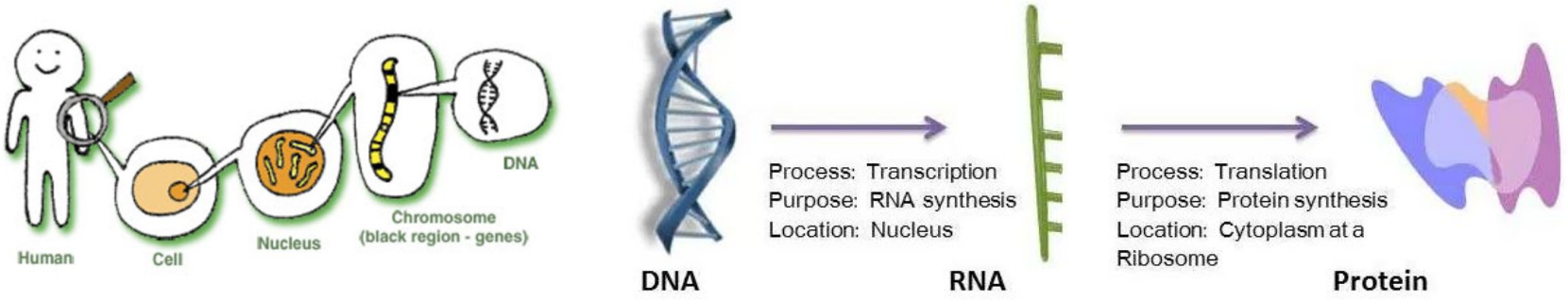
Introduction

This is to give anyone a high-level rough guide to Gene Expression. It was originally written in 2017 and therefore is out of date, but the high-level ideas are hopefully still valid.

What is a gene?

The Central Dogma (Francis Crick, 1958)

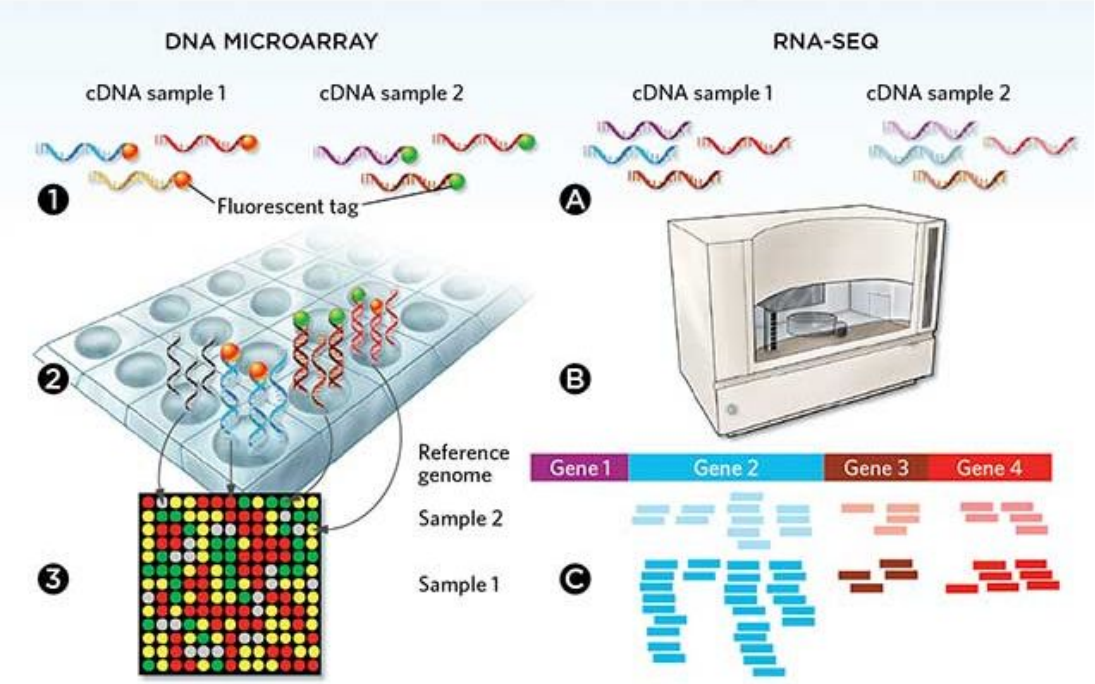


thatbiologist.wordpress.com

- DNA contains the original codes for making the proteins that living things need.
- mRNA (messenger RNA) is a copy of a gene located on the DNA molecule.
- mRNA will leave the nucleus of the cell and the ribosome will create the protein based on that mRNA.
- So we can measure the amount of mRNA in a cell as a surrogate measure of gene expression (ie. an expressed gene will make a copy of its mRNA, whereas a silent gene won't).

Measuring mRNA

- There are machines that can measure the mRNAs in a sample.
- For a long time, microarray were used for this, but now RNASeq machines have superseded the microarrays.

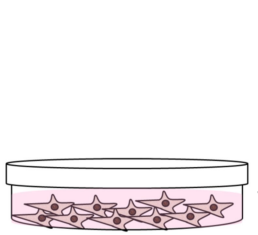


Kate Yandell, www.the-scientist.com

- scRNASeq machines can work on a single cell as its sample, rather than a collection of cells.

Primer - What is a dataset?

For genomics data analysis purposes, a dataset is the **single output** from a machine used to make the measurement (eg. gene expression).



sample

sequencer

	10w-AML1_S7	10w-AML2_S8	10w-AML3_S9	2w-A
ENSMUSG00000089699	0	0	0	0	0
ENSMUSG00000085623	0	0	0	0	0
ENSMUSG00000090031	10	5	4	6	5
ENSMUSG00000089982	0	0	0	0	0
ENSMUSG00000089787	0	0	0	0	0
ENSMUSG00000089914	0	0	0	0	0
ENSMUSG00000037432	159	189	264	277	272
ENSMUSG00000086064	0	0	0	3	0
ENSMUSG00000089767	2	0	0	2	0
ENSMUSG00000087374	0	0	0	0	0
ENSMUSG00000085391	4	4	7	2	10
ENSMUSG00000085280	1	0	0	0	0
ENSMUSG00000087131	0	0	1	0	1
ENSMUSG00000060424	0	0	0	0	0
ENSMUSG00000093778	0	0	0	0	0

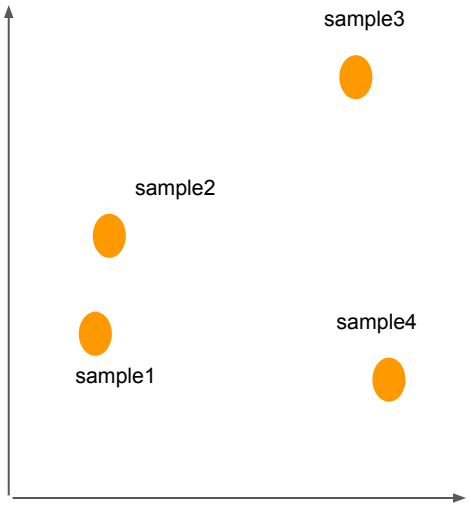
dataset

So you could end up with multiple datasets from same starting samples if they were split up and run separately on the machine.

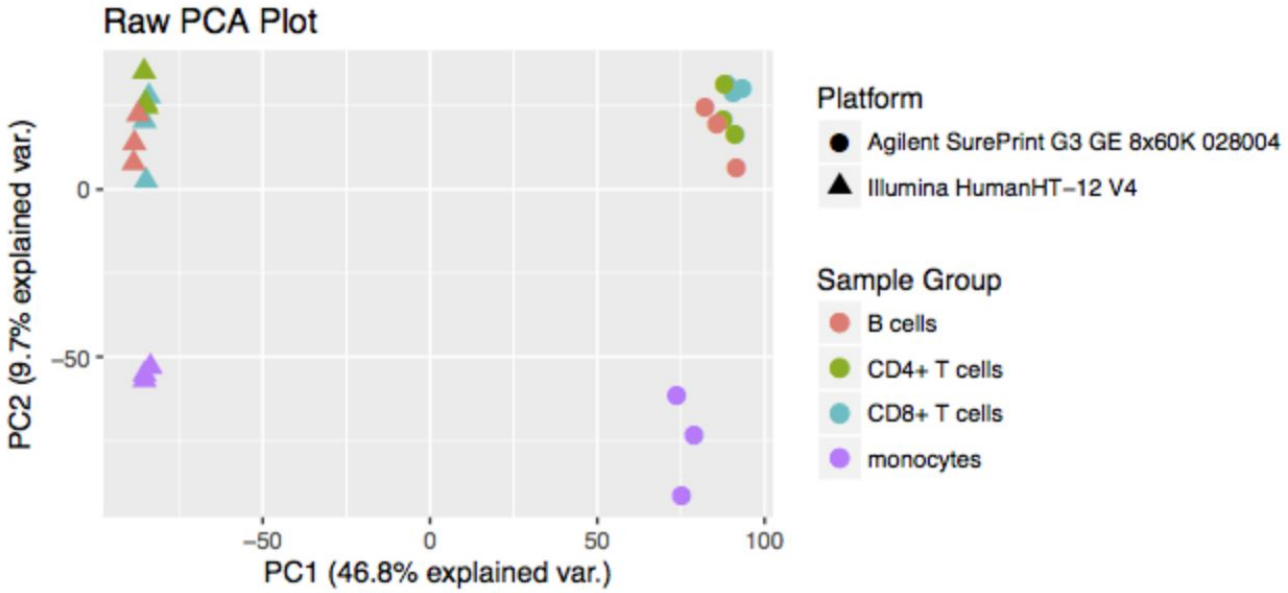
Primer - What is a PCA?

PCA, principal components analysis, is a commonly used technique to address variability in your data. Usually if two points are close together on a PCA plot, it means they are close in terms of data.

	sample1	sample2	sample3	sample4
gene1	2	4	0	8
gene2	5	10	0	0
gene3	0	0	1	0



Primer - What is the problem with merging datasets?



PCA of the samples when two datasets are concatenated (no correction) shows clear separation based on dataset and not on cell type.

ie. The biological signal is swamped by the technical artifact of these samples being run separately.

Primer - What is normalisation?

Data normalisation changes values of a dataset or across multiple datasets in order to make fair comparisons.

Eg. In this toy example, is gene2 more highly expressed in sample2 than in sample1? Perhaps not if you take total values into account for each sample.

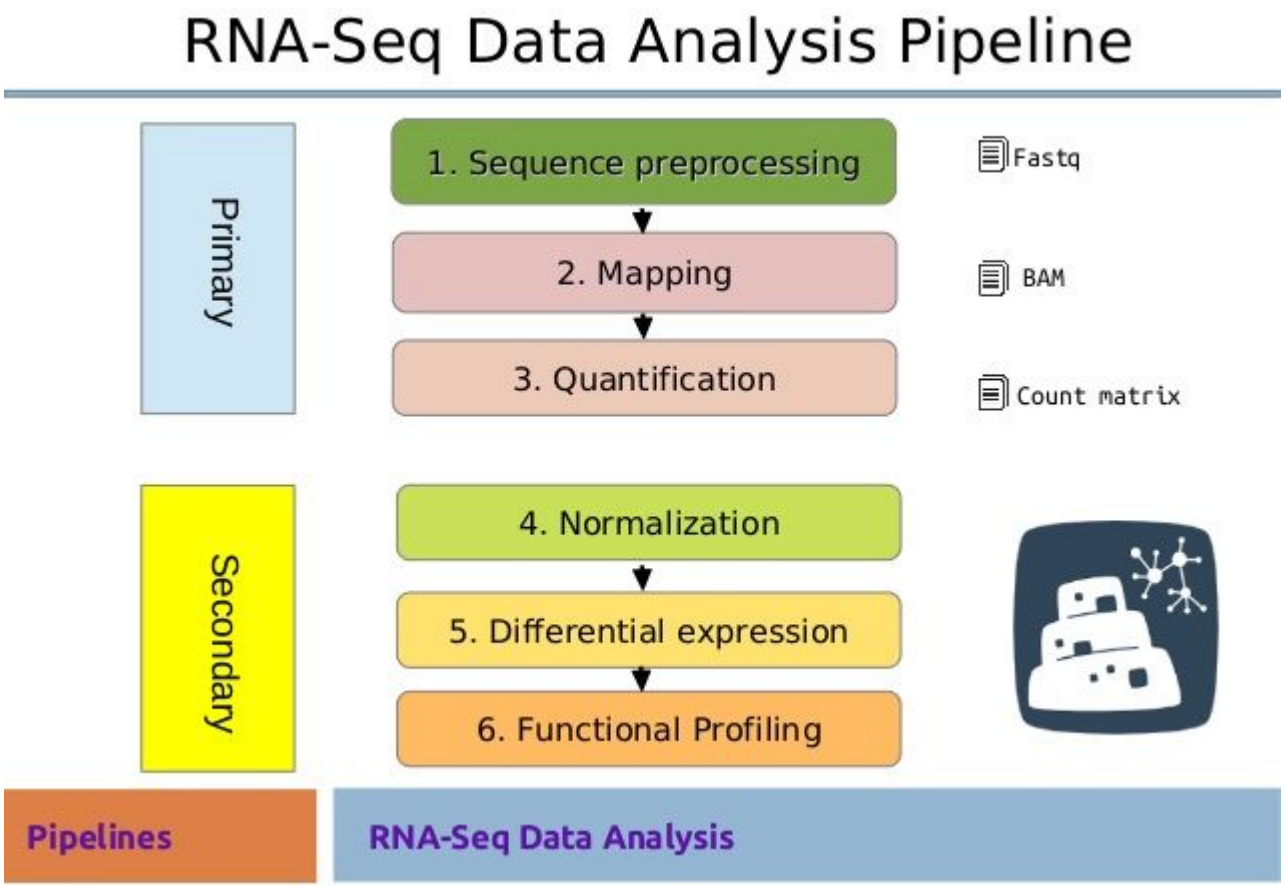
There are many different normalisation methods, some more mature than others, depending on context - no single method for merging datasets.

	sample1	sample2	sample3	sample4
gene1	2	4	0	8
gene2	5	10	0	0
gene3	0	0	1	0

Microarray vs bulk vs single cell RNA Seq

	Microarray	Bulk RNASeq	Single cell RNASeq
File Size (largest we have seen)	1GB	1TB	5TB+
Genes	Some genes*	All genes	Some genes**
How	*Probes that measure genes are created by humans. Probes light up when complementary RNA found	Fragment the RNA and sequence these as reads. Map the reads to some genome.	**Same as bulk RNASeq, but very lowly expressing genes can't pool together so may not be detected.
Data pre-processing difficulty	Easy	Average (mature)	Bigger, more complex and immature
Data analysis difficulty	Easy	Average (mature)	Very involved + immature
Row id (data file)	Probe (map to gene)	Gene	Gene

FASTQ (raw) to count file(small)



https://www.slideshare.net/YaoyuEWang/rnaseq-analysis-ngsapplication2

Acknowledgements

UoM

Christine Wells

Rowland Mosbergen

Tyrone Chen

Isha Nagpal

Sadia Waleem

Huan Wang

Jarny Choi

Elizabeth Mason

Chris Pacheco Rivera

UoM (Melbourne Integrative Genomics)

Kim-Anh Lê Cao

AIBN (UQ)

Othmar Korn

Ariane Mora

Steve Englart

Travelling

Florian Rohart