Lecture 2: Early Transcriptomic Strategies

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

March 16th, 2020

Overview

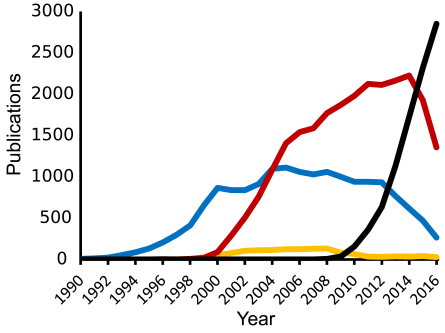
The Motivation

- The transcriptome is a highly dynamic set of molecules
- Small changes can potentially have significant ramifications
  - e.g. a "Master Regulator" can determine cellular fate
- RNA molecules are small
  - How do we find what's in our sample?
  - How do we quantify RNA?
  - And how do we compare one or more groups?

Technological Developments

- Technological developments are constant
- Technologies are often transient
- Key technologies are:
  1. Real Time Polymerase Chain Reaction (RT-PCR)
  2. Expressed Sequence Tags (EST)
  3. Serial/Cap Analysis of Gene Expression (SAGE/CAGE)
  4. Microarray technologies
  5. Sequencing technologies
- Analytic methodologies *often lag technologies*

A Simplified History



EST (blue); SAGE / CAGE (yellow); Microarrays (red); RNA Seq (black)[1]

[1] Rohan Lowe et al. "Transcriptomics technologies". In: *PLOS Computational Biology* 13.5 (May 2017), pp. 1–23. DOI: 10.1371/journal.pcbi.1005457. URL: https://doi.org/10.1371/journal.pcbi.1005457.

Measuring Single Genes

The Northern Blot

- One of the earliest strategies[2]
- Developed as an extension of the Southern Blot[3] (DNA)
- Gel Electrophoresis-based strategy
  - Based on size differentiation and probe sequences

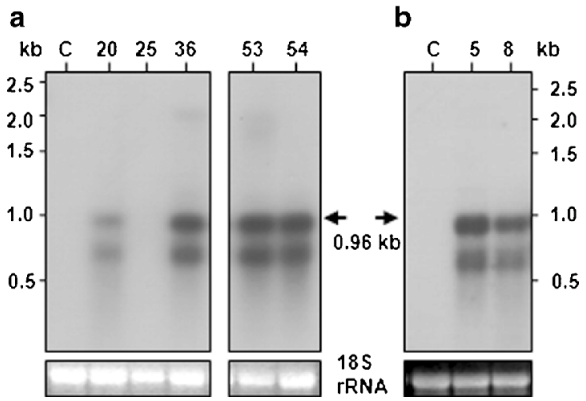
[2] J. C. Alwine, D. J. Kemp, and G. R. Stark. "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes". In: *Proc. Natl. Acad. Sci. U.S.A.* 74.12 (1977), pp. 5350–5354.

[3] E. M. Southern. "Detection of specific sequences among DNA fragments separated by gel electrophoresis". In: *J. Mol. Biol.* 98.3 (1975), pp. 503–517.

The Northern Blot

- RNA is extracted then denatured
- RNA is size separated using Gel Electrophoresis
- RNA is transferred to a "blotting membrane"
- Treat the membrane with a labelled probe
  - Probes are complementary to the "target sequence"
  - Probes are labelled with fluorescent dye or radioactive atoms

The Northern Blot



Taken from Ramkumar et al., "Effect of orientation of transcription of a gene in an inverted transferred DNA repeat on transcriptional gene silencing in rice transgenics-a case study"

The Northern Blot

- Prominent usage *before* genomes were sequenced
- Can possibly detect different isoforms
- Crude quantitation using Densitometric Analysis
  - What limitations might this have?

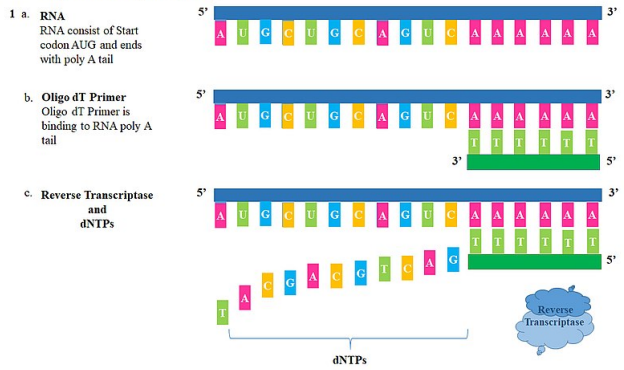
RT-qPCR

- Reverse Transcriptase quantitative PCR
  - Sometimes called: qPCR, RT-PCR
- Often considered to be the "gold standard" for quantitation
- Targets a *specific transcribed region* via *specific primers*
  - Primers must be individually designed
  - Primers often span exon-exon junctions

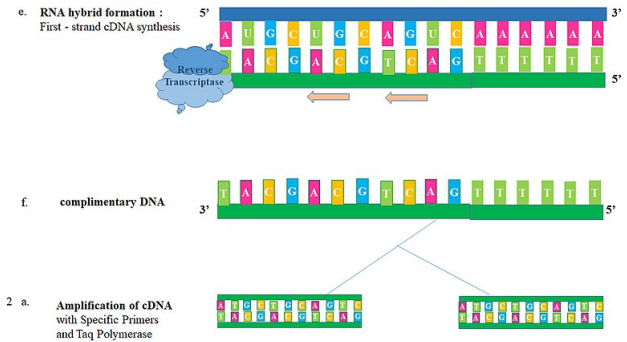
RT-qPCR

1. *Reverse Transcriptase* converts RNA to cDNA
   - Primers are required: Can target poly-A or random
2. Sequence-specific primers amplify the target fragment in cycles
   - Fluorescent dye is commonly incorporated during amplification
3. Abundance of target will grow exponentially ($\times 2$) for each amplification cycle
4. The cycle where abundance reaches the "limit of detection" is estimated (C_T)

RT-qPCR

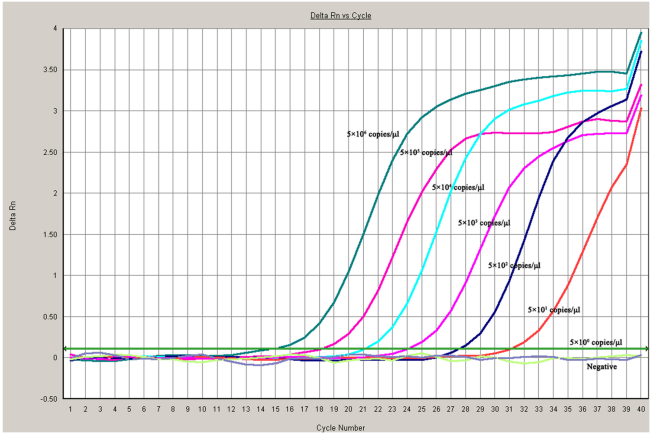


RT-qPCR



©Lokesh Thimmana, under the guidance of Dr. G. Mallikarjuna, Assistant Professor, Molecular Biology, Agri Biotech Foundation.

RT-qPCR



This is a 10-fold dilution series[4]

[4] Ma Mingxiao et al. "TaqMan MGB Probe Fluorescence Real-Time Quantitative PCR for Rapid Detection of Chinese Sacbrood Virus". In: *PLoS ONE* 8.2 (Feb. 2013), pp. 1–7. DOI: 10.1371/journal.pone.0052670. URL: https://doi.org/10.1371/journal.pone.0052670.

RT-qPCR

- Can be used with a standard curve and dilution series to estimate absolute quantity of an RNA *within a sample*
- Can be used to compare *across samples for relative abundance*

RT-qPCR

- Can be used with a standard curve and dilution series to estimate absolute quantity of an RNA *within a sample*
- Can be used to compare *across samples for relative abundance*

What may be a fundamental issue when comparing across samples?

Normalisation

- There may be pipetting and other technical differences between samples
  - These are **non-biological** in origin
- To correct for these we can **normalise** our data
- In RT-qPCR this is often done using "housekeeper" genes
  - We choose genes which *should not* change between samples/groups
  - These are commonly structural genes such as *ACTN*β or *GAPDH*

Estimating Change In Expression

- Relative abundances are often referred to as fold-change (FC)
  - Down regulation is squeezed between 0 and 1
  - Up regulation ranges from 1 to ∞
- We often use $\log_2$ fold-change to get a better scale, e.g.
  - A 2-fold increase in abundance: $\log_2 2^1 = 1$
  - A 2-fold decrease in abundance: $\log_2 \frac{1}{2} = \log_2 2^{-1} = -1$
  - No change in abundance $\log_2 1 = \log_2 2^0 = 0$
- This is often abbreviated as *logFC*

Estimating Change In Expression

- For *RT-qPCR* the estimate of *logFC* is known as $\Delta\Delta C_T$
- To calculate this, we calculate **two** changes in C_T
  1. ΔC_T relative to the housekeeper(s)
  2. $\Delta\Delta C_T$ across samples for our gene/fragment of interest
- The first step corrects for technical errors
- The second step estimates our true change in abundance

Estimating Change In Expression

Within each sample

$$\Delta C_T = C_{t[\text{gene}]} - C_{t[\text{HK}]}$$

Across samples/groups

$$\Delta\Delta C_T = -(\Delta C_{T[\text{group1}]} - \Delta C_{T[\text{group2}]})$$

This formulation assumes *equal amplification efficiency* for all primers/genes (i.e. $\textit{Efficiency} = 2$)

Estimating Change In Expression

- Housekeeper genes must be matched to the "gene of interest" **within each sample** and **within each qPCR reaction**
- Choosing > 1 housekeeper gene is advised
- Measurements are often taken in triplicate/quadruplicate for each sample (reactions sometimes fail)

Both Northern blots and RT-qPCR use targeted primers, but in very different ways

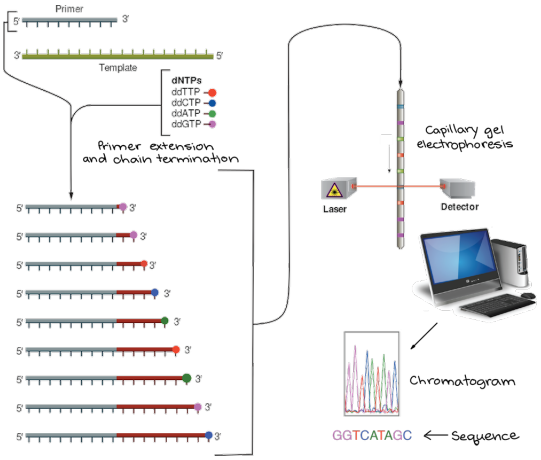
Measuring Multiple Genes

Expressed Sequence Tags

- The first attempt at capturing the larger transcriptome was via Expressed Sequence Tags[5] (ESTs) in 1991
  - Sequenced 609 mRNA human brain mRNA sequences
  - ESTs were generated by reverse transcribing poly-A selected mRNA, amplified using random primers
  - Used ESTs $\sim 100 - 800$nt
  - Obtained actual sequences using Sanger Sequencing
- >10 years before the Human Genome Project completed
- Just **discovering** genes was a huge priority

[5] Mark D. Adams et al. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project". In: *Science* 252.5013 (1991), pp. 1651–1656. ISSN: 00368075, 10959203. URL: http://www.jstor.org/stable/2876333.

Expressed Sequence Tags



Source: https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing

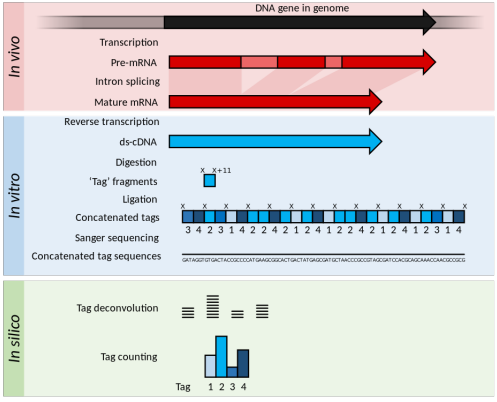
Serial Analysis of Gene Expression

Serial Analysis of Gene Expression[6] (SAGE) was the first attempt to quantify expression on a larger scale

1. Conversion of mRNA to ds-cDNA using biotinylated primers (often poly-T)
2. cDNA is bound to beads using biotin and cleaved
3. 11-mer "tags" were produced after cleavage and concatenated
4. Sequenced by Sanger Sequencing
5. Tags were "de-convoluted" and counted

[6] V. E. Velculescu et al. "Serial analysis of gene expression". In: *Science* 270.5235 (1995), pp. 484–487.

Serial Analysis of Gene Expression



Also see: `https://www.scq.ubc.ca/wp-content/uploads/2006/07/SAGE3b.gif`

Serial Analysis of Gene Expression

- The word "tag" is still commonly used in some NGS manuals and software
- The term "Digital Gene Expression" arose during this era
  - Is sometimes shortened to DGE, but **does not** stand for *Differential* Gene Expression.
- SAGE doesn't rely on probes targeting known sequences
- Variants on the technique are still used[7]
  - Even used these concatenated tags in early NGS contexts[8]

[7] A. M. Zawada et al. "Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease". In: *Epigenetics* 9.1 (2014), pp. 161–172.

[8] H. Matsumura et al. "SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays". In: *Nat. Methods* 3.6 (2006), pp. 469–474.

Cap Analysis of Gene Expression

- A variant technique is *Cap Analysis of Gene Expression*[9]
- Targets Transcription Start Site (TSS) of mRNA via the 5' cap
  - Specifically for identification of the exact TSS and analysis of promoters
- Original 27nt long, but now only limited by NGS length
- Heavily used in FANTOM (Functional ANnoTation Of the Mammalian genome) project

[9] R. Kodzius et al. "CAGE: cap analysis of gene expression". In: *Nat. Methods* 3.3 (2006), pp. 211–222.

SAGE Vs CAGE

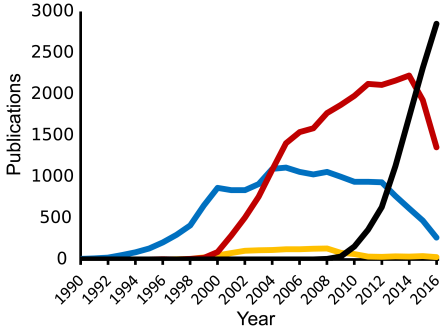
- Primers which target the poly-A sequence will capture *mature* mRNA
  - mRNA will also be intact (i.e. not degraded)
- CAGE targets transcriptional initiation
  - Transcripts may not be "mature"
  - 5' Cap must be in place (i.e. not degraded)
- **Both techniques** still involve concatenation of "tags"

Microarray Technology

Microarrays

- Microarrays effectively ushered in the modern era of transcriptomics
- Purely interested in *relative abundances*
- Could measure expression levels for 1000's of genes simultaneously, for *the first time*
- Were essentially glass slides with probes affixed to them

Microarrays



EST (blue); SAGE / CAGE (yellow); Microarrays (red); RNA Seq (black)[10]

[10] Rohan Lowe et al. "Transcriptomics technologies". In: *PLOS Computational Biology* 13.5 (May 2017), pp. 1–23. DOI: 10.1371/journal.pcbi.1005457. URL: https://doi.org/10.1371/journal.pcbi.1005457.

Microarrays

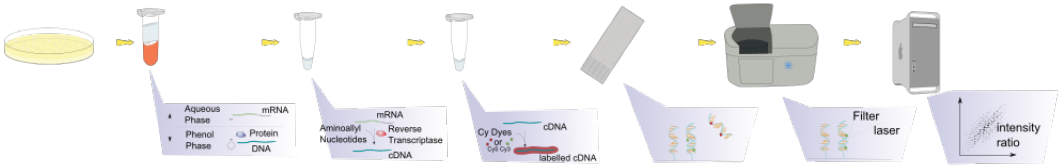
- Once again depends on reverse transcriptase for mRNA $\rightarrow$ cDNA
- **No reliance on Sanger Sequencing**
- Used probes (like a Northern blot) but the **cDNA is labelled and the probes are spatially fixed**
  - Probes must be designed beforehand
  - Probes are fixed to the array in *known locations*

Microarrays

1. Fluorescent labelling during mRNA conversion to cDNA
2. Complimentary probes bind target sequences (hybridisation)
3. Fluorescence detection at each probe

Fluoresence Intensity $\propto$ mRNA abundance

Microarrays

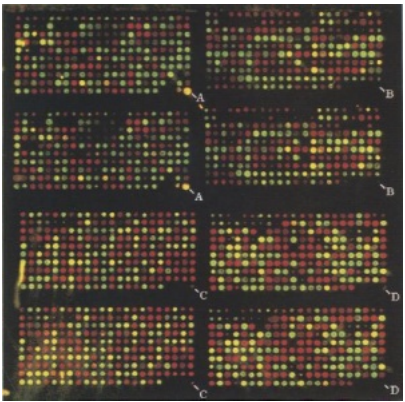


Source: https://commons.wikimedia.org/wiki/File:Microarray_exp_horizontal.svg

Two Colour Microarrays

- Probes with known sequences are at known locations
  - Probes were 65mer complimentary cDNA
  - Originally printed in local facilities
- Samples are labelled with *either* Cy3 (Green @ 570nm) or Cy5 (Red @ 670nm)
- Both samples are hybridised to array
- Relative Red/Green intensities were of interest
- Gave an estimate of logFC within each array

Two Colour Microarrays



A section of a two colour array[11]

[11] D Shalon, S J Smith, and P O Brown. "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.". In: *Genome Research* 6.7 (1996), pp. 639–645. DOI: 10.1101/gr.6.7.639. eprint: http://genome.cshlp.org/content/6/7/639.full.pdf+html. URL: http://genome.cshlp.org/content/6/7/639.abstract.

Two Colour Microarrays

- Probes are "printed" to the array
  - Print tips can get clogged
- Able to be customised for your own experiment
  - We need a mapping file for probe location to target sequence
- Both colours were scanned individually
  - One scan detects red only, the next detects green only
  - Each scan would have to be aligned with the other

Two Colour Microarrays

- Spots were detected using astronomical software
  - Detection of true signal above background (DABG)
- "Spots" could be of variable size
- Dye bias was noted *implies* experiments often used dye swaps
  - One sample might be labelled with red on one array, then labelled with green on the next

Single Channel Microarrays

- 3' Arrays (Affymetrix) became the dominant transcriptomic technology until RNA seq
- Probes target the 3' end of transcripts - reduce issues with RNA degradation
- Single channel (i.e. single colour)
- One sample per arrays
- $\sim$1,000,000 $\times$ 25-mer probes

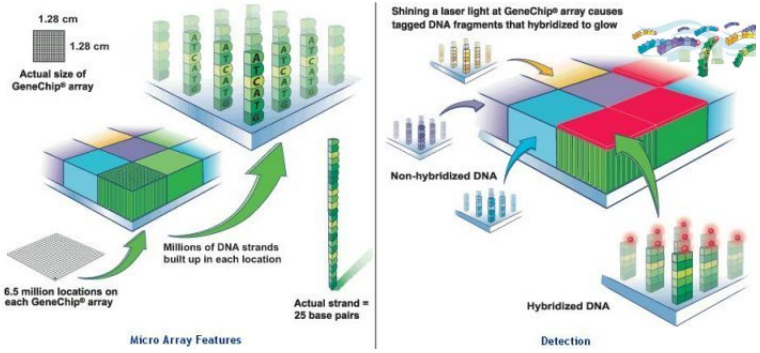
Single Channel Microarrays

Source: https://commons.wikimedia.org/wiki/File:Affymetrix-microarray.jpg

Single Channel Microarrays

- Manufacture used photolithography
- Greater density of probes than two-colour arrays
  - Shorter probes but far more of them
- Also need a mapping file from location to probe sequence

Single Channel Microarrays



Source: https://universe-review.ca/R11-16-DNAsequencing.htm

Single Channel Microarrays

- Each 3' exon would be targeted by 11 unique probes
- The set of 11 probes would be collected together as a single "probeset"
- Alternate isoforms with different 3' exons could be detected easily as they would have distinct probesets