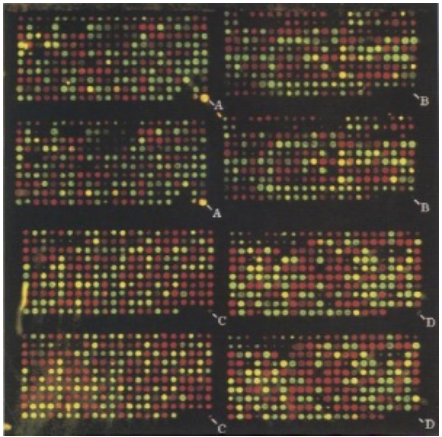


Figure 9: Section of two-colour array taken from Shalon, Smith, and Brown (1996)

- Truly launched the modern transcriptomics era
- Quantified thousands of transcripts simultaneously
- Relied on development of Human Genome Project (+ other organisms)
- Analysis in R/Bioconductor
  - Rv1.0.0 (2000)
  - Bioconductor (Gentleman et al. 2004)
  - Modern statistical high-throughput models developed

References

Adams, Mark D., Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R. Merril, et al. 1991. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science* 252 (5013): 1651–56. http://www.jstor.org/stable/2876333.

Anderson, Christine, and Lisa Bartee. 2016. *Mt Hood Community College Biology 102*. Open Oregon Educational Resources.

Chan, Jia Jia, and Yvonne Tay. 2018. "Noncoding RNA:RNA Regulatory Networks in Cancer." *International Journal of Molecular Sciences* 19 (5). https://doi.org/10.3390/ijms19051310.

Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biol.* 5 (10): R80.

Shafee, Thomas, and Rohan Lowe. 2017. "Eukaryotic and Prokaryotic Gene Structure." *WikiJournal of Medicine*, January. https://doi.org/10.15347/WJM/2017.002.

Shalon, D, S J Smith, and P O Brown. 1996. "A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization." *Genome Research* 6 (7): 639–45. https://doi.org/10.1101/gr.6.7.639.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. "Serial analysis of gene expression." *Science* 270 (5235): 484–87.