THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Australian Institute for Bioengineering and Nanotechnology

Gene Expression – Part 1 Technologies for Transcriptomics

Associate Prof Jess Mar

Australian Institute for Bioengineering & Nanotechnology Level 4 West

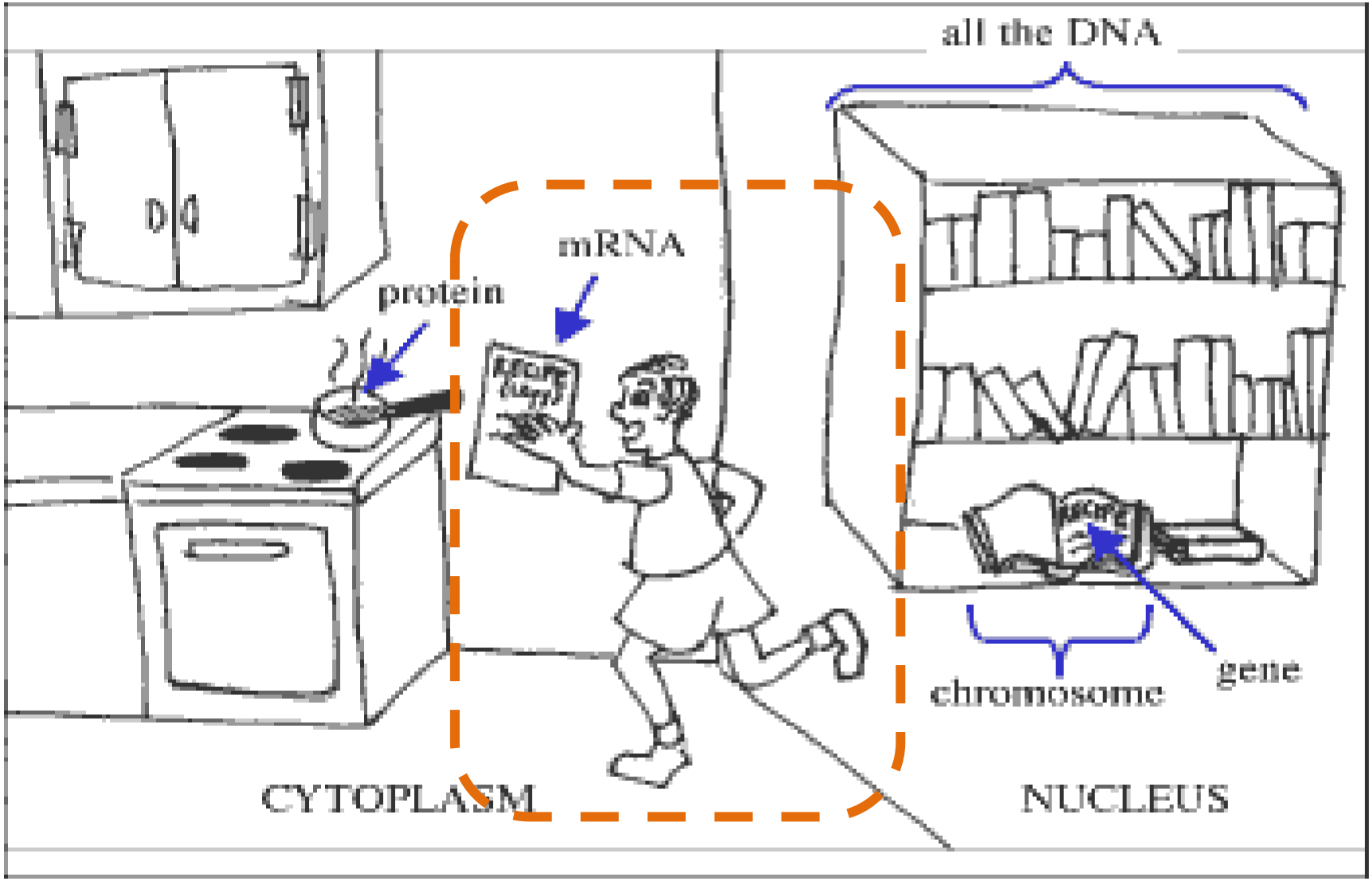
jmar@uq.edu.au

https://aibn.uq.edu.au/mar

@jessicacmar

SCIE2100/BINF6000 – Semester 1, 2021

The *Central Dogma of Molecular Biology* has guided our interpretation of the genome

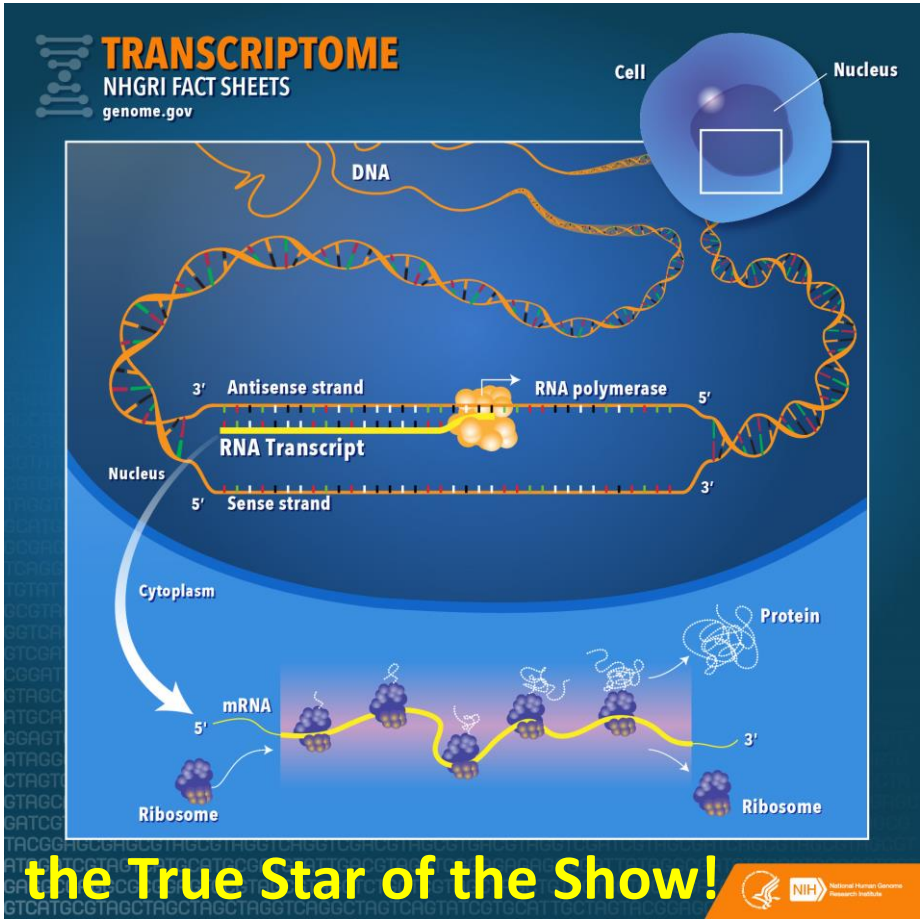


Source: http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

Introducing the Transcriptome

Definition: the transcriptome is defined as the entire set of transcripts produced inside a cell which is subject to a developmental stage or physiological condition.

- Transcripts represents RNA molecules including **mRNA, rRNA, tRNA, and other non-coding RNA like lncRNA & microRNA**.
- Genes are expressed or transcribed in response to a specific **cell stage** or **physiological condition**.
- Understanding transcriptional changes represents a window into figuring out the functional rules of the genome.



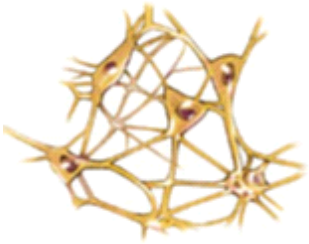
NHGRI Transcriptome Fact Sheet

The human body is made up of different kinds of cells

Within a single human, over 200 highly-specialized cell types exist!

Fertilized Egg

Neural Cells



Pluripotent Stem Cells

Cardiac Muscle

Almost all cell types share the same genome.

Epigenetic modification and **transcription factor networks** generate the mechanism for cell type-specific diversity.

A cell type's unique properties are regulated by its transcriptional program.

Understanding DNA variation is important but not the complete picture

In the context of crime, cats, or anything else (equally) important!

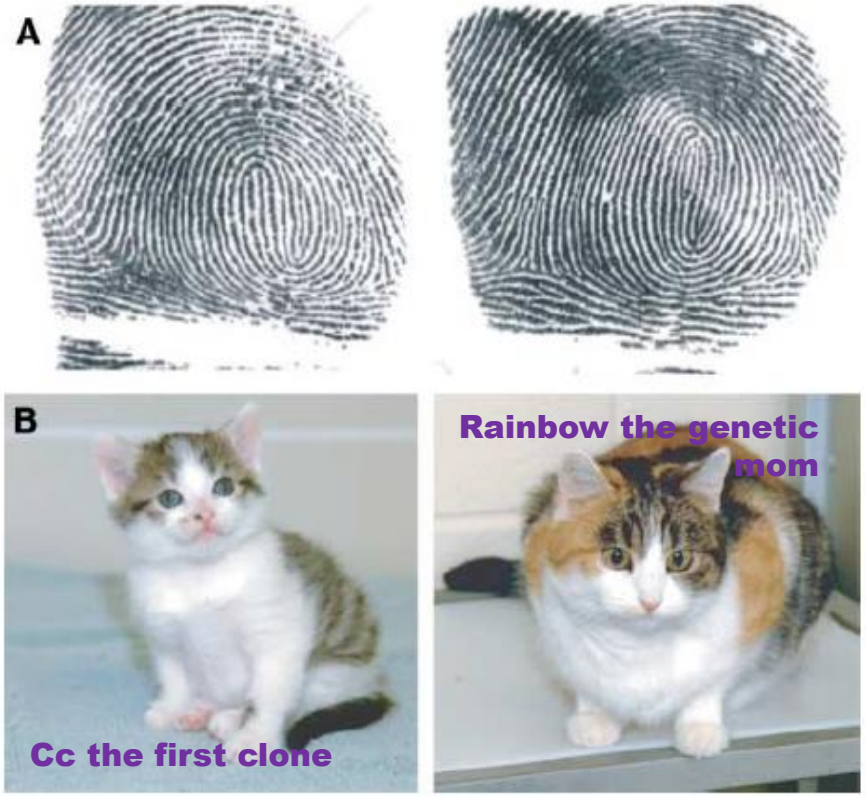
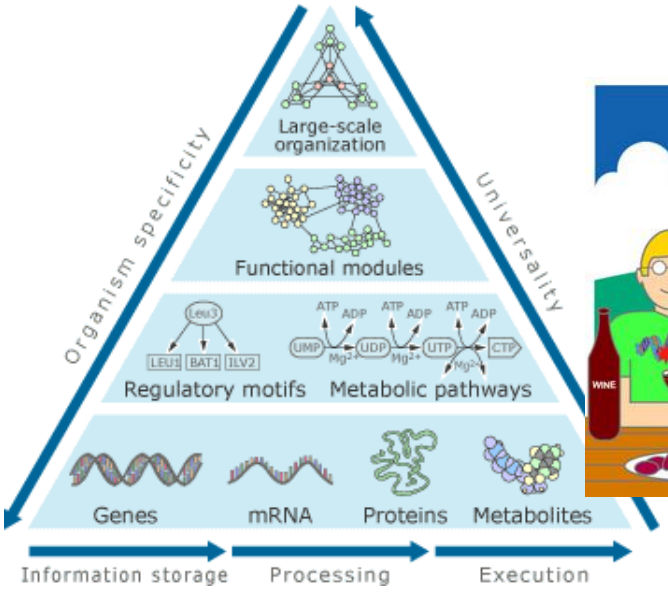


Fig. 1. Examples of possible stochastic influences on phenotype. (A) The fingerprints of identical twins are readily distinguished on close examination. Reprinted from (37) with permission from Elsevier. (B) Cc, the first cloned cat (left) and Rainbow, Cc's genetic mother (right), display different coat patterns and personalities (38). Photo credit, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University.

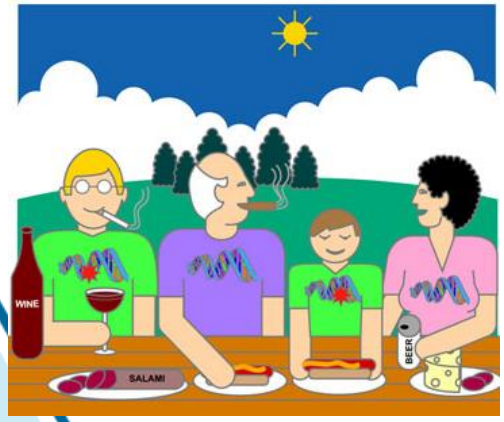
Raser & O'Shea (2005). Noise in Gene Expression: Origins, Consequences, and Control. Science.

Biology is a complex system

The more we sequence, the more we recognize that there are **MANY** factors that contribute to how biology is regulated!



Interpreting multi-layer signals.



Gene x gene
Gene x environment
…

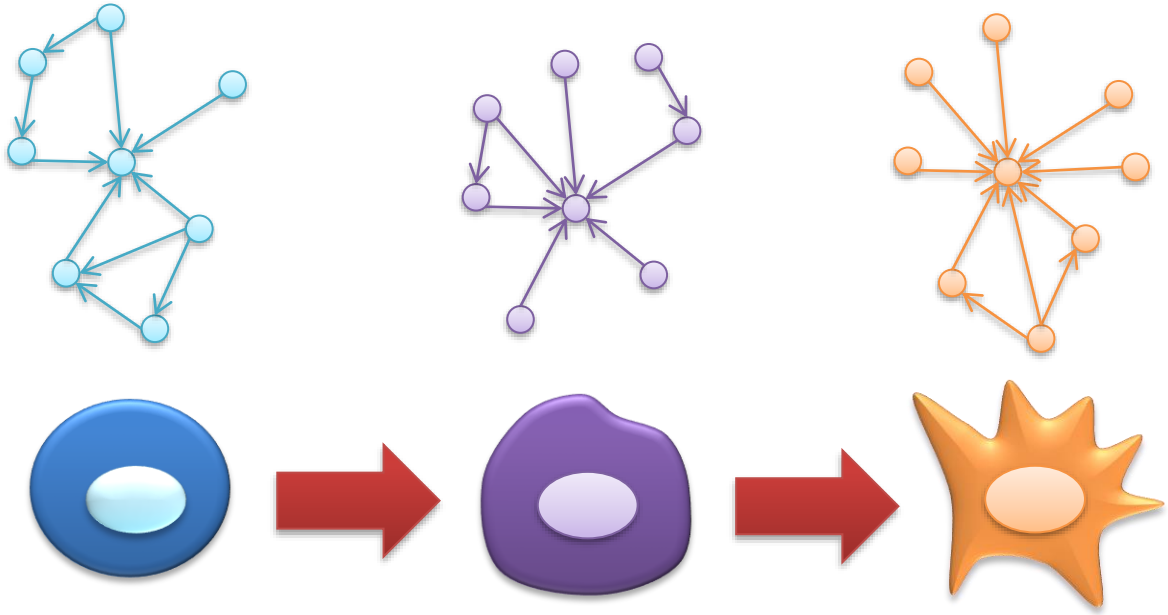
The goal of systems biology is to make sense of this complexity by developing predictive models and generating hypotheses that can be experimentally validated.

Cellular phenotypes are controlled by gene regulatory networks

Genes work together in coordinated context-specific interactions that regulate a cellular phenotype.

Networks are graphical (graph + probability) models that capture our knowledge of which interactions define a phenotype.

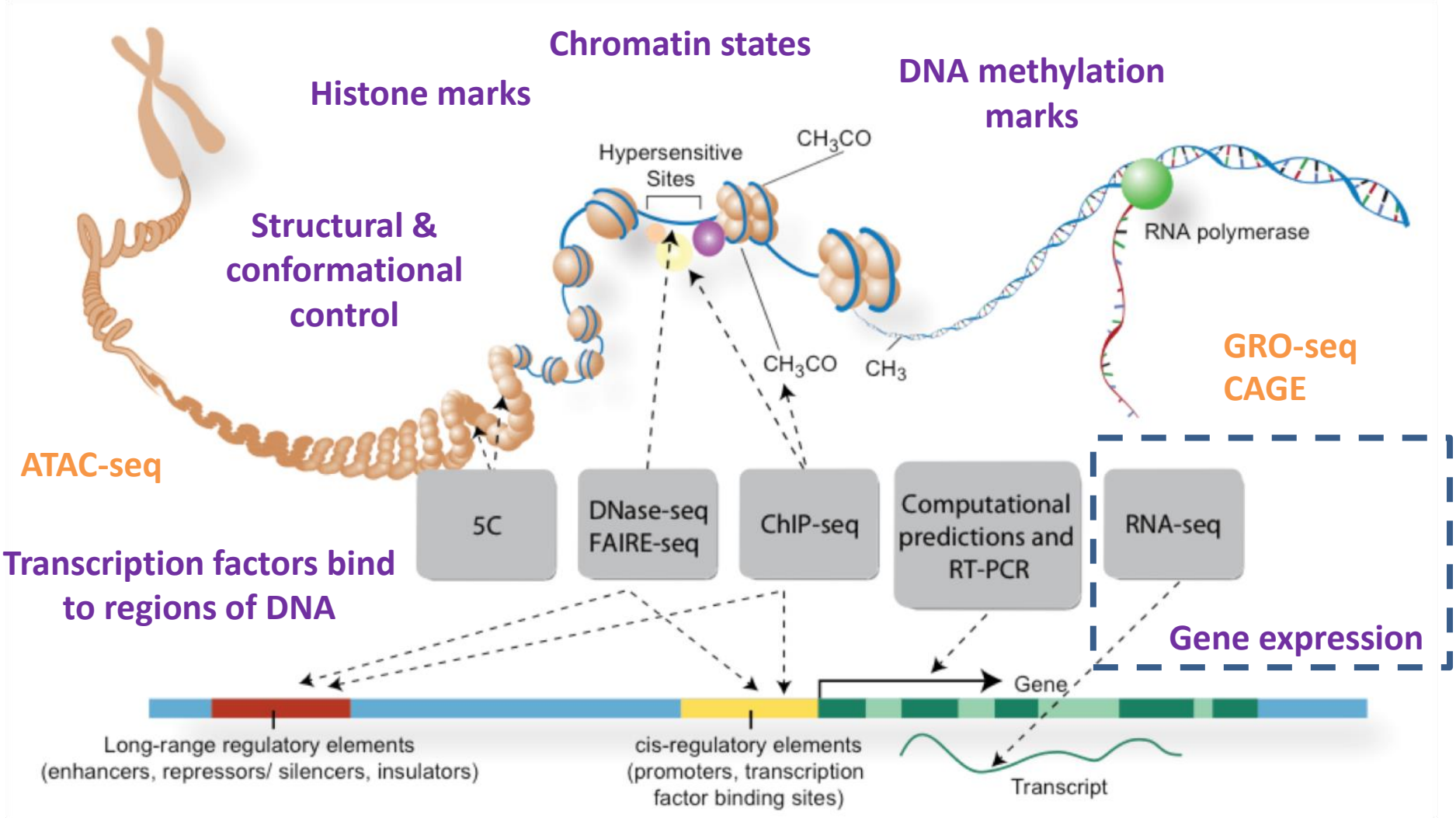
Regulatory networks



totipotent cell ***pluripotent cell*** ***lineage-committed cell***

Our job as bioinformaticians is to integrate datasets to infer what these network models look like.

Improvements in technology continue to show that the genome is a very complex entity!

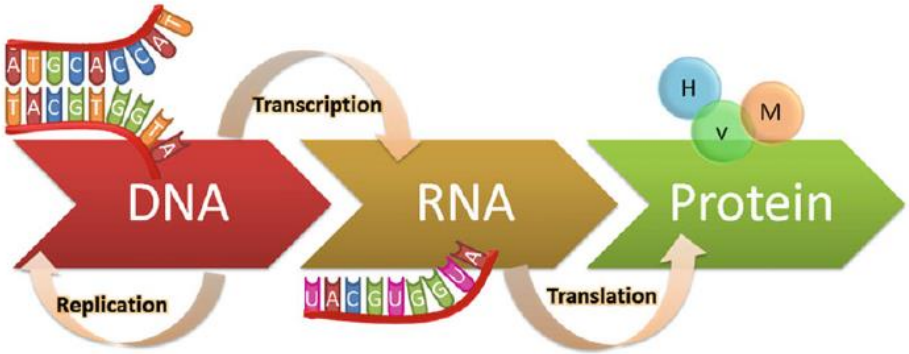


ENCODE Project https://www.encodeproject.org/

The rationale for studying RNA

- RNA is another biological molecule, just like DNA.
- A protein-coding gene is made into a transcript (e.g. mRNA) before being turned into a protein inside a cell.
- Transcripts are copies of a gene, just like photocopies of a document.
- Cells order mRNA copies of a gene to be made when that gene is required for a specific task or purpose.
- The more copies (number of mRNAs), the more this gene is in demand!

Studying mRNA counts is informative for understanding how biological processes are regulated.



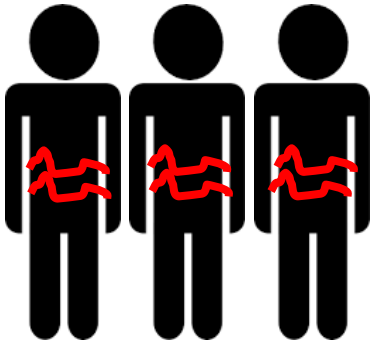
Overview of Gene Expression Analysis: Transcriptomics (2018).

Measuring DNA variants produces binary or categorical types of data

Example: Lactose intolerance is a condition where people are unable to digest lactose (a sugar found in milk and dairy products).

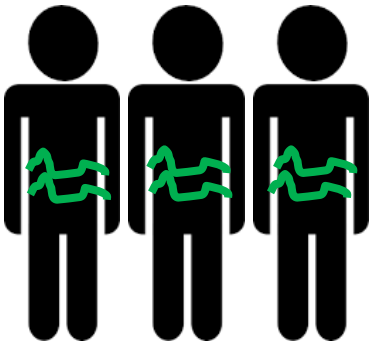
Lactose is broken up by an enzyme called lactase.

The LCT gene is responsible for making lactase.

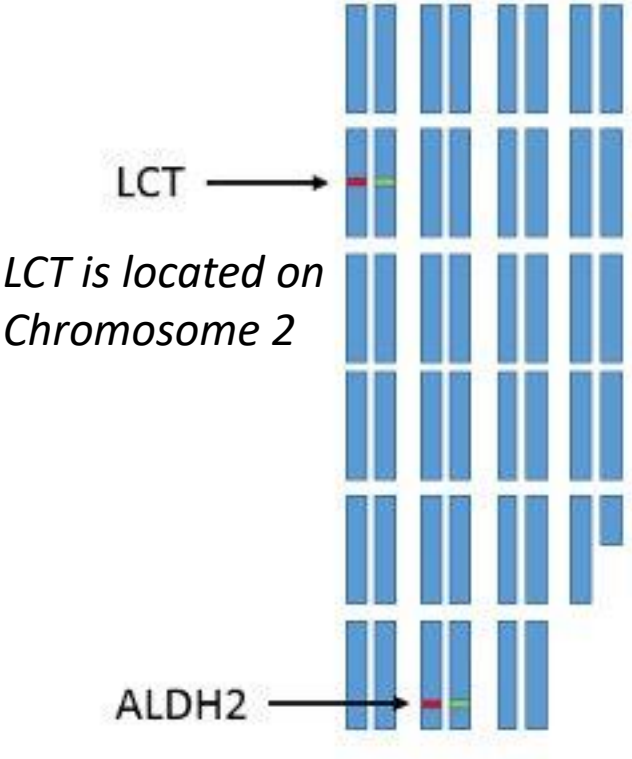


Homozygous Recessive (inherited lactose intolerance)

vs



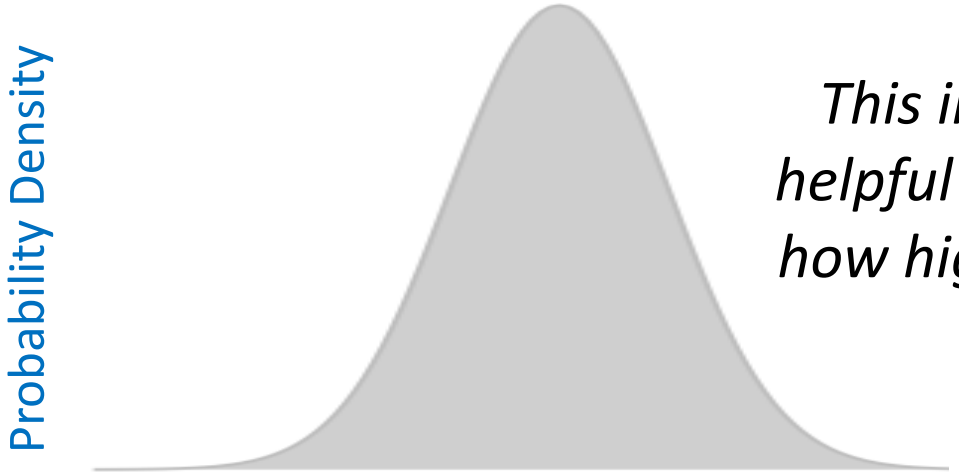
Heterozygous (usually no symptoms)



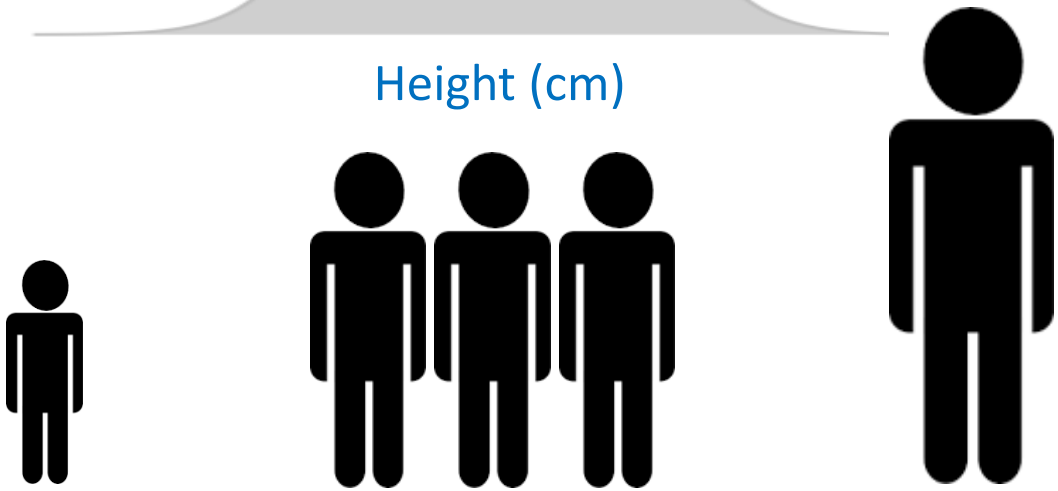
LCT is located on Chromosome 2

Measuring certain biometric variables produces continuous types of data

What is the average height of the student population at the University of Queensland?



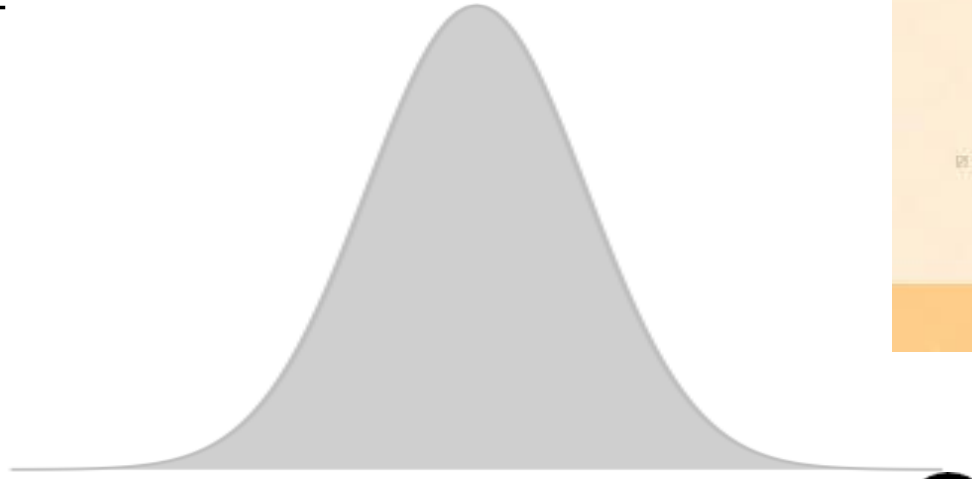
This information is helpful when deciding how high to make the doors!



Gene expression is a continuous (count) variable

Continuous data necessitates a different set of statistical analysis compared to binary/categorical data like DNA variation (see Part 2 next week!).

LCT gene example:



Counting the number of LCT mRNA molecules.

Some technologies for measuring gene expression

- ***One to a few genes at a time***
  - Northern blotting
  - PCR-based approaches – e.g. quantitative RT-PCR

- ***Lots of genes in one go!***
  - Fluidigm (hundreds)
  - **DNA microarrays** (thousands to tens thousands)
  - Serial Analysis of Gene Expression (tens thousands)
  - Chromium 10X sequencing (thousands)

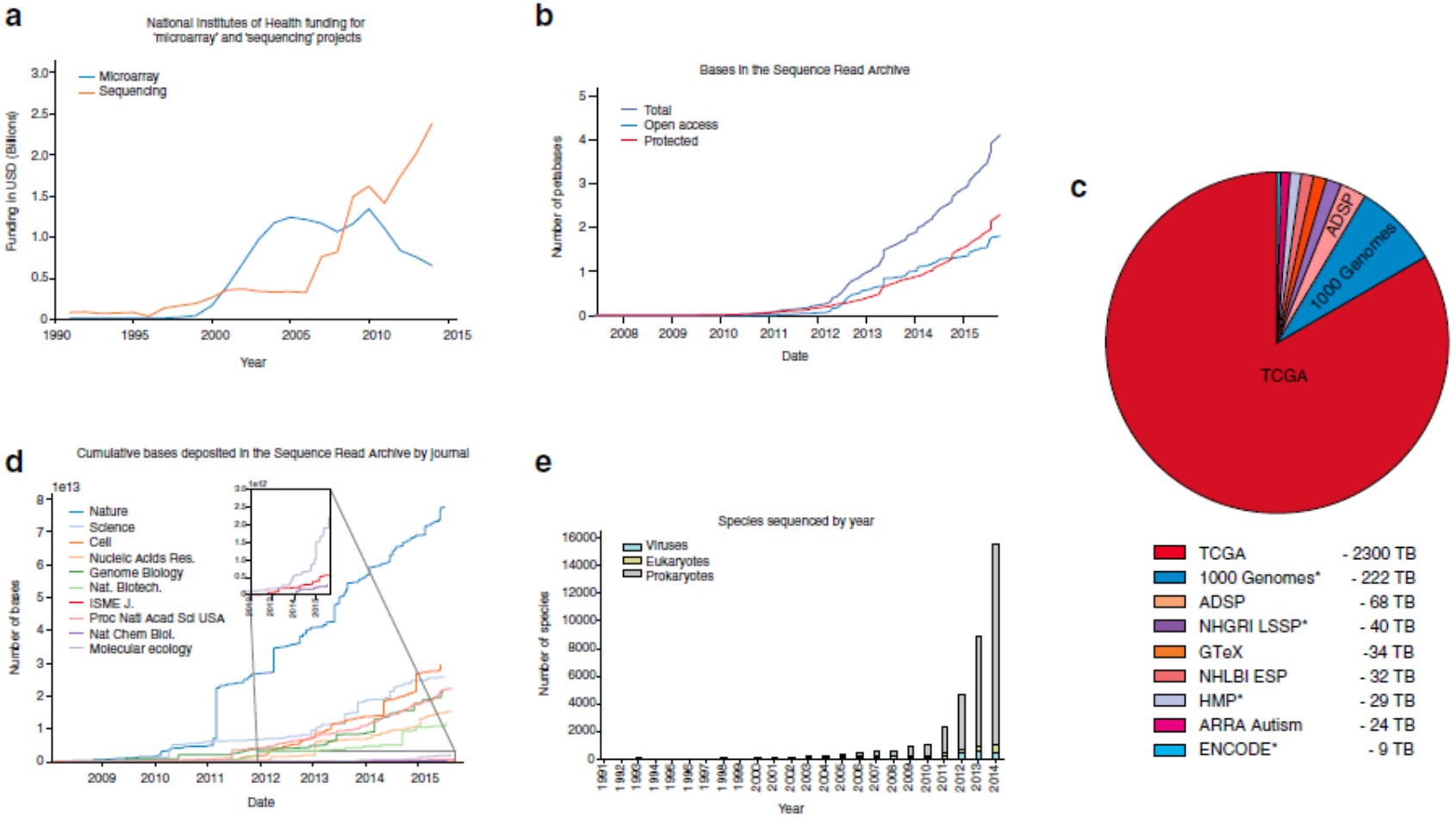
- ***All genes in one go!***
  - **RNA-sequencing**
  - Capped Analysis of Gene Expression (CAGE)

Questions Addressed Today (& Next Week)

- What are the most common platforms for collecting high-throughput gene expression data?
- What are the key steps in analyzing RNA-sequencing and microarray data?
- How can we learn about biology through analyzing gene expression data?

Let's start with technology

Sequencing technologies are the driving force behind the big data revolution



Muir et al. **(2016).** *Genome Biology.* The real cost of sequencing: scaling computation to keep pace with data generation.

RNA-sequencing is the heavy-weight amongst transcriptomics methods

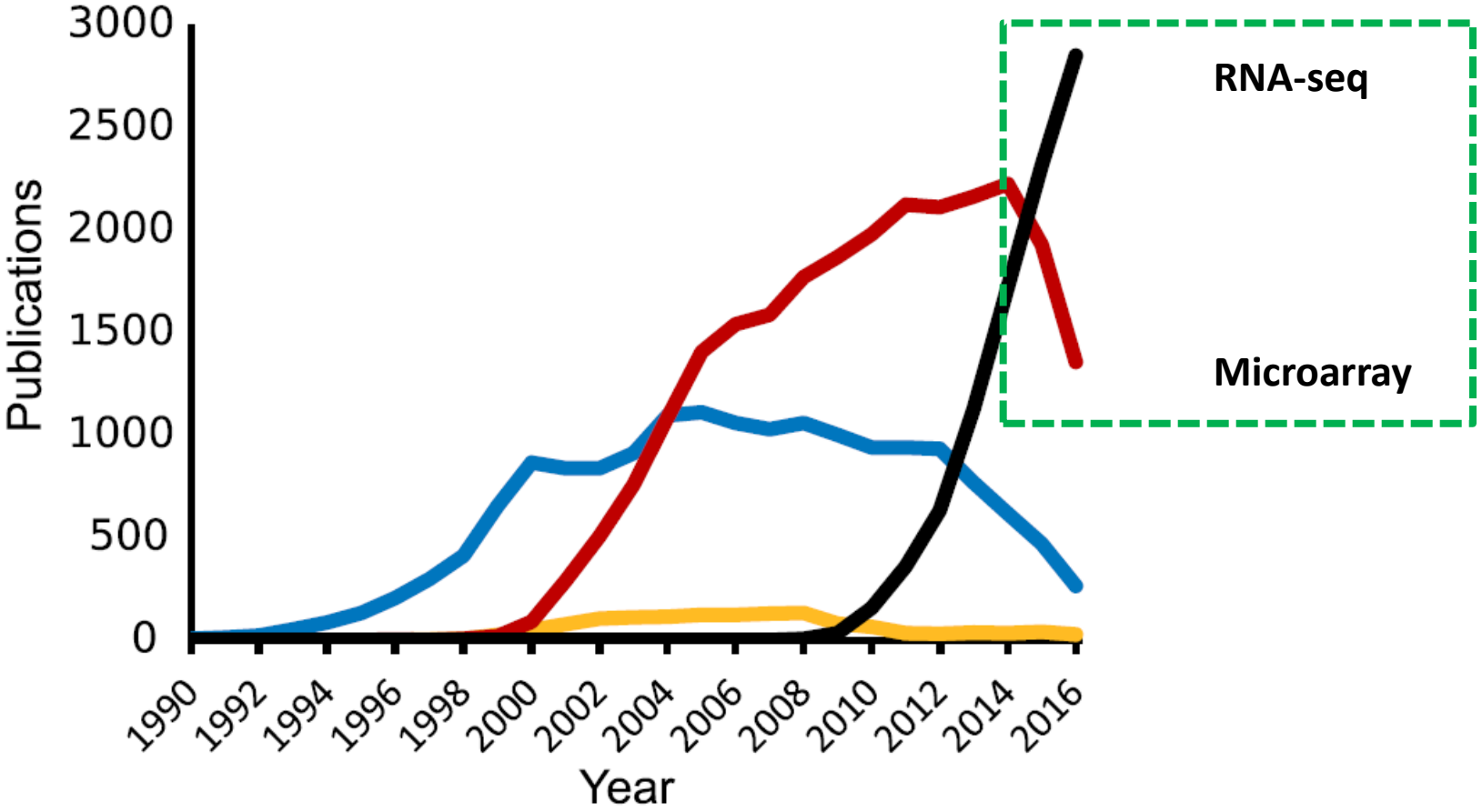
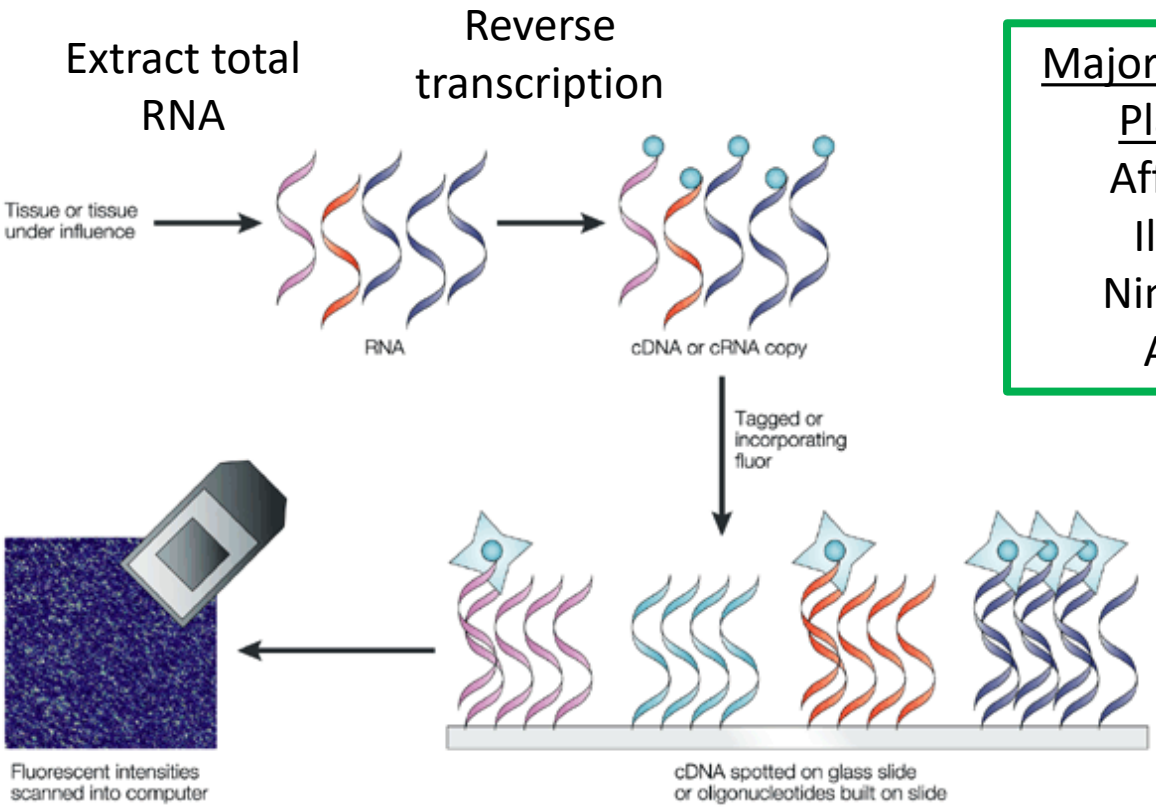


Fig 1. Transcriptomics method use over time. Published papers since 1990, referring to RNA sequencing (black), RNA microarray (red), expressed sequence tag (blue), and serial/cap analysis of gene expression (yellow)[12].

https://doi.org/10.1371/journal.pcbi.1005457.g001

Lowe et al. (2017). *PLoS Computational Biology.* Transcriptomics technologies.

But big data really started with microarrays…

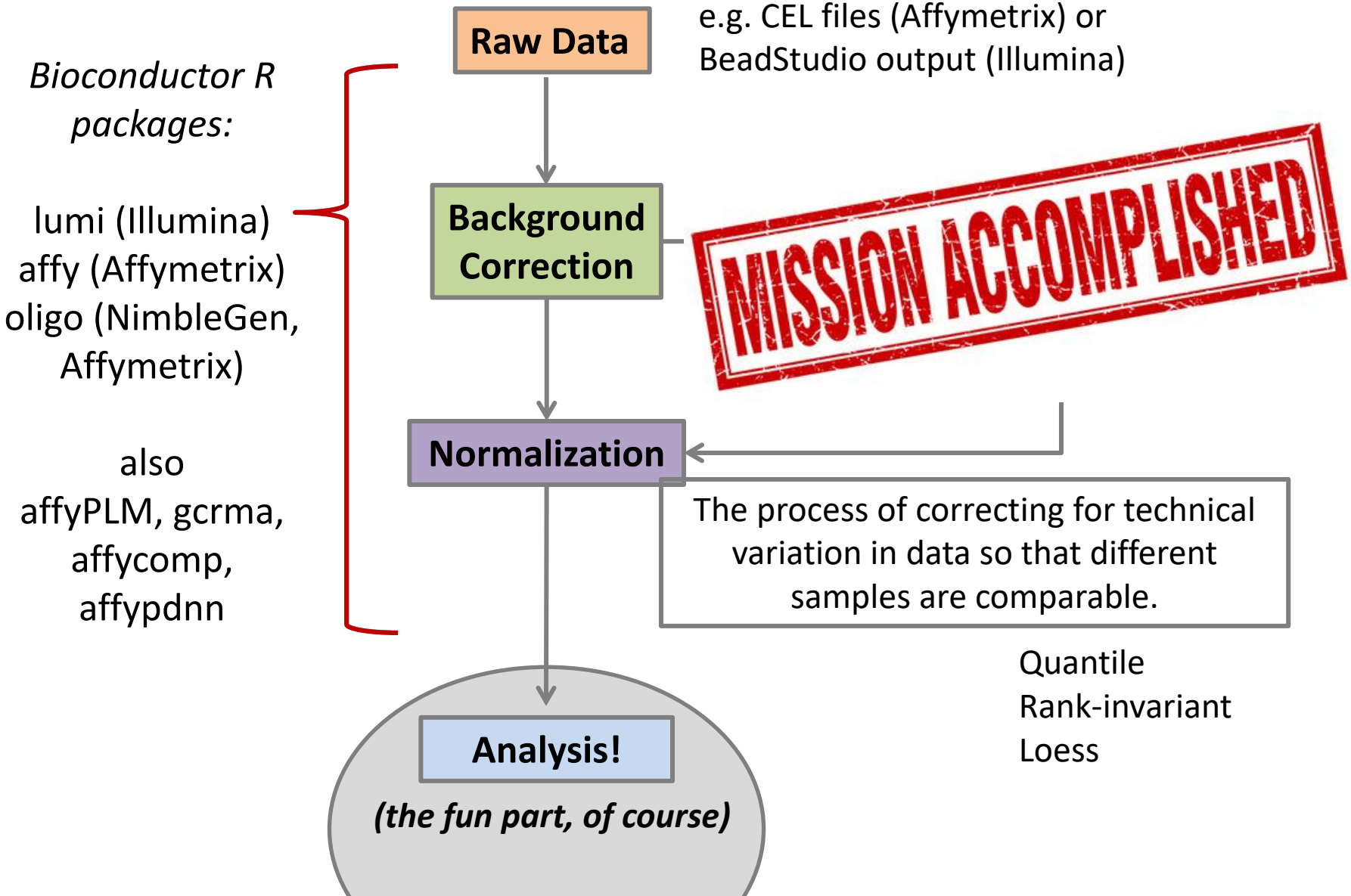


Major Microarray Platforms
Affymetrix
Illumina
NimbleGen
Agilent

Great (short) video intro!
https://www.youtube.com/watch?v=0ATUjAxNf6U

Butte. **(2002).** *Nat Rev Drug Discovery.* The use and analysis of microarray data.

Standard pipeline for microarray pre-processing



RNA-sequencing technology basics

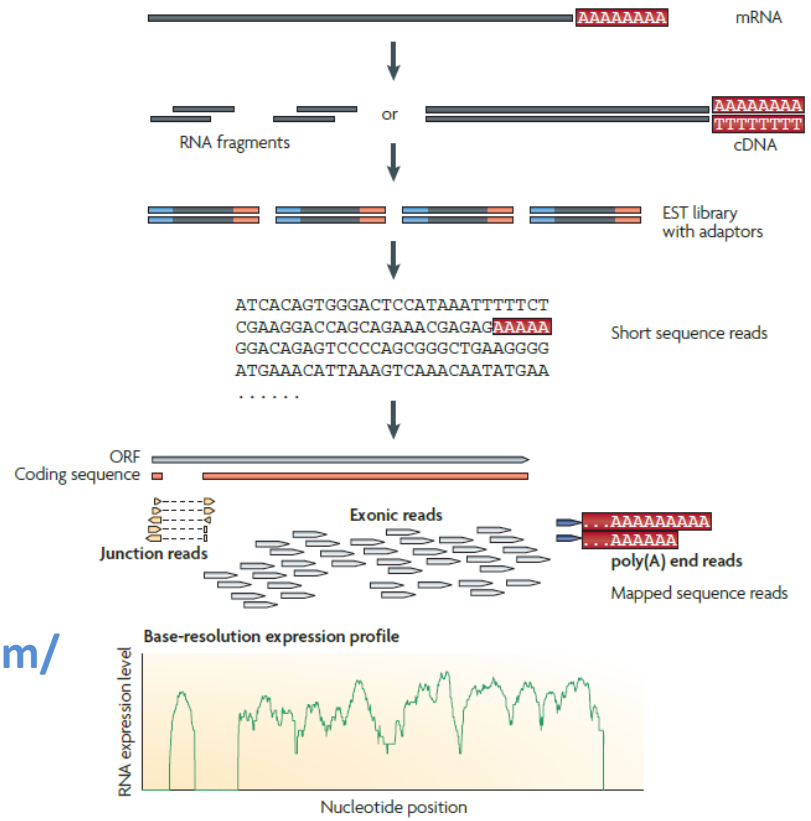
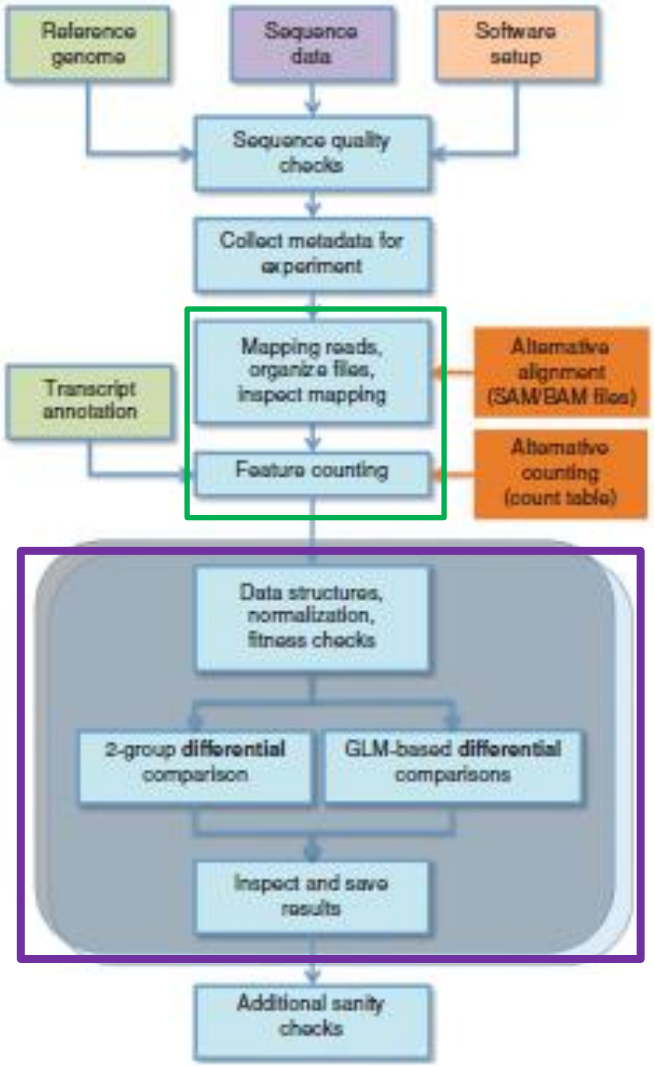


Figure 1 | **A typical RNA-Seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.

Illumina video: https://www.youtube.com/watch?v=fCd6B5HRaZ8

Wang et al. **(2009).** *Nature Reviews Genetics.* RNA-seq: a revolutionary tool for transcriptomics.

RNA-seq analysis pipeline (simplified version)



Bioconductor R packages:
edgeR
DESeq, DESeq2

Specialized Platforms:
monocle (single cell)

Wang et al. **(2009).** *Nature Reviews Genetics.* RNA-seq: a revolutionary tool for transcriptomics.

RNA-sequencing: breaking down the steps!

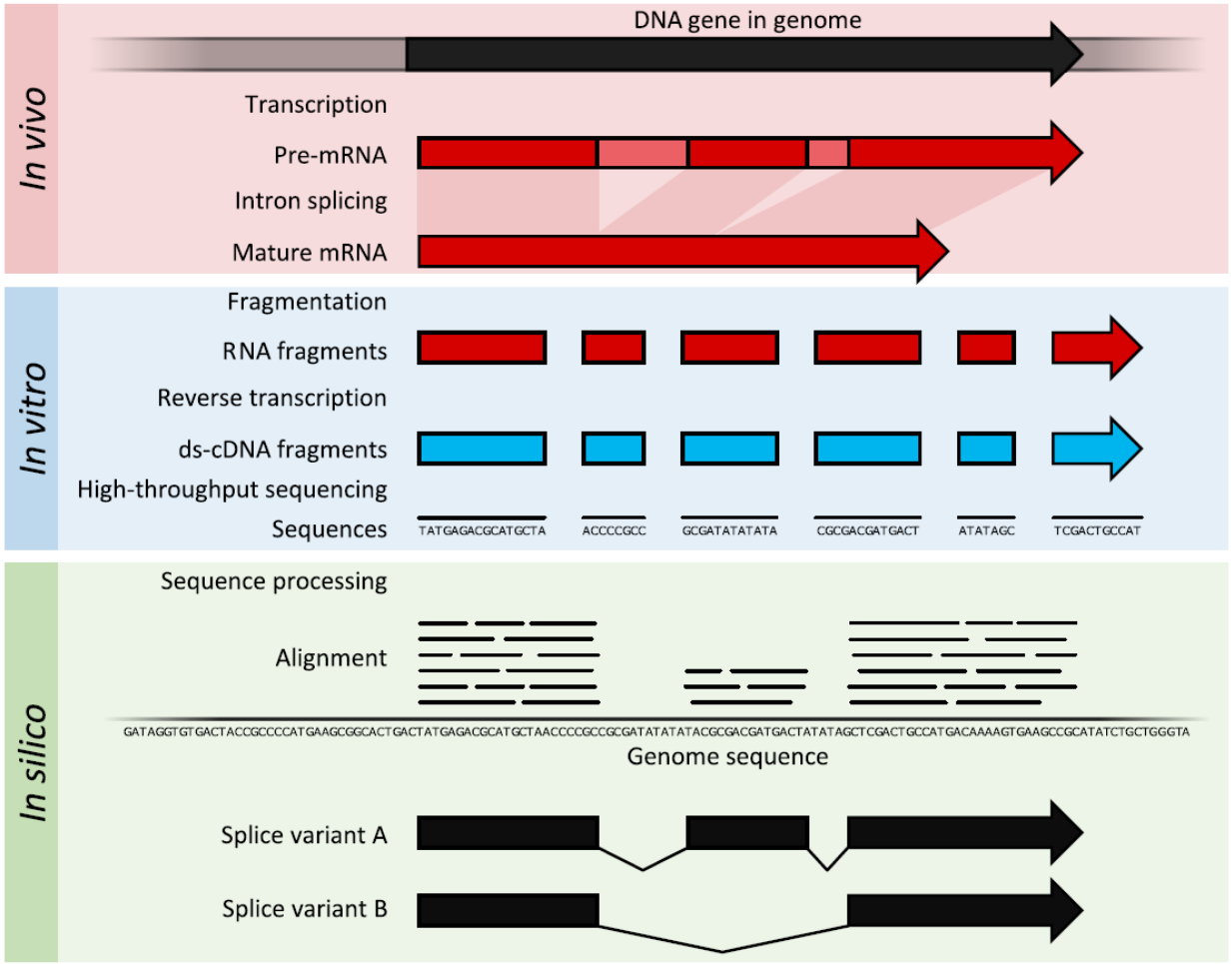


Fig 4. Summary of RNA sequencing. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable double-stranded–cDNA (ds-cDNA; blue). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. These data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.

https://doi.org/10.1371/journal.pcbi.1005457.g004

Lowe et al. **(2017).** *PLoS Computational Biology.* Transcriptomics technologies.

Microarray vs RNA-sequencing

Table 1. Comparison of contemporary methods [23] [24] [10].

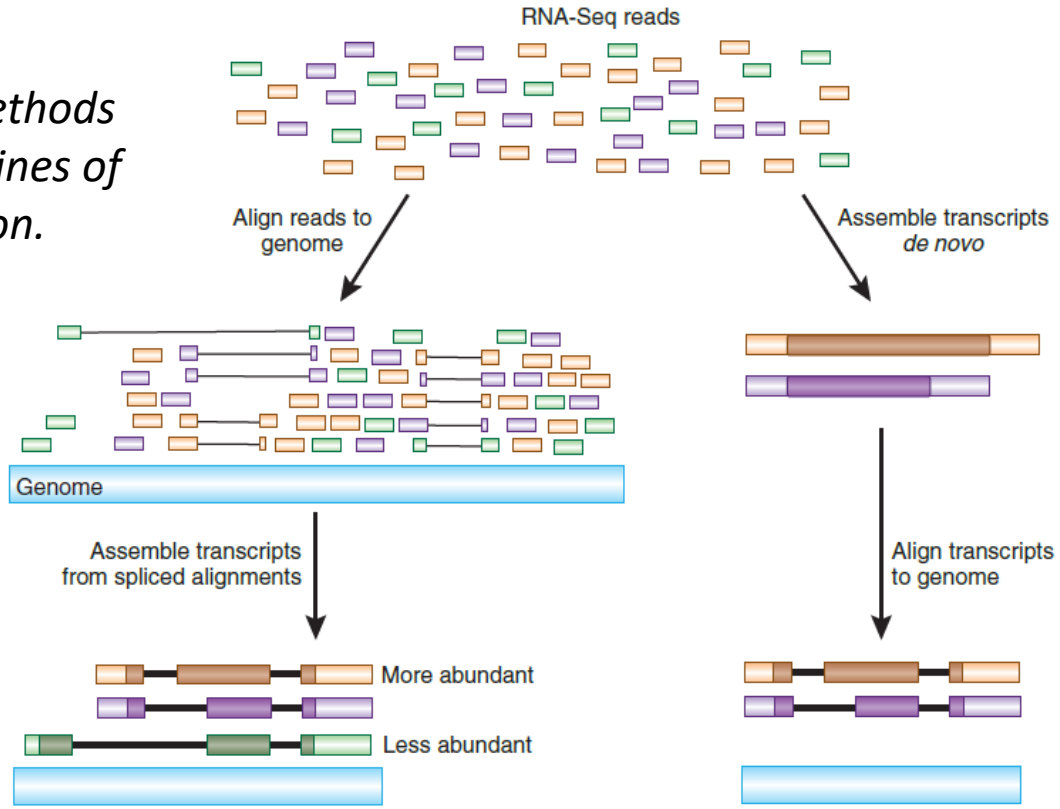
Method	RNA-Seq	Microarray
Throughput	High [10]	Higher [10]
Input RNA amount	Low ~ 1 ng total RNA [25]	High ~ 1 µg mRNA [26]
Labour intensity	High (sample preparation and data analysis) [10][23]	Low [10][23]
Prior knowledge	None required, though genome sequence useful [23]	Reference transcripts required for probes [23]
Quantitation accuracy	~90% (limited by sequence coverage) [27]	>90% (limited by fluorescence detection accuracy) [27]
Sequence resolution	Can detect SNPs and splice variants (limited by sequencing accuracy of ~99%) [27]	Dedicated arrays can detect splice variants (limited by probe design and cross-hybridisation) [27]
Sensitivity	10^{-6} (limited by sequence coverage) [27]	10^{-3} (limited by fluorescence detection) [27]
Dynamic range	$>10^{5}$ (limited by sequence coverage) [28]	10^{3}–10^{4} (limited by fluorescence saturation) [28]
Technical reproducibility	>99% [29][30]	>99% [31][32]

- The size of the advantage depends on the ***biological question*** being asked!
- In general, key advantages (Table 1) have driven the uptick in RNA-seq versus microarray-based applications.

Lowe et al. **(2017).** *PLoS Computational Biology.* Transcriptomics technologies.

Using RNA-seq to learn about the transcriptome

Sequencing methods allow for new lines of investigation.



Quantify differential expression, isoform abundance, non-coding RNAs, RNA editing, splicing, SNPs, etc.

Haas and Zody. **(2010).** *Nature Biotechnology.* Advancing RNA-seq analysis.

Despite the different applications, the preprocessing steps are generally similar

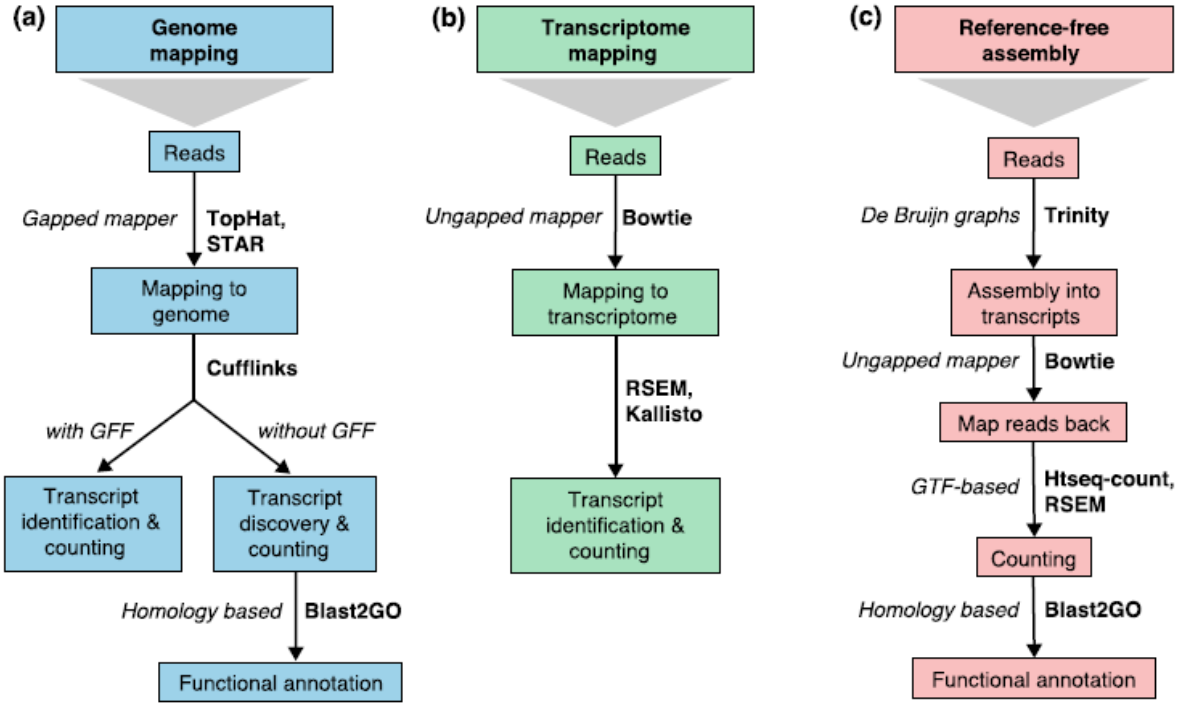


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in (**b**) followed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used at each analysis step are indicated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

Conesa et al. **(2011).** *Genome Biology.* A survey of best practices for RNA-seq data analysis.

Technology platforms for next-generation sequencing

Table 2. Sequencing technology platforms commonly used for RNA-Seq [72][73].

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

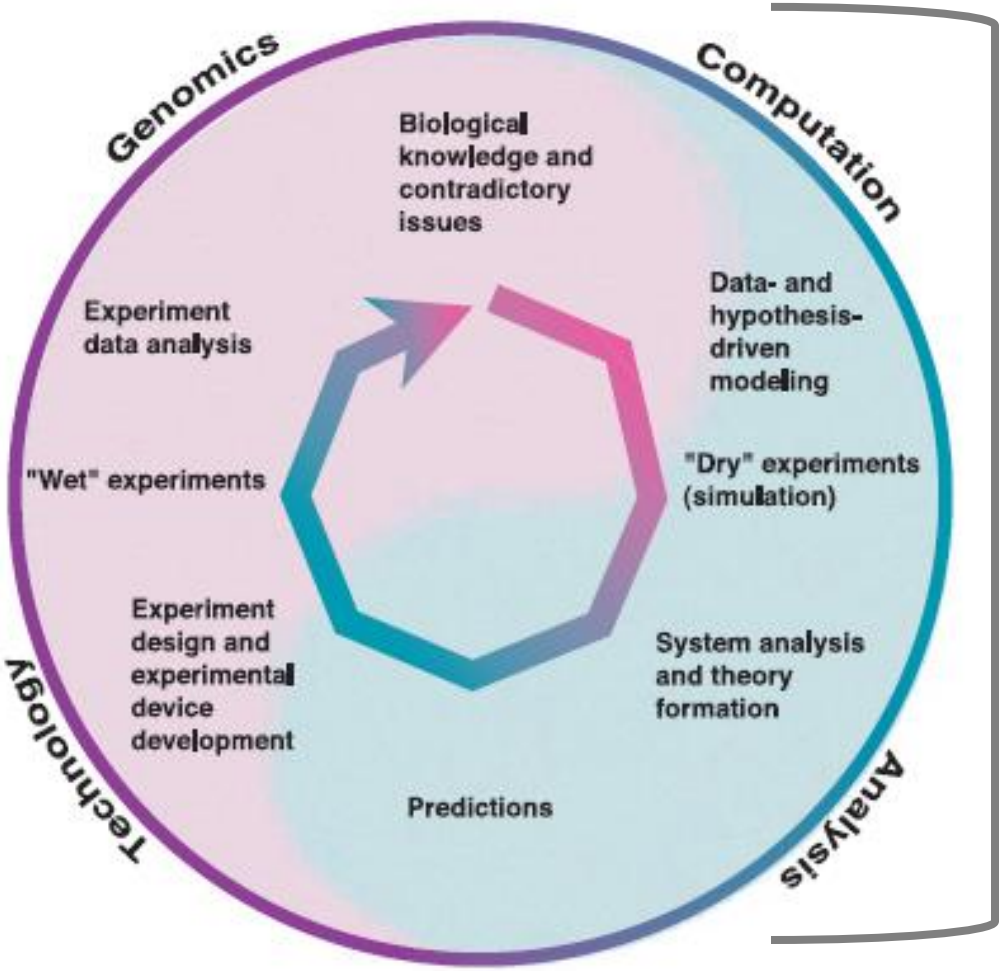
https://doi.org/10.1371/journal.pcbi.1005457.t002

New platforms and technologies continue to be released at rates that no publication can capture accurately.

Lowe et al. **(2017).** *PLoS Computational Biology.* Transcriptomics technologies.

Resources to study gene expression
–
what's in this for you?

How systems biology fits into the general scientific method

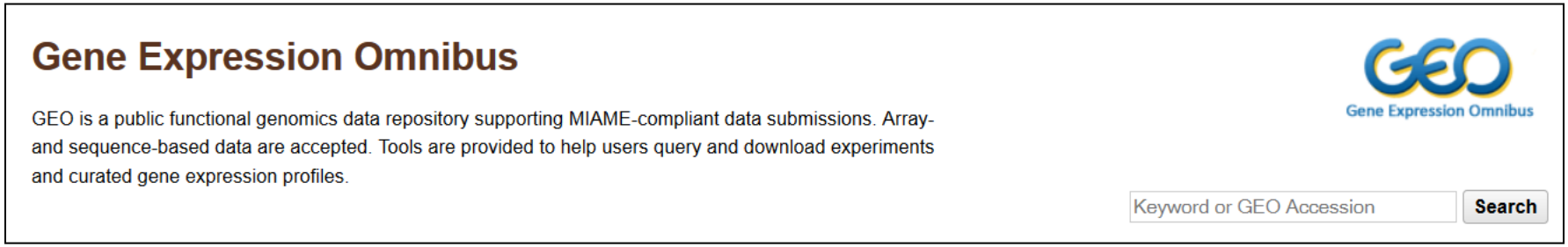


1. Identify a testable hypothesis.
2. Identify the data set(s) or experimental design under which this hypothesis can be tested.
3. Conduct statistical analysis.
4. Interpret results and acknowledge limitations.

Kitano, Science (2002): vol. 295 no. 5560.

Got questions? Ask the data!

A wealth of gene expression data is publicly available in the form of repositories, the largest being GEO, and is hosted by the NCBI at the NIH.



- All expression data stored in GEO follows a standard format called MIAME which was created to ensure the experimental details were available.
- Easy links to the paper that describes the study that generated the data.
- Data can be downloaded in normalized (good to go!), or raw formats.

NCBI has many databases that store specific genomic information.

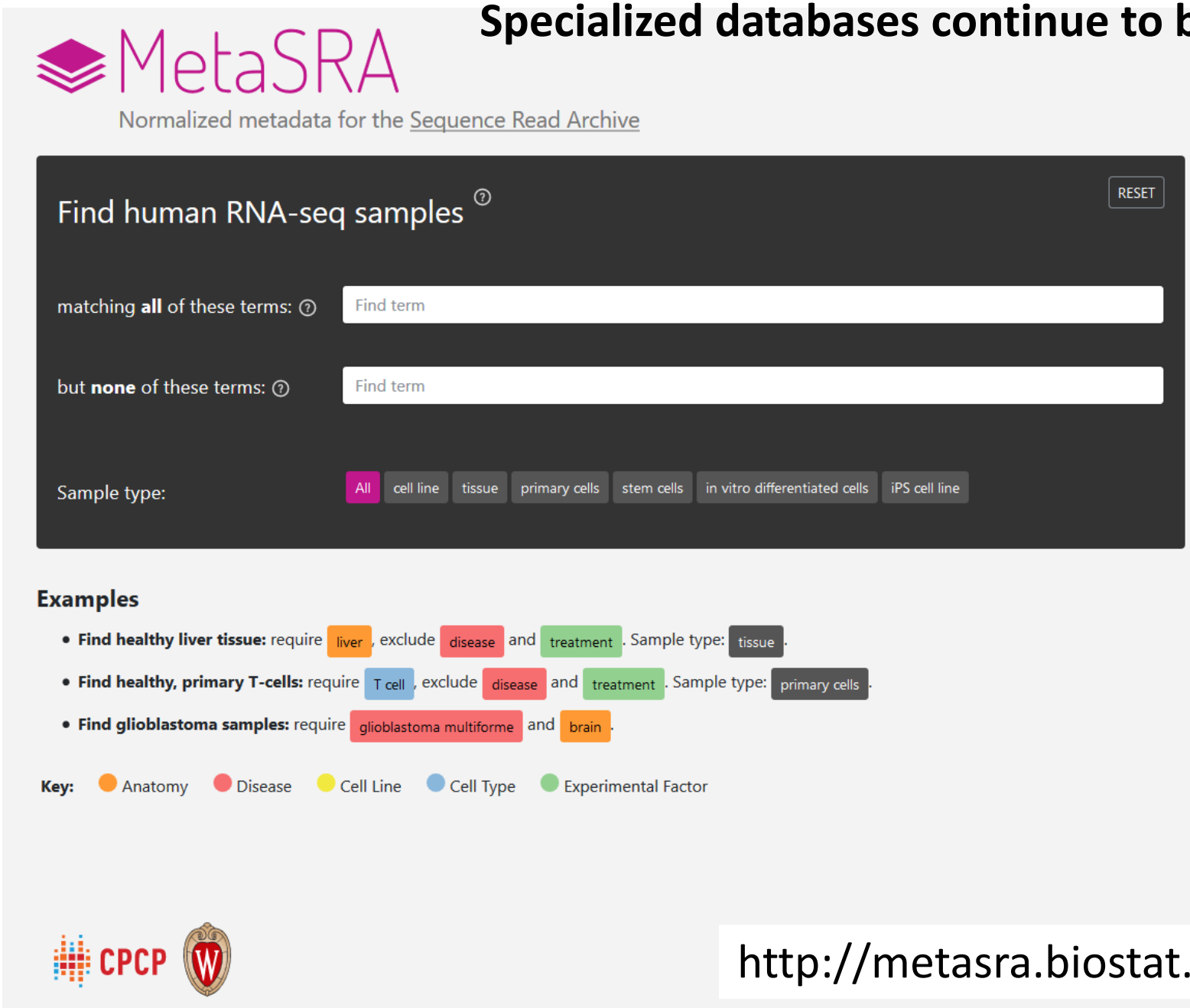
dbSNP

dbGAP: database of Genotypes and Phenotypes

SRA (Short Read Archive)

http://www.ncbi.nlm.nih.gov/

Specialized databases continue to be added…



CPCP

http://metasra.biostat.wisc.edu/

Social Genomics – Loneliness, Happiness and Science?!

An emerging area of social science deals with the intersection of happiness/loneliness and the impact on human health. More recently, this field has taken a quantitative molecular approach, giving rise to **“social genomics”**.

Loneliness Is Bad For You, And This Study May Help Explain Why

Feeling lonely may trigger changes in our cells that could make us more susceptible to illness.

11/28/2015 08:53 am ET

Jacqueline Howard
Senior Science Editor, The Huffington Post

EVGENIASH VIA GETTY IMAGES

Loneliness can affect the production of white blood cells in our bodies, study shows.

Forbes / Pharma & Healthcare

NOV 24, 2015 @ 08:00 AM 15,913 VIEWS

Loneliness Destroys Physical Health From The Inside Out

David DiSalvo
CONTRIBUTOR

I write about science, technology and the cultural ripples of both.

FULL BIO >

Opinions expressed by Forbes Contributors are their own.

Loneliness can increase the risk of pr
study supported by the National Inst
of the potential for loneliness to dam

What the research team found is tha
strongly linked to two critical physio
immune systems and increased cellu
loneliness affects the expression of g
transcriptional response to adversity

The longer someone experiences lon
genes related to white blood cells (ak
infections) and inflammation. CTRA
simultaneously increasing the geneti
at the cellular level rather than the s
happening within the body’s cells.

The combination of the two effects is
with a slow erosion of cellular health
problems, some of which worsen ove

The study also found that CTRA and
CTRA gene expression more than a y
more than a year later. In other word

The Physical Effects Of Loneliness Include Weakened Immune Systems, Premature Death

AFP/Relaxnews

Posted: 11/24/2015 10:50 am EST | Updated: 11/24/2015 10:59 am EST

Do Social-Environmental Factors Influence Our Gene Expression?

Proponents of human social genomics argue that social-environmental conditions can affect differential expression of genes in leukocytes.

Circulating leukocytes are a mixed population of cells, responsible for pathogen recognition, immune response and tissue repair.

Adverse social conditions (loneliness, bereavement, depression) have shown to affect leukocyte gene expression.

How do social environments regulate immune function?

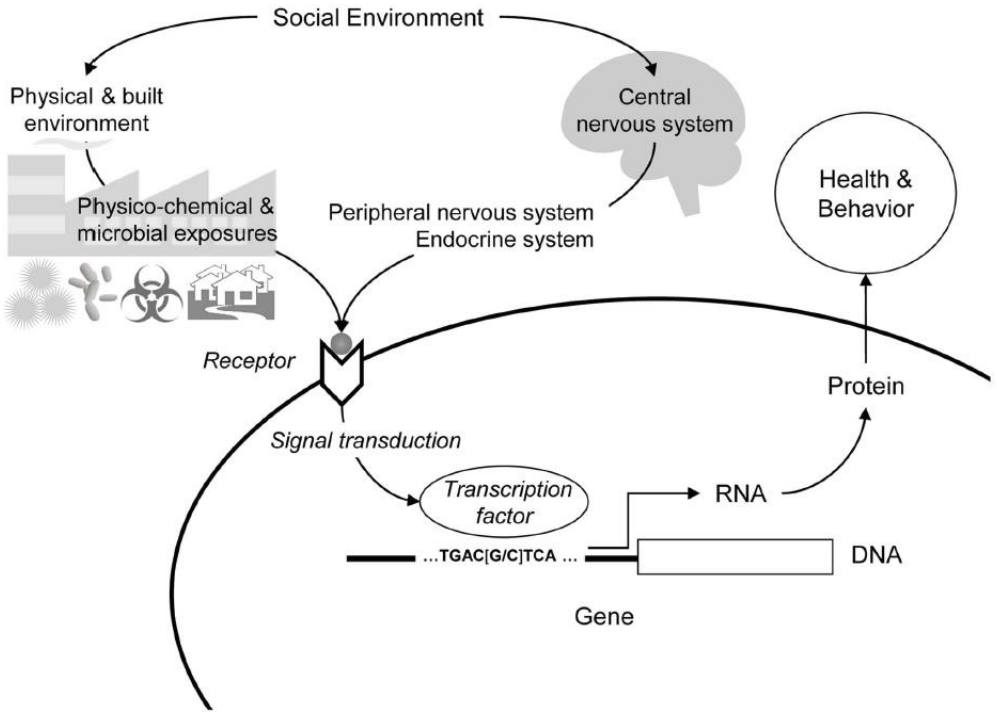


Figure 1. Social regulation of human gene expression. Social environments can influence human gene expression via physicochemical processes (e.g., toxins, pollutants, and microbes) and psychological processes (e.g., experiences of threat or uncertainty) that trigger neural and endocrine responses (e.g., activation of the sympathetic nervous system). In both cases, biochemical mediators engage cellular receptor systems, which activate intracellular signal transduction pathways culminating in the activation (or repression) of transcription factors that proximally regulate the transcription of genes bearing response elements for that particular factor. The gene regulatory "wiring diagram" that maps specific biochemical signals to specific gene expression responses represents an evolved genomic program that was presumably adaptive under ancestral conditions but may have distinct maladaptive effects in the very different social environments of contemporary human life.
doi:10.1371/journal.pgen.1004601.g001

Cole SW. (2014). Human Social Genomics. *PLoS Genetics.*

Retrieving data from GEO is straightforward

Research

Social regulation of gene expression in human leukocytes

Steve W Cole[*†‡], Louise C Hawkley[§], Jesusa M Arevalo[*], Caroline Y Sung[†], Robert M Rose[¶] and John T Cacioppo[§]

Published: 13 September 2007

Genome ***Biology*** 2007, **8:**R189 (doi:10.1186/gb-2007-8-9-r189)

Received: 2 March 2007
Revised: 30 July 2007
Accepted: 13 September 2007

Open Access

from the Methods section

Social isolation

Subjectively experienced social isolation was assessed by the UCLA-R Loneliness scale [53] at each yearly study visit. Biological samples from 10 individuals who consistently scored in the top 15% of the loneliness distribution during study years 1, 2, and 3, and 10 individuals who consistently scored in the bottom 15% during years 1, 2, and 3, were selected for analysis after matching for age, gender, and ethnicity. Two samples from low-lonely individuals and four samples from high-lonely individuals yielded insufficient RNA for reliable gene expression assay, and analyses are thus based on fourteen individuals (eight low-lonely, six high-lonely). Objective social isolation was assessed by the social network index [54].

puncture sample and isolated by ficoll density gradient centrifugation; RNAlater/RNeasy, Qiagen, Valencia, CA, USA), and 5 μg of the resulting RNA was assayed using Affymetrix U133A high-density oligonucleotide arrays [58] in the UCLA DNA Microarray Core as previously described [41,59]. Robust multiarray averaging [60] was applied to quantify expression of the 22,283 assayed transcripts, and differentially expressed genes were identified as those showing ≥30% difference in mean expression levels in samples from high- versus low-lonely individuals (corresponding to a FDR of 10%) [59]. Functional characteristics of individual genes were identified through GO annotations, Gene References into Function annotations, and PubMed literature links retrieved through NCBI Entrez-Gene [61]. Functional commonalities among multiple differentially expressed genes were identified using GOstat [62] with default stringency parameters (Benjamini FDR <0.10) [44]. A full list of differentially expressed genes is provided in Additional data file 1, and raw data are deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO Series GSE7148).

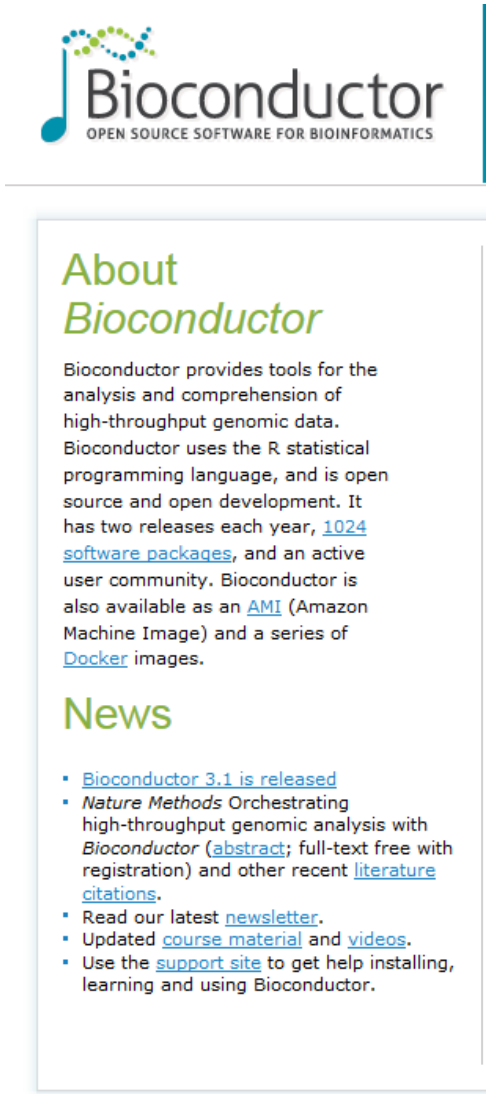
To install any Bioconductor package:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("GEOquery")
```

The Bioconductor Project – A Bioinformatics Standard

- This project has become the standard repository for R software that deals with all things **bio**.
- A big theme of Bioconductor has been the standardization of data classes to make analysis of –omic data easier, more robust and **more reproducible**.
- The project makes available packages that deal with:
  - Annotation
  - Statistical Methods
  - Pre-processing Approaches
- *Vignettes will change your life!*

http://bioconductor.org



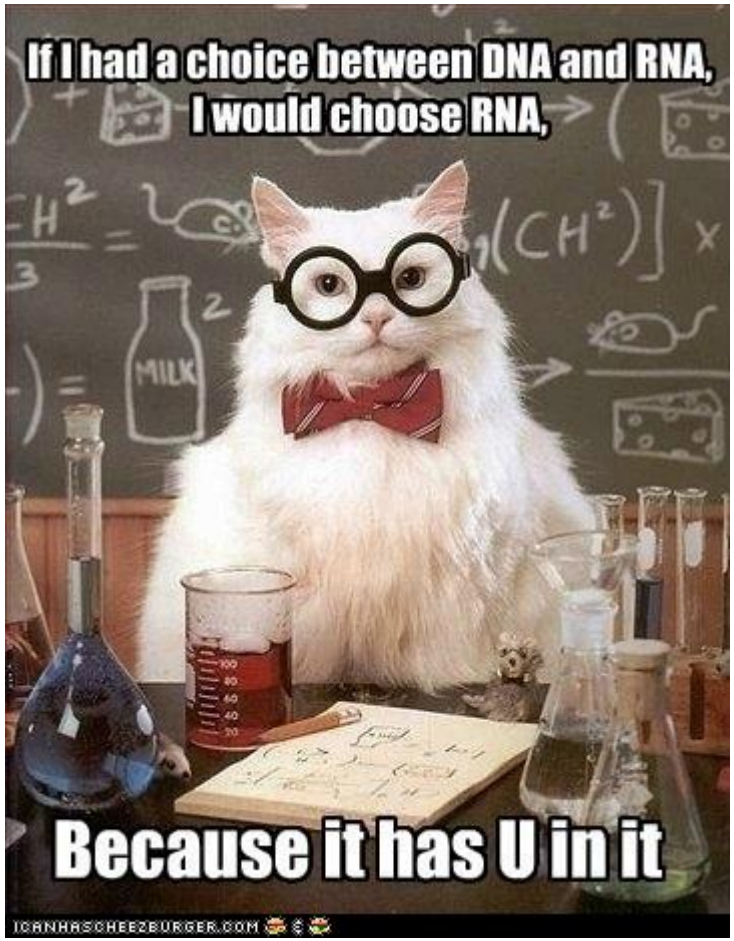
Lecture Summary

- RNA-sequencing and microarrays are generally used for high-throughput gene expression data, with the former eclipsing the latter.
- Pre-processing of RNA-seq data requires alignment of reads, transcript identification and quantification.
- RNA-seq allows us to capture information about mRNAs but also other RNA molecules.

Additional Take-aways

- RNA-seq allows us to capture information about mRNAs but also other RNA molecules.
- For a beginner, it is easier to learn how the pipelines work by starting with bulk RNA-seq data.
- Innovative methods exploit the advantages of gene expression datasets – this is a space for amazing science that you could even do from home!
- Bioinformatics-driven science is a reality.
- Opportunities exist to use publicly accessible datasets to act on your curiosity!

So, what are you curious about?



Please check out Part 2 and see you next week for our online lectorial!

AIBN (Building 75) Level 4 West
Email: j.mar@uq.edu.au

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Australian Institute for Bioengineering and Nanotechnology

Gene Expression – Part 2
Big Data Analysis

Associate Prof Jess Mar

Australian Institute for Bioengineering & Nanotechnology Level 4 West

j.mar@uq.edu.au

https://aibn.uq.edu.au/mar

@jessicacmar

SCIE2100/BINF6000 – Semester 1, 2021

Questions Addressed Today (& Last Week)

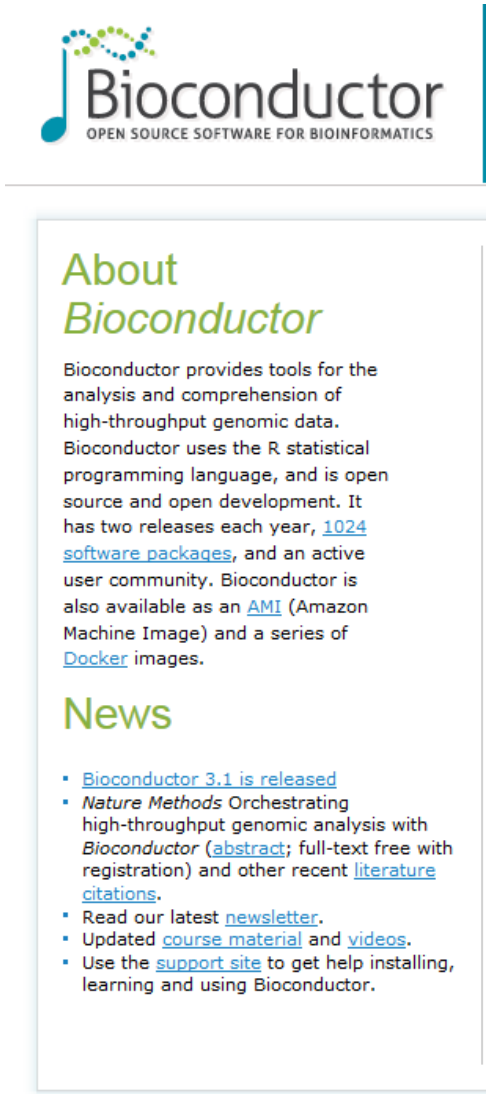
- What are the most common platforms for collecting high-throughput gene expression data?
- What are the key steps in analyzing RNA-sequencing and microarray data?
- **How can we learn about biology through analyzing gene expression data?***

***This is a whirlwind tour of some examples. Consider this a starter flight of bioinformatics analyses to pique your curiosity!**

The Bioconductor Project – A Bioinformatics Standard

- This project has become the standard repository for R software that deals with all things **bio**.
- A big theme of Bioconductor has been the standardization of data classes to make analysis of –omic data easier, more robust and **more reproducible**.
- The project makes available packages that deal with:
  - Annotation
  - Statistical Methods
  - Pre-processing Approaches
- *Vignettes will change your life!*

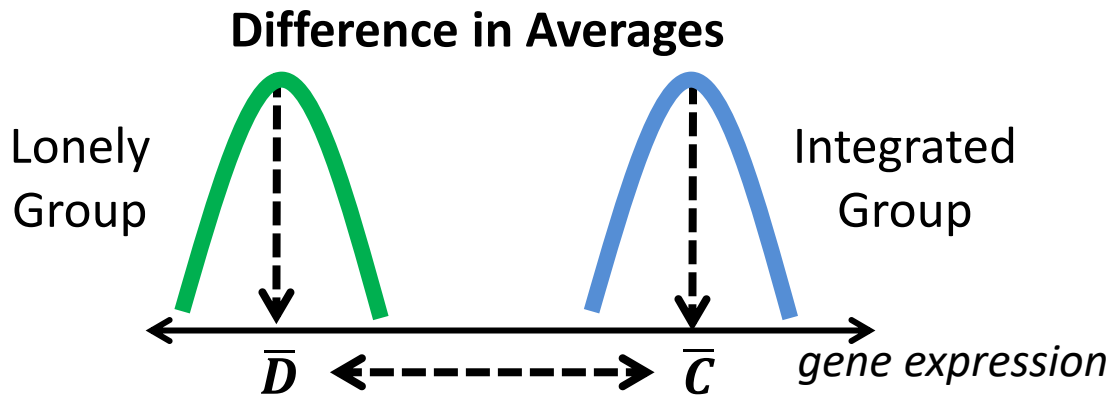
http://bioconductor.org



Standard approaches to analyzing large-scale gene expression data begin with identifying what's different…

Which Genes Are Different between Two Phenotypes?

Example: the primary goal of the study was to assess differential gene expression in leukocytes between lonely and integrated people.



What do we know about our patient data?
Where is this information stored?
How can we identify which columns correspond to which patients?

Social Genomics – Loneliness, Happiness and Science?!

An emerging area of social science deals with the intersection of happiness/loneliness and the impact on human health. More recently, this field has taken a quantitative molecular approach, giving rise to **“social genomics”**.

Loneliness Is Bad For You, And This Study May Help Explain Why

Feeling lonely may trigger changes in our cells that could make us more susceptible to illness.

11/28/2015 08:53 am ET

Jacqueline Howard
Senior Science Editor, The Huffington Post

EVGENIASH VIA GETTY IMAGES

Loneliness can affect the production of white blood cells in our bodies, study shows.

Forbes / Pharma & Healthcare

NOV 24, 2015 @ 08:00 AM 15,913 VIEWS

Loneliness Destroys Physical Health From The Inside Out

David DiSalvo
CONTRIBUTOR

I write about science, technology and the cultural ripples of both.

FULL BIO >

Opinions expressed by Forbes Contributors are their own.

Loneliness can increase the risk of pr
study supported by the National Inst
of the potential for loneliness to dam

What the research team found is tha
strongly linked to two critical physio
immune systems and increased cellu
loneliness affects the expression of g
transcriptional response to adversity

The longer someone experiences lon
genes related to white blood cells (ak
infections) and inflammation. CTRA
simultaneously increasing the geneti
at the cellular level rather than the s
happening within the body’s cells.

The combination of the two effects is
with a slow erosion of cellular health
problems, some of which worsen ove

The study also found that CTRA and
CTRA gene expression more than a y
more than a year later. In other word

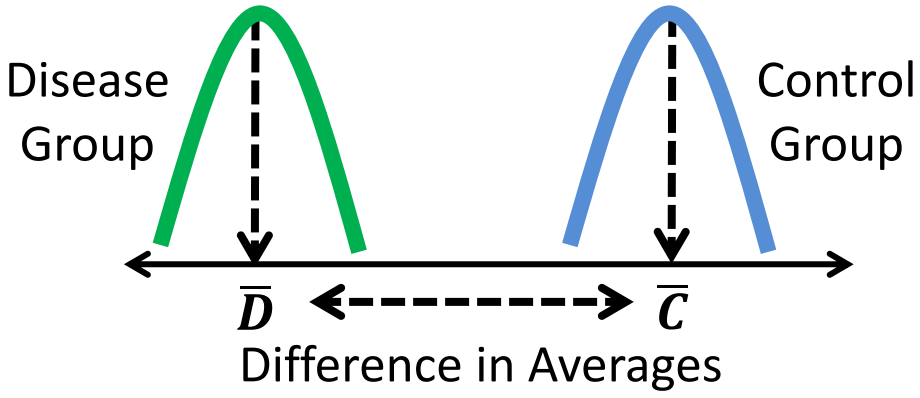
The Physical Effects Of Loneliness Include Weakened Immune Systems, Premature Death

AFP/Relaxnews

Posted: 11/24/2015 10:50 am EST | Updated: 11/24/2015 10:59 am EST

The T-test

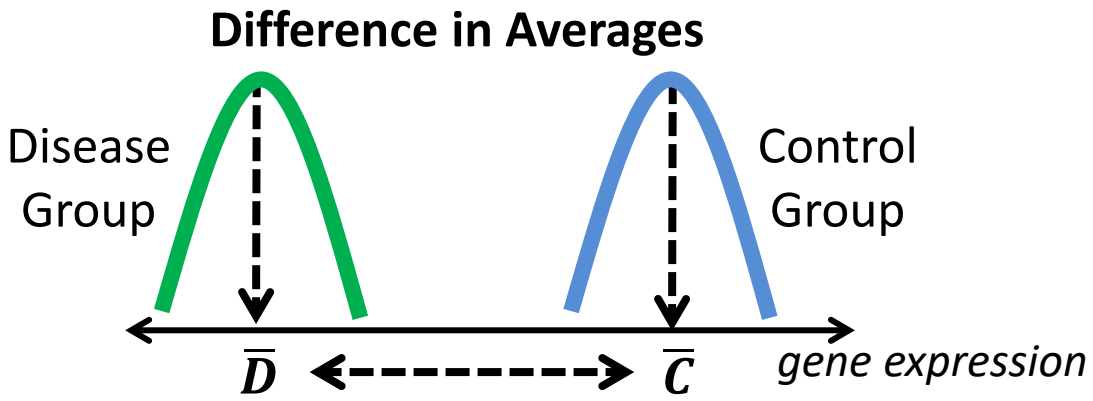
$$T_{(gene)} = \frac{\bar{D} - \bar{C}}{f(Var(D, C))}$$



Historical Note

- Gosset worked for the Guiness Brewery in Dublin, Ireland.
- He adopted the pseudonym of "Student" because his employer viewed the use of statistics as a trade secret.
- Gosset's job was to apply *biochemistry* + *statistics* to an industrial problem.

Assessing Differential Expression with a T-test



Some sample R code:

```
# for first gene
> t.test(edat[1,lonely.index=="HighLonely"],
      edat[1,lonely.index=="LowLonely"])
```

Assessing Differential Expression with a T-test

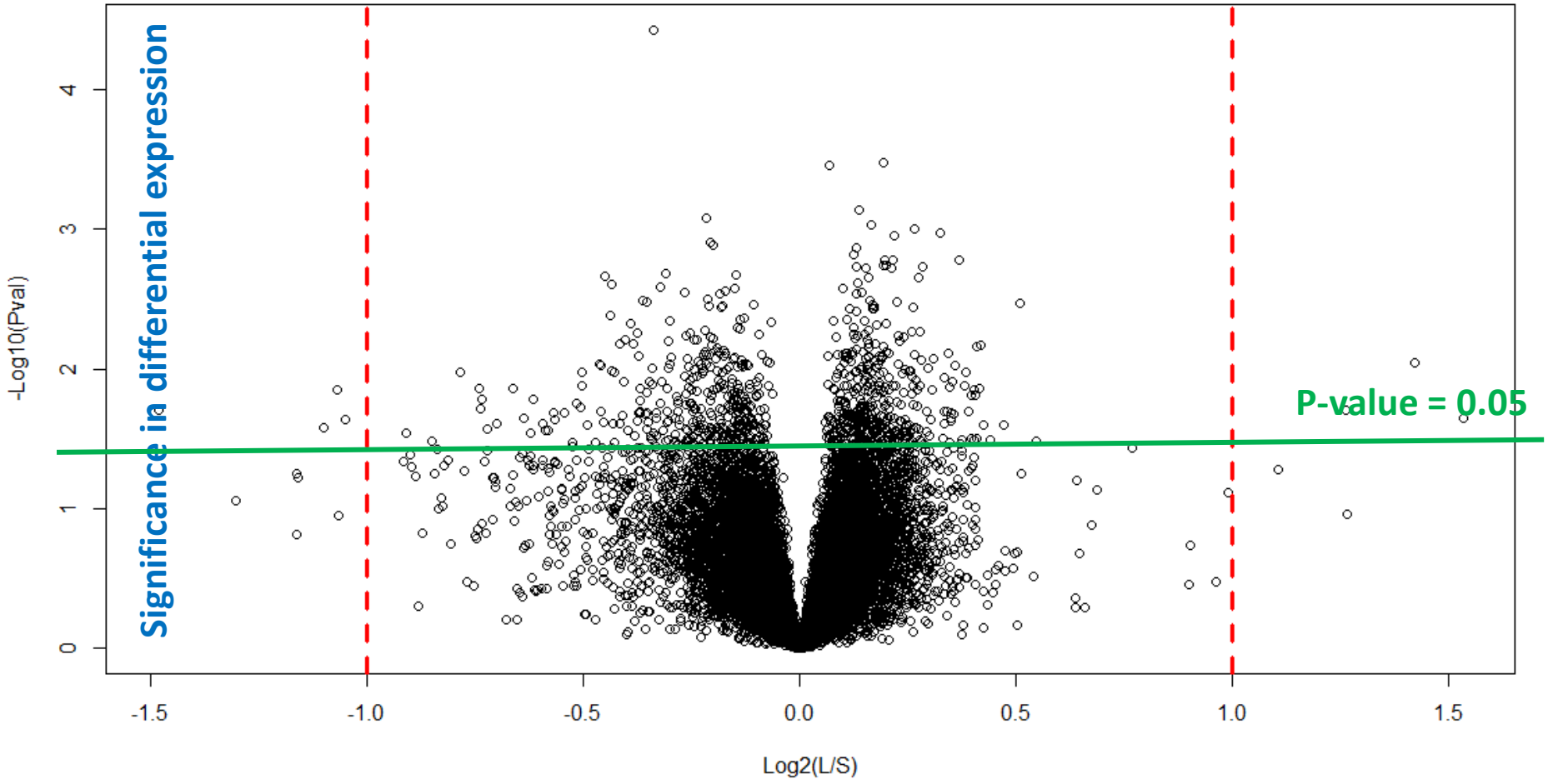
```
> runTP <- function(x,y){
                    res <- t.test(x[y=="HighLonely"],
                                  x[y=="LowLonely"])
                    p <- res$p.value
                    return(p)
                }
> tpvals <- apply(edat, 1, runTP, y=lonely.index)
> length(tpvals)
> sum(tpvals < 0.05)
```

How many genes are significant after multiple testing correction?

```
> tapvals <- p.adjust(tpvals, "BH")
> sum(tapvals < 0.05)

> summary(tapvals)
```

A volcano plot is a device that let's us assess the overall distribution of differential gene expression

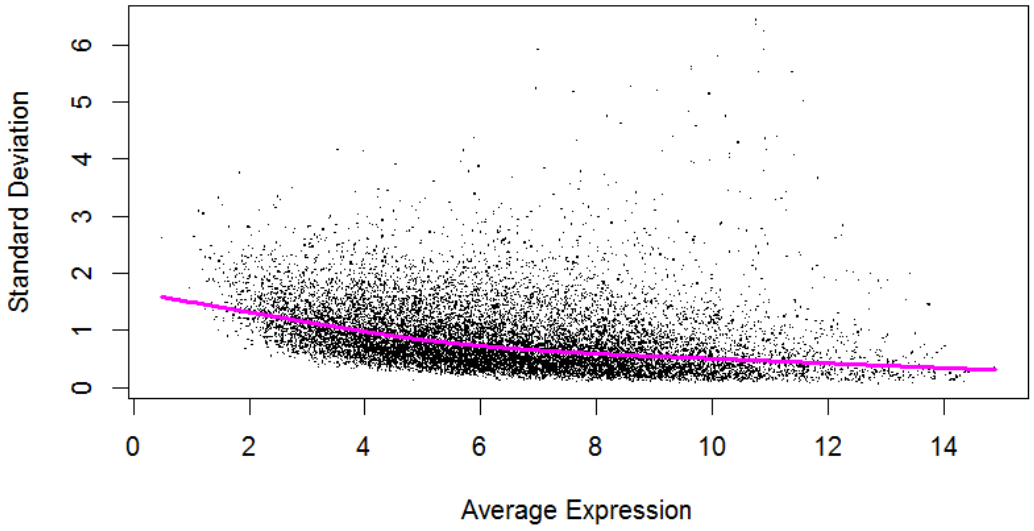


Log_2 Expression Fold Change shows the direction of gene expression in one condition relative to another

Testing for differential expression using limma*

- When dealing with –omic level platforms, we are working with high-dimensional data, and tiny quantities of biological material.
- Noisy data and false positives are therefore bound to occur.
- Limma uses an empirical Bayes method to estimate differential expression by minimizing the variance estimate.
- This results in a moderated T-statistic:

$$T_{(gene)} = \frac{\bar{D} - \bar{C}}{f(Var(D, C) + \alpha)}$$



*limma is a R/Bioconductor package that is used for microarray and RNA-seq data analysis.

Integrating gene expression with other types of –omics data

Integrating gene expression data to understand biology

Do we see similar gene expression patterns in the lonely cohorts profiled in PNAS (2011, 2015) and Genome Biology (2007).

Which pathways (pro-inflammatory?) have different expression in lonely versus non-lonely people?

1. Meta-Analysis
Building evidence for consistent trends across multiples lines of data sources and experiments.

2. Integrating with external sources of information.
Interpreting results using pathways, gene sets or other properties of interest from the literature.

3. Integrating different types of genome-wide data.
Modeling related high-throughput data sets to identify multi-level regulatory events.

Are genes with differential expression in the lonely versus non-lonely people associated with SNPs or CNVs?

Pathways and ontologies

Efforts have been made to systematically characterize our knowledge of biological pathways and processes into public databases.

KEGG: Kyoto Encyclopedia of Genes and Genomes

Initially set up to characterize metabolic pathways, but now represents all cellular pathways. Low coverage of the genome, but high quality gene sets. `In R: KEGGREST`

Reactome

Pathway information is manually curated and peer-reviewed, can be downloaded in different formats and cross referenced to other databases. `In R: reactome.db`

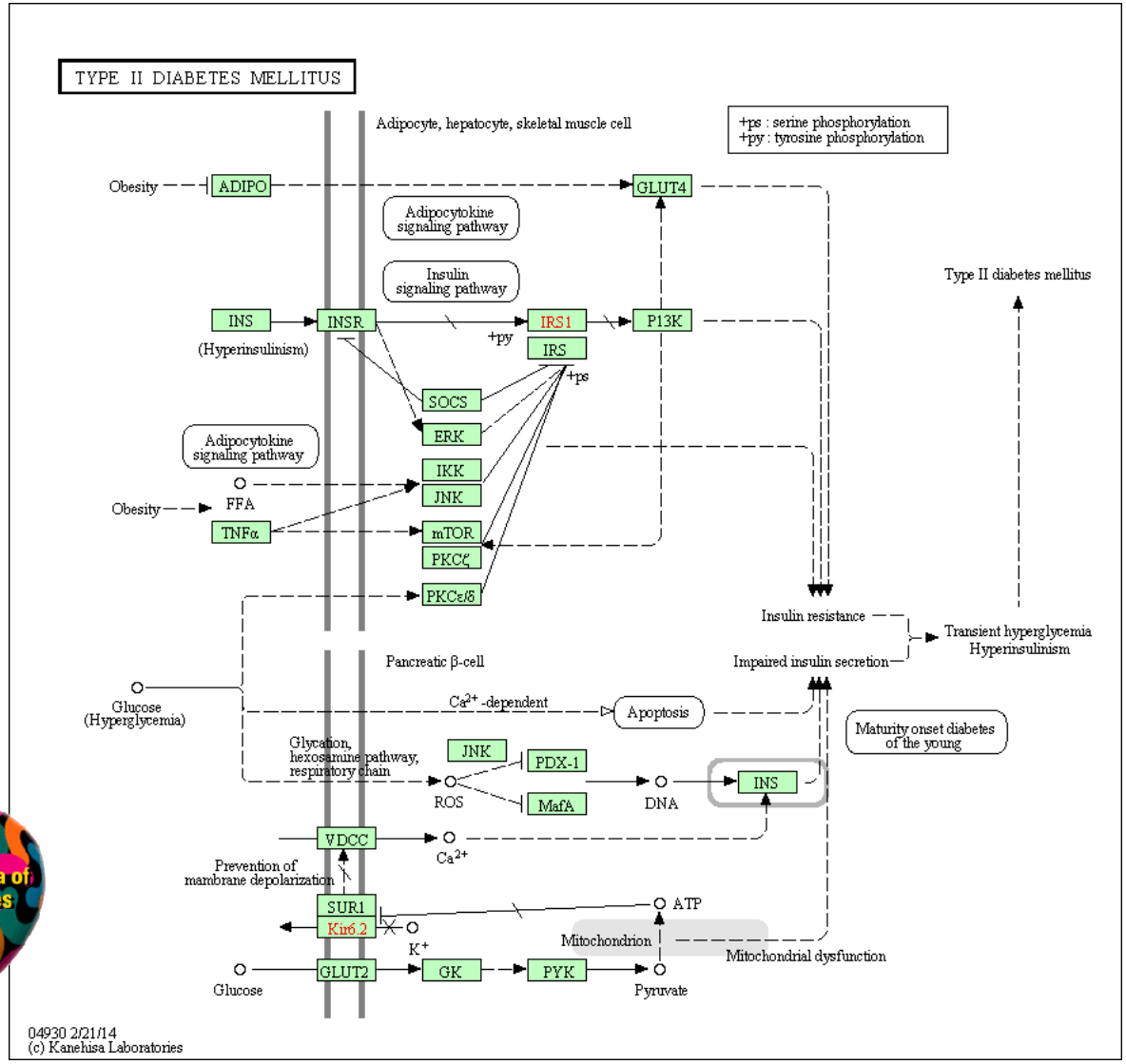
Gene Ontology

Hierarchical definitions by biological process (BP), molecular function (MF), cellular component (CC). Genes can be filtered on evidence codes representing the reliability of the assignment. `In R: GOstats`

MSigDB

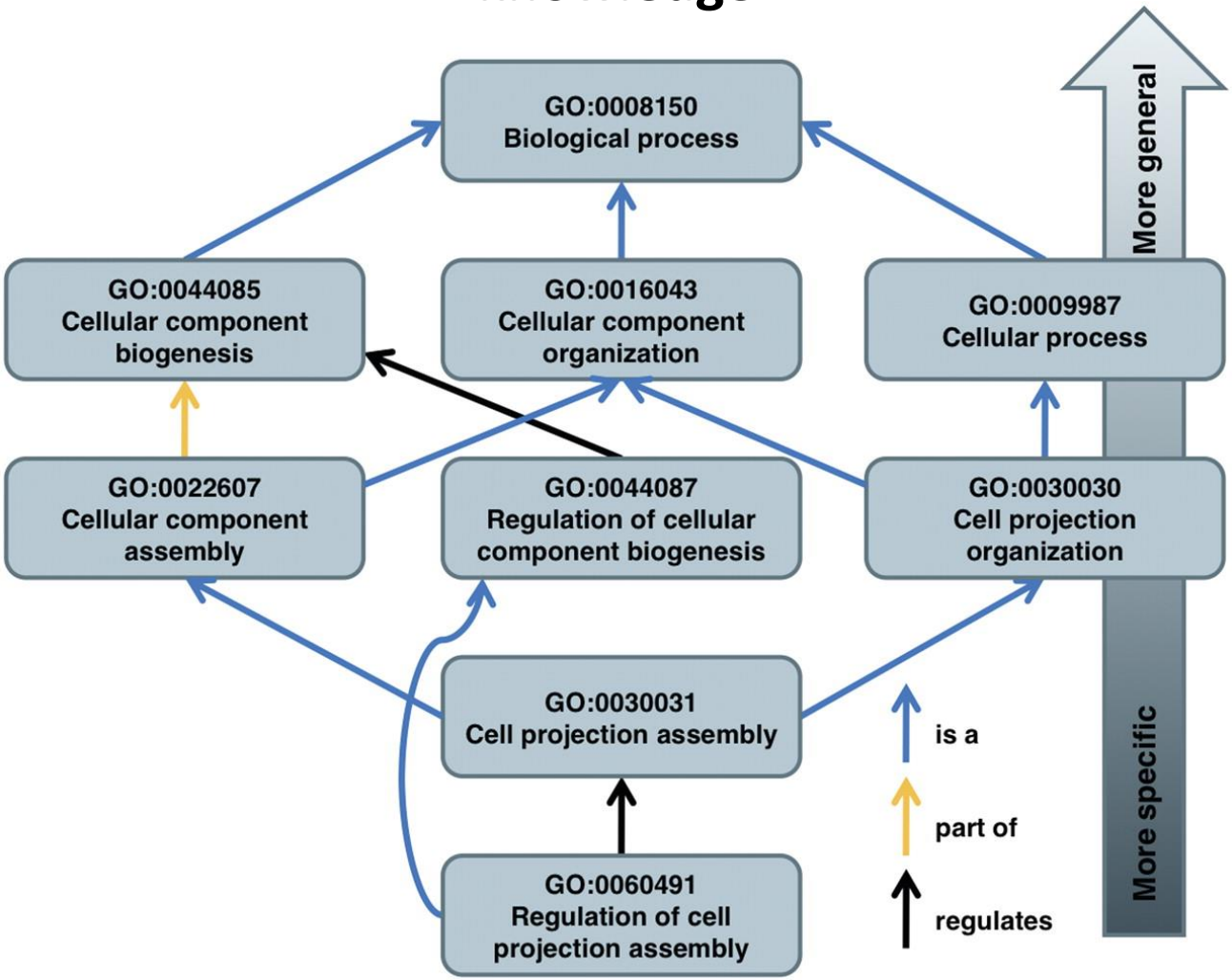
One of the most comprehensive sources of gene set information; there are 7 major groups, some of which overlap with the above resources.

KEGG Pathway: Type II Diabetes (human)



Kyoto Encyclopedia of Genes and Genomes

Gene Ontology: a computational representation of biological knowledge

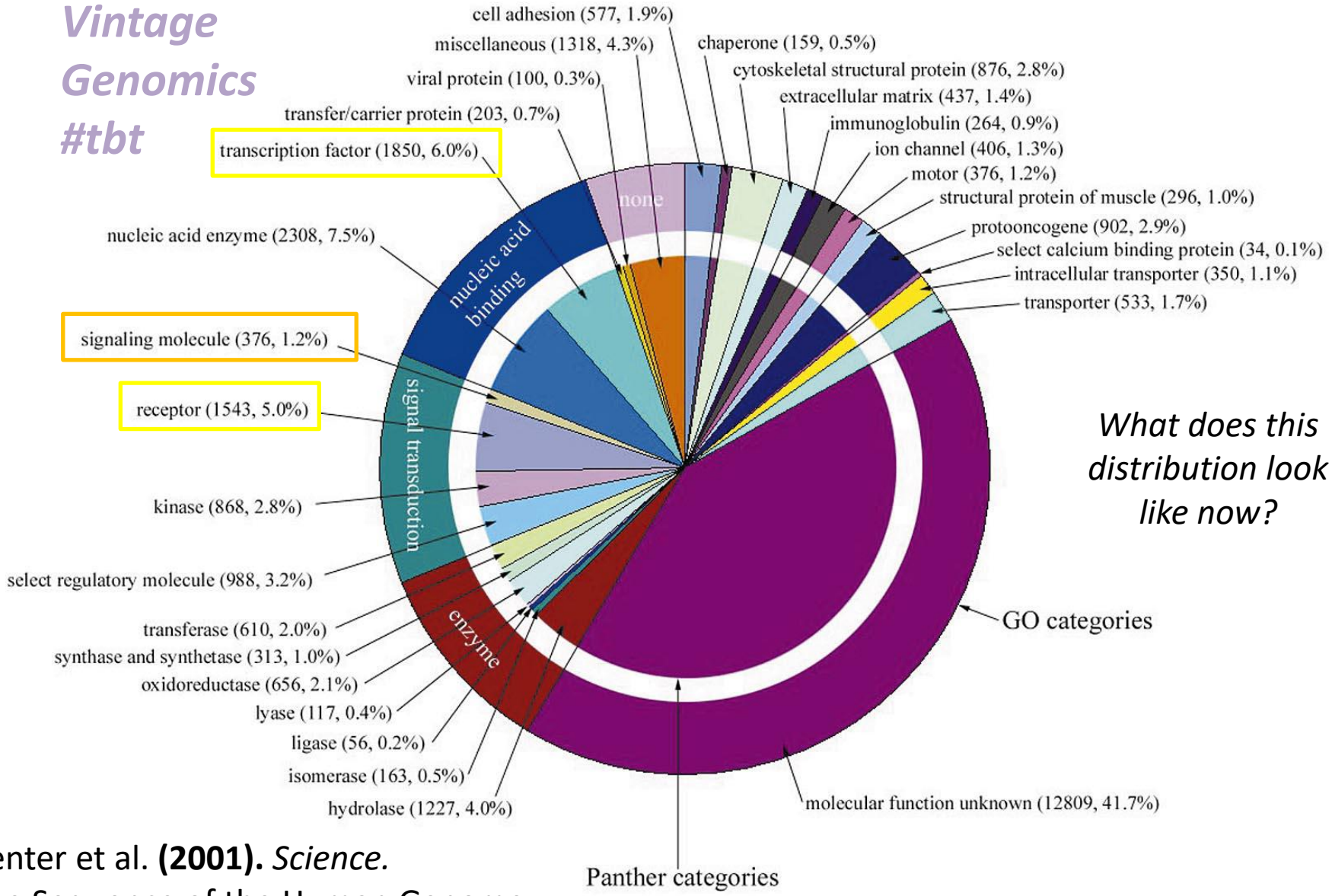


Louis du Plessis et al. Brief Bioinform 2011;12:723-735

Briefings in Bioinformatics

Distribution of Human Genes in GO:MF (20 years ago!)

Vintage Genomics #tbt



What does this distribution look like now?

Venter et al. **(2001).** *Science.*
The Sequence of the Human Genome.

Where does the information come from?

GO evidence codes and their abbreviations.

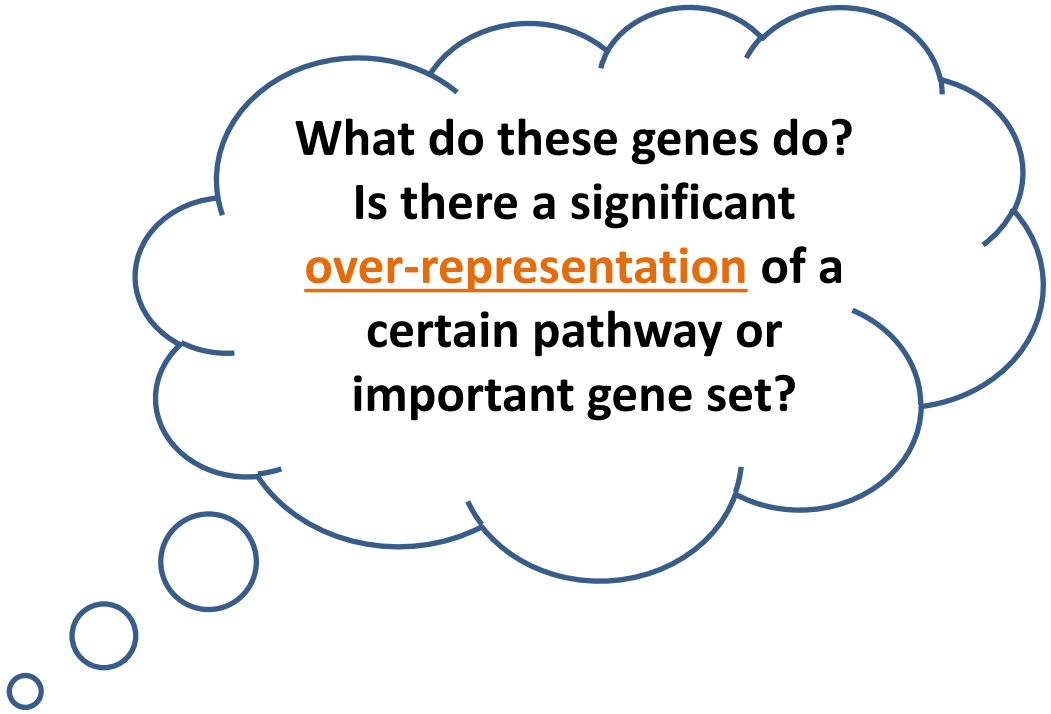
Experimental Evidence Codes		Computational Analysis Evidence Codes	
EXP	Inferred from Experiment	ISS	Inferred from Sequence or Structural Similarity
IDA	Inferred from Direct Assay	ISO	Inferred from Sequence Orthology
IPI	Inferred from Physical Interaction	ISA	Inferred from Sequence Alignment
IMP	Inferred from Mutant Phenotype	ISM	Inferred from Sequence Model
IGI	Inferred from Genetic Interaction	IGC	Inferred from Genomic Context
IEP	Inferred from Expression Pattern	RCA	Inferred from Reviewed Computational Analysis
Author Statement Evidence Codes		**Curator Statement Evidence Codes**	
TAS	Traceable Author Statement	IC	Inferred by Curator
NAS	Non-traceable Author Statement	ND	No biological Data available
Automatically-assigned Evidence Codes		**Obsolete Evidence Codes**	
IEA	Inferred from Electronic Annotation	NR	Not Recorded

du Plessis et al. *(2011).* **Briefings in Bioinformatics.** 12:723-735

Briefings in Bioinformatics

Integrating gene lists of interest with pathway information provides biological/mechanistic context

ZFPM1	NLGN2
EXOC6	GLIPR2
COX4I1	PLXDC2
ECH1	MGP
ZMAT3	BTF3L4
ECM2	**OR11L1**
PORCN	EGFLAM
IL13RA1	NELFB
RPPH1	NR2F2
SCRN1	TMSB15B
TRAK1	SNAPC4
HBEGF	DKK3
WDR12	STX2
RFX1	**HSPA1A**



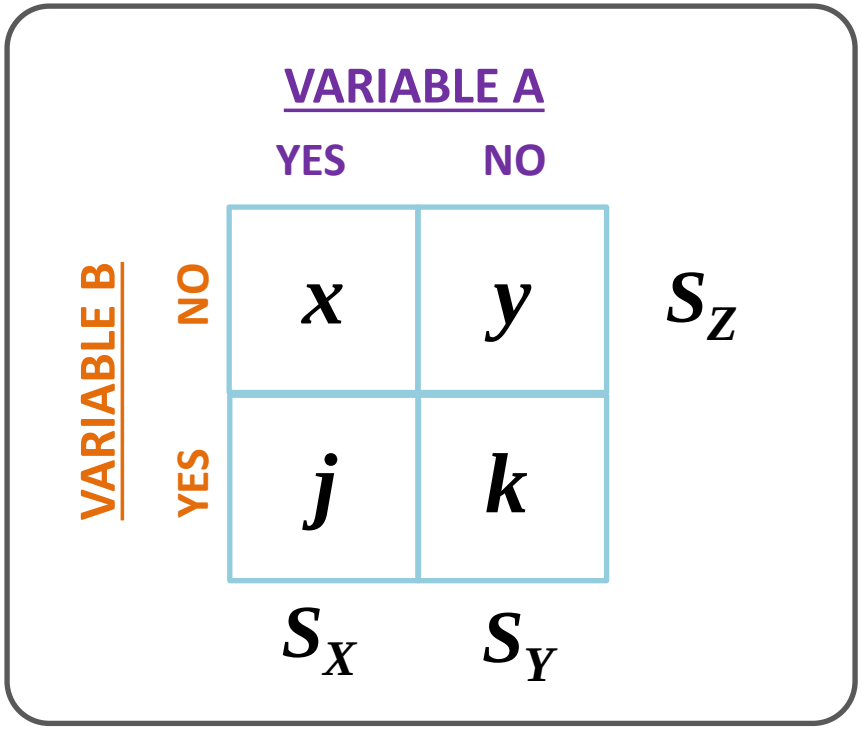
Over-representation analysis: Fisher's Exact Test

Tests the association between two variables using a **Hypergeometric distribution**.

Fisher's Exact Test tests the enrichment of seeing an overlap between two variables.

It can also be used to test the goodness of fit exactly.

Used for small numbers, but actually works for any size.



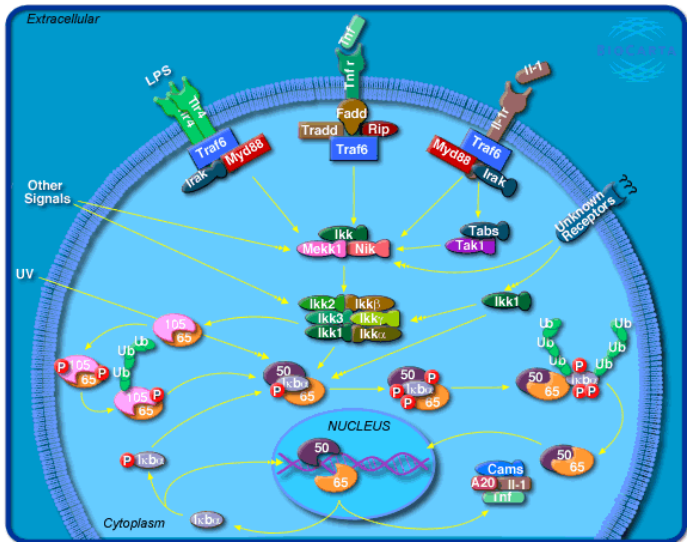
$$P(X = x) = \frac{\binom{S_X}{x}\binom{S_Y}{y}}{\binom{S_X + S_Y}{S_Z}} = \frac{\binom{S_X}{x}\binom{S_Y}{y}}{\binom{S_X + S_Y}{x + y}}$$

for $\max(0, S_Z - y) \leq x \leq \min(S_Z, S_X)$

Testing for enrichment of a single pathway in a given gene list

Consider a list of genes (e.g. loneliness study/cancer biomarker/your dream experiment). The goal is to examine whether this list is enriched for genes in the NFκB pathway.

NFκB Pathway (BioCarta)



	Interesting Gene List	Not in Interesting Gene List	
Genes in NFκB Pathway	70	y	150
Genes Not in NFκB Pathway	j	k	
	100	120	

2 x 2 contingency table

Hypergeometric random variable

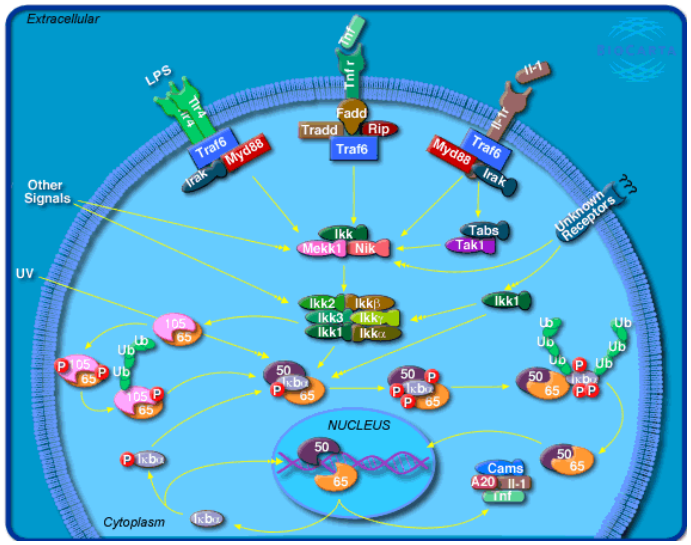
$$P(X = 70) = \frac{\binom{100}{70}\binom{120}{80}}{\binom{220}{150}} =$$

probability of seeing 70 genes that belong to the NFκB pathway **AND** in biomarker gene list.

$$\text{P-value} = \quad P(X \geq 70) = \sum_{i=70}^{100} \frac{\binom{100}{i}\binom{120}{150-i}}{\binom{220}{150}}$$

Applying Fisher's Exact Test in R

NFκB Pathway (BioCarta)

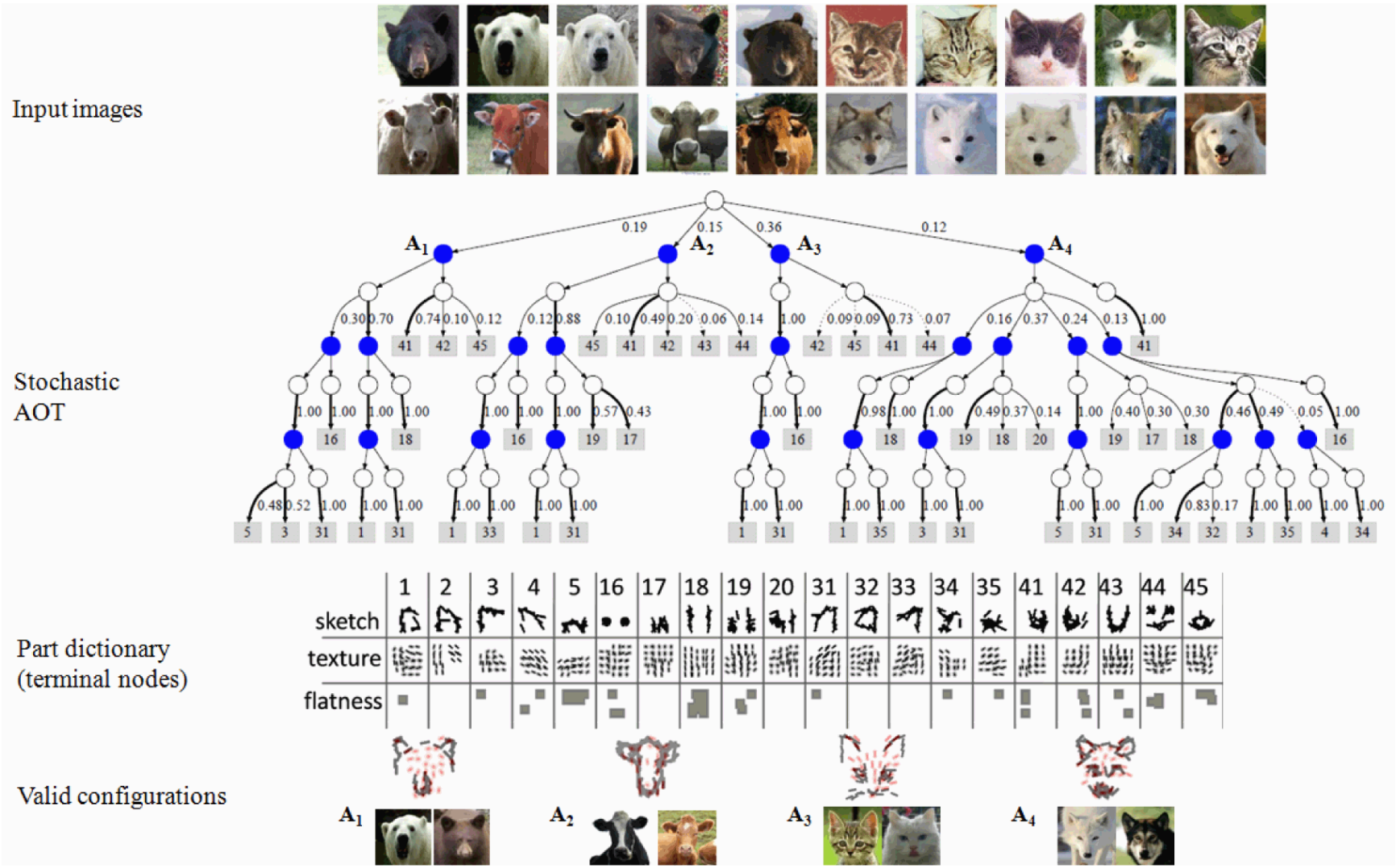


	Interesting Gene List	Not in Interesting Gene List	
Genes in Pathway Z	70	y	150
Genes Not in Pathway Z	j	k	
	100	120	

```
> tab <- cbind(c(70, 100-70), c(80,40))
> fish.res <- fisher.test(tab, alt="great")
> fish.res$p.value                    # P-value
```

Identifying groups of genes based on clusters of expression profiles

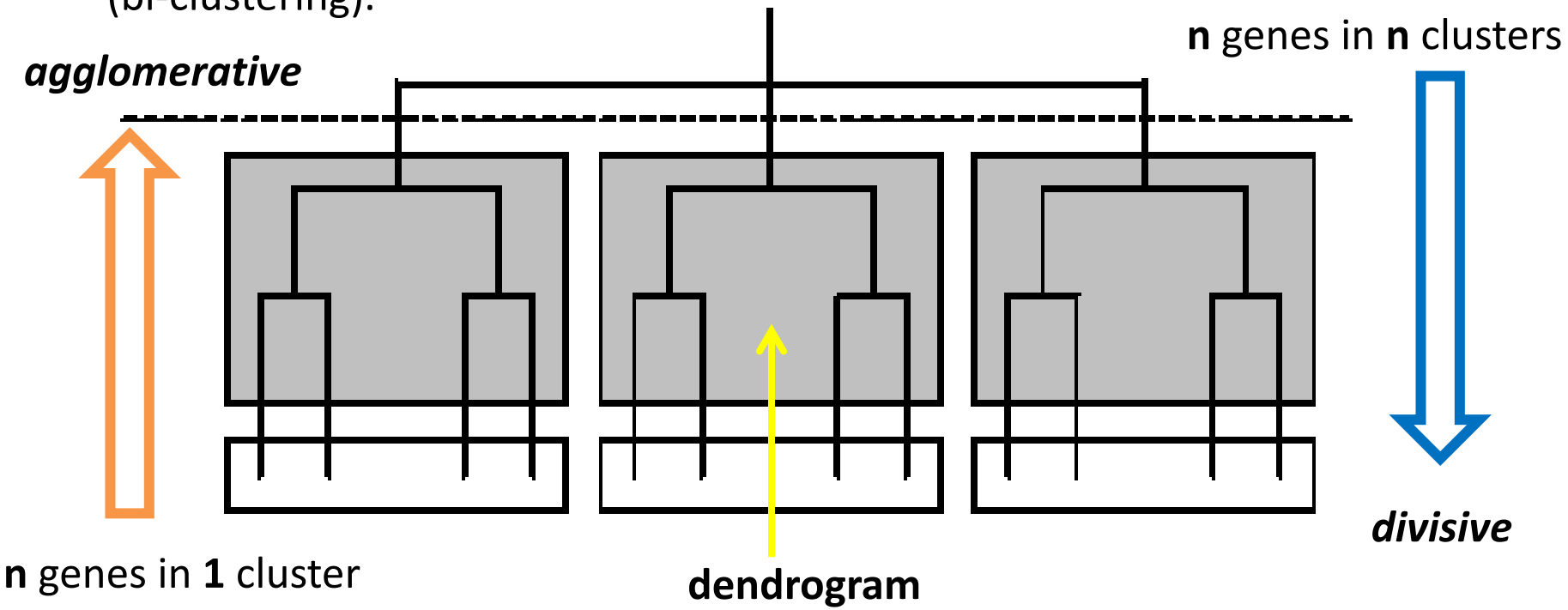
Unsupervised learning is the task of identifying patterns in the presence of many data variables where the number of patterns is also not known.



Si & Zhu. (2013). *IEEE Transactions on Pattern Analysis & Machine Intelligence*

Hierarchical clustering

- Constructs a hierarchy of clusters.
- Nodes in the dendrogram can be either genes, or samples or both (bi-clustering).



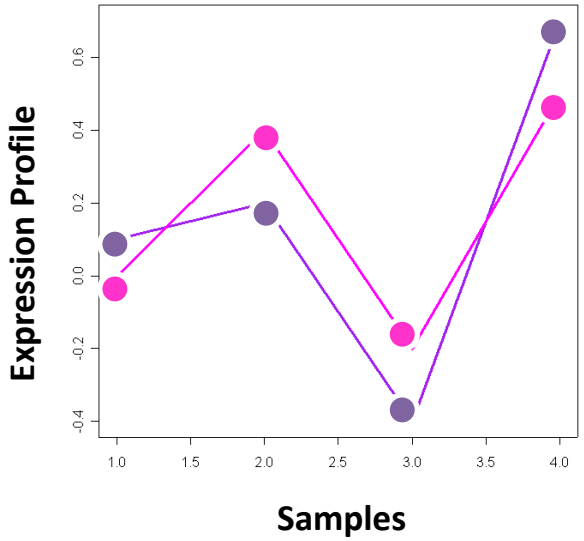
- We join nodes based on the notion of maximum **'similarity'**.
- Equivalently, we break nodes based on minimal similarity

Source: J-Express Manual

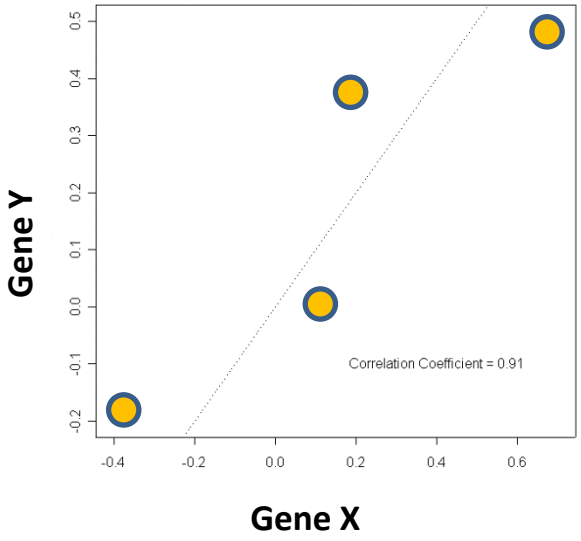
Measures of similarity – what counts as the same versus different?

Consider expression profiles of Gene X and Gene Y: how do we score their similarity?

Euclidean Distance



Correlation Coefficient

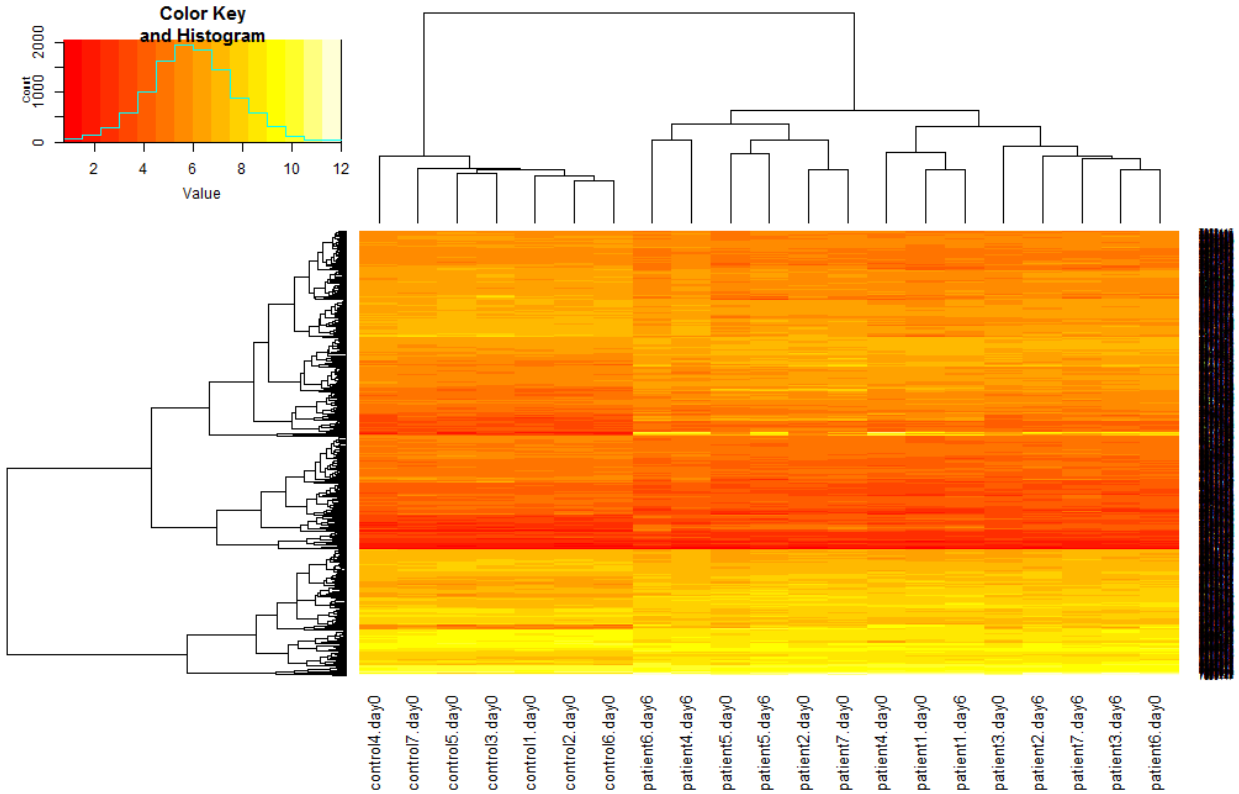


$$d^2(p,q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_i - q_i)^2 + \ldots + (p_n - q_n)^2.$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Heatmaps!

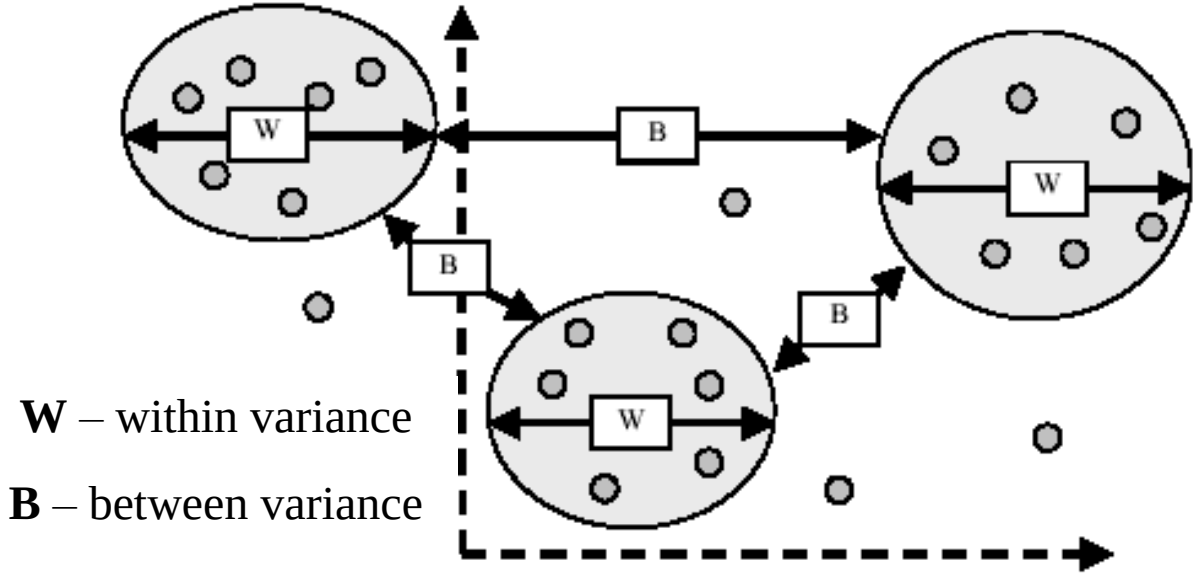
```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("gplots")
> library(gplots)
> heatmap.2(edat.sig, trace="none", margins=c(8,8))
```



More information via this [helpful tutorial](#).

Partitioning methods: k-means clustering

- Identifying the distinct set of expression profiles represented in the data set.
- Grouping genes based on their similarity to cluster profile.



Reference: J-Express manual

General Framework of a K-means Algorithm

Step 0: Start with a defined number of clusters.
Step 1: Initialize clusters; usually based on agglomerative hierarchical clusters.

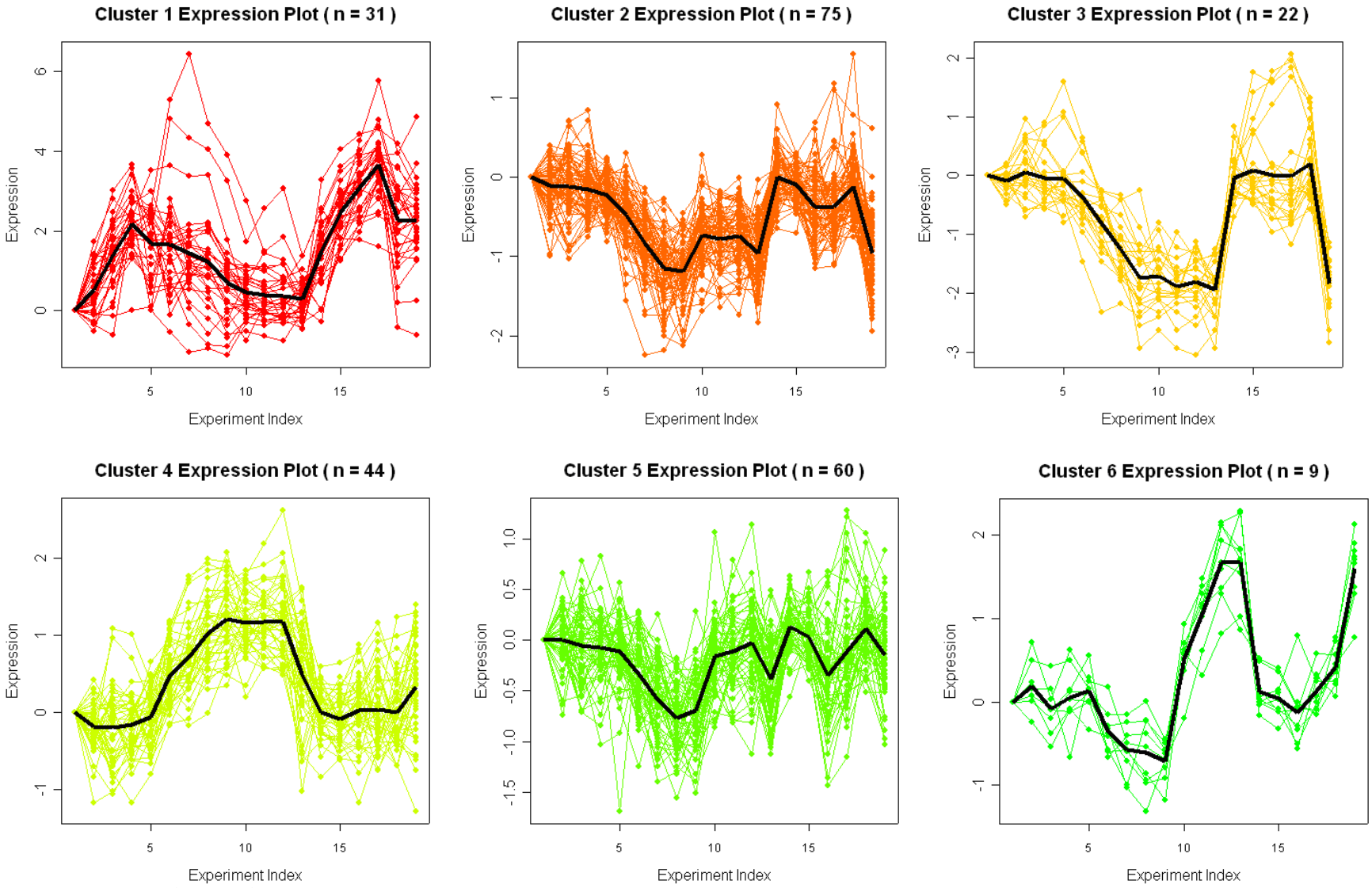
Means = K-means.

Medians = K-medoids, PAM

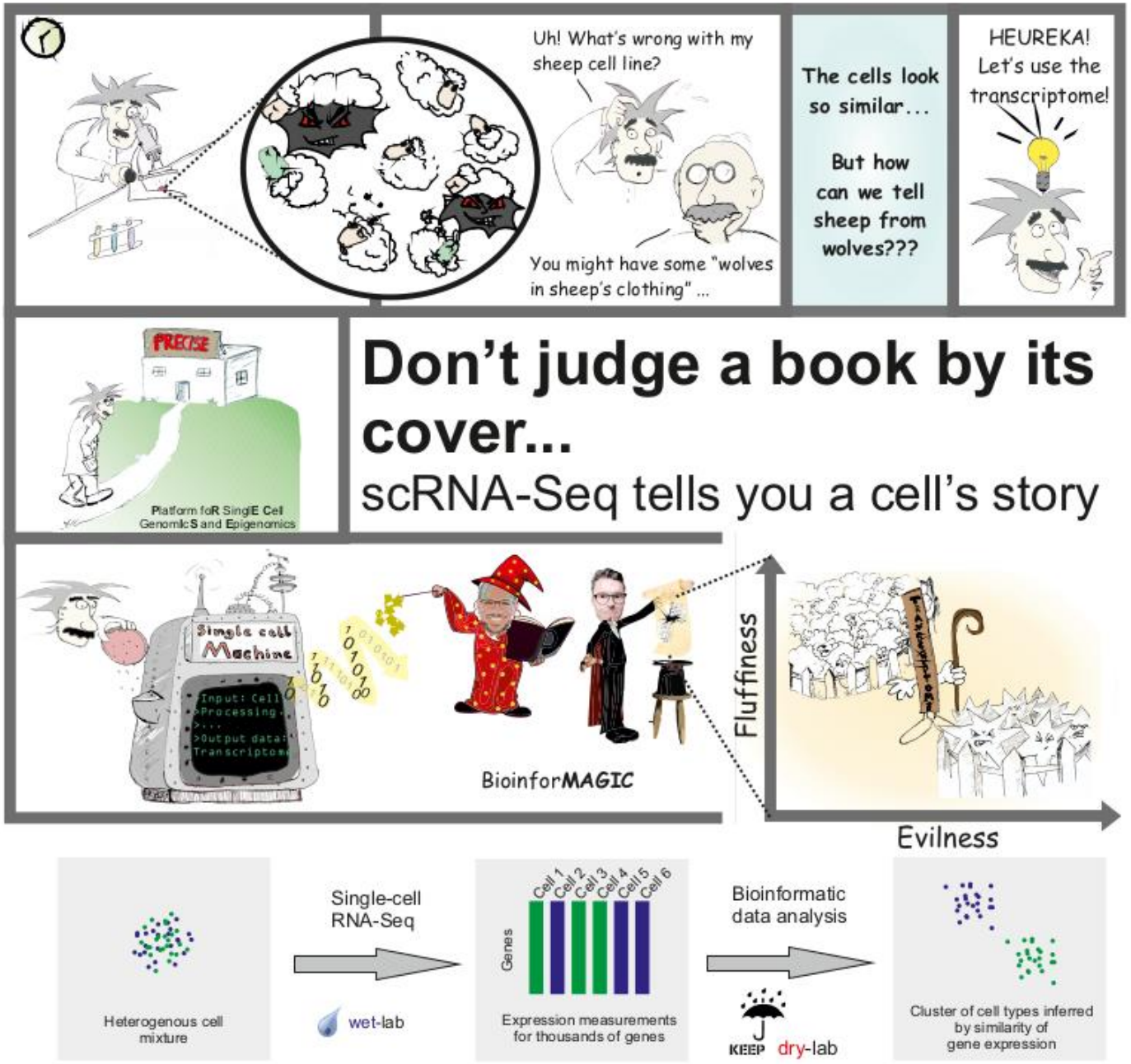
Step 2: Random sort of list, assign each gene to a cluster based on distance metric.
Step 3: Assess convergence criteria. If convergence achieved, stop. Otherwise repeat.

Mapping Genes to their Roles in the Cell Cycle

241 *Saccharomyces Cerevisiae* genes from a time course experiment into 6 clusters.



Spellman et al. **(1998).** *Mol Bio Cell.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.



http://www.immunosensation-blog.de/dont-judge-book-cover-scrna-seq-tells-cells-story/

Biology occurs on many different scales

Breast Cancer: a disease-related example

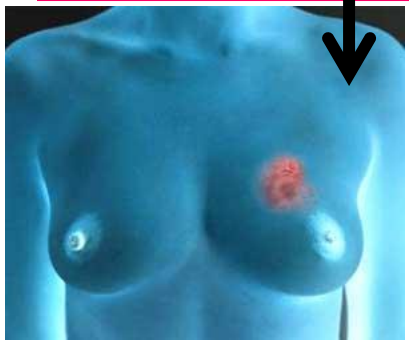
Multi-cellular organization

Behavior/physiology of cells:
Staging – TNM classification; 0 to 4

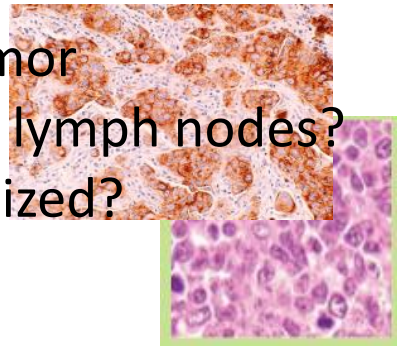
Single cell features:
Histopathology & Grade

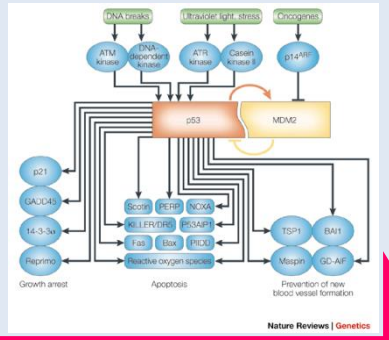
Behavior/aberration of gene profiles:
Oncotype DX, MammaPrint.

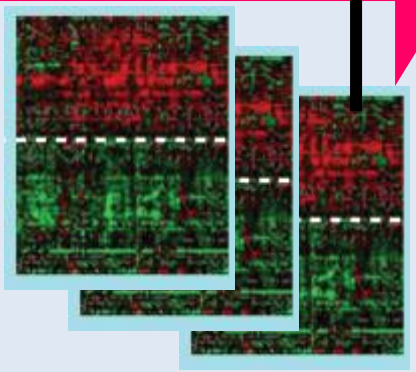
Tissues | **Cells** | **Sub-cellular molecules (genes, proteins)**



T = size of tumor
N = spread to lymph nodes?
M = metastasized?

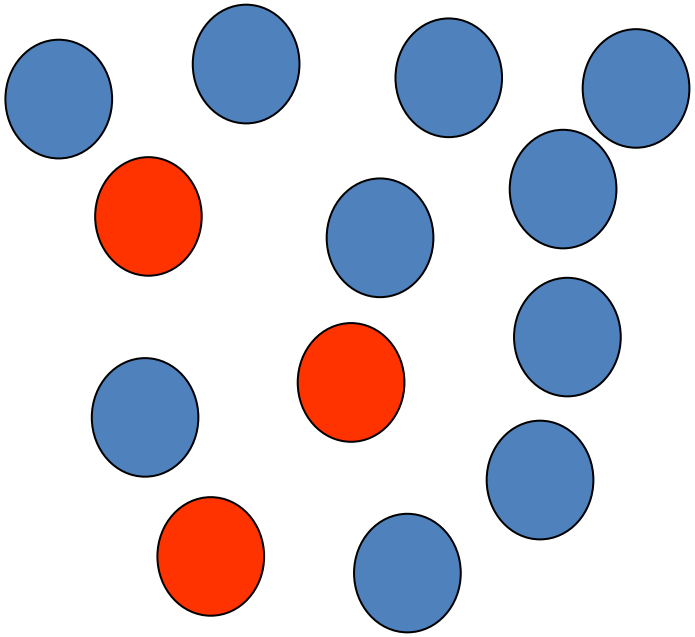




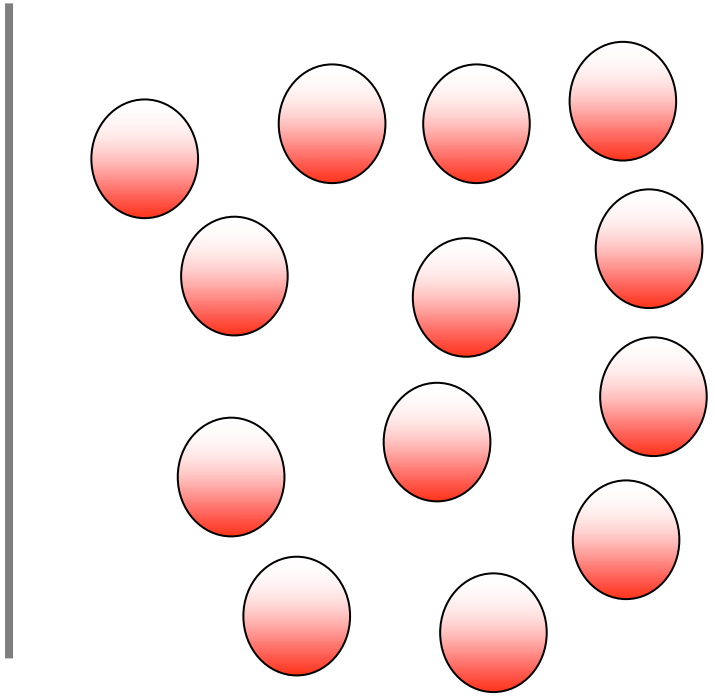


Genome-wide Microarrays & Sequencing

Cell populations are inherently heterogeneous



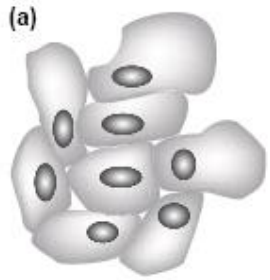
Ensemble methods survey the "average" transcriptome: microarrays, qPCR, RNA-seq



Single cell sequencing is changing the way in we think about transcriptional regulation.

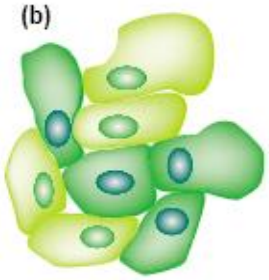
Conceptualizing gene expression in single cells

1980s:
Before single cell assays were invented:



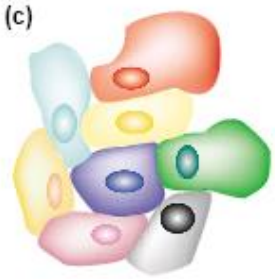
Cells were thought to be identical.

In situ hybridization in 1989 gave snapshots of individual nuclei.



Genes are either "on" or "off".

Single-cell gene expression profiling in 2001.

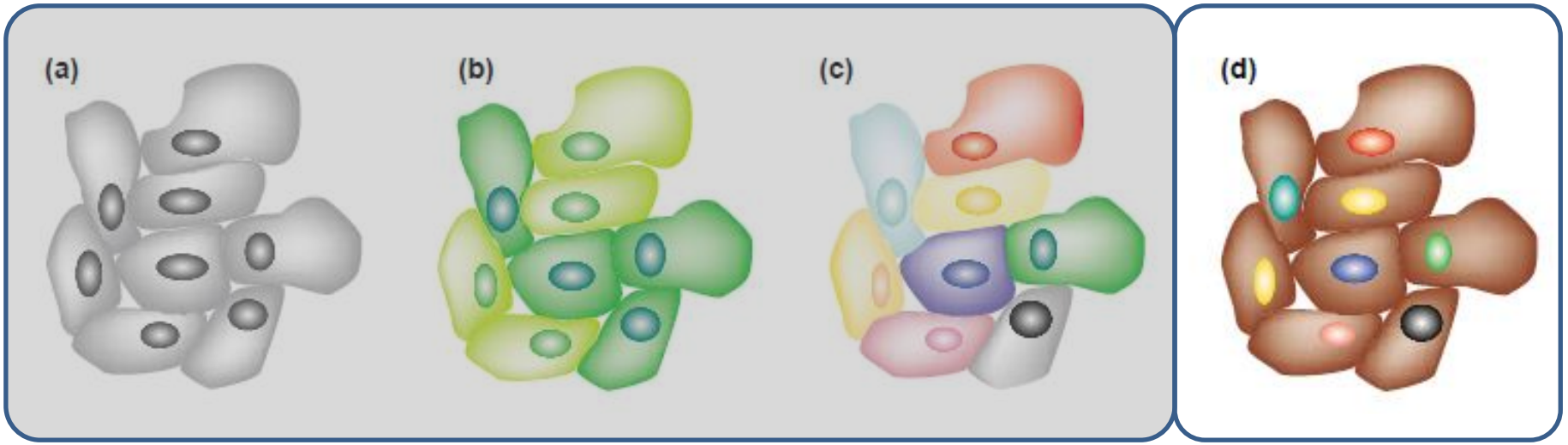


Cells express genes heterogeneously around a distribution of levels.

Levsky and Singer. (2003). *Gene expression and the myth of the average cell.* Trends in Cell Biology.

Understanding the functional effects of variability is the next frontier

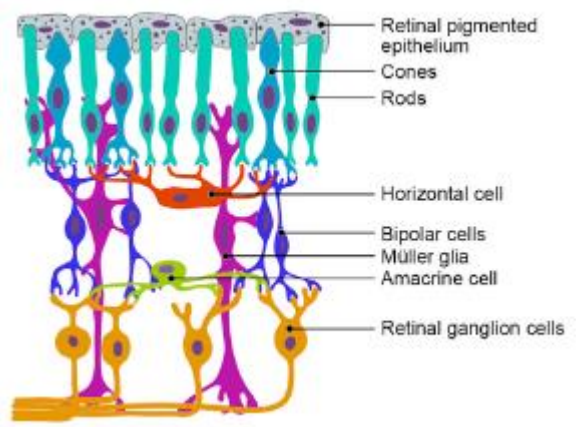
How are cells able to tolerate gene expression variability and maintain similar physiological function?

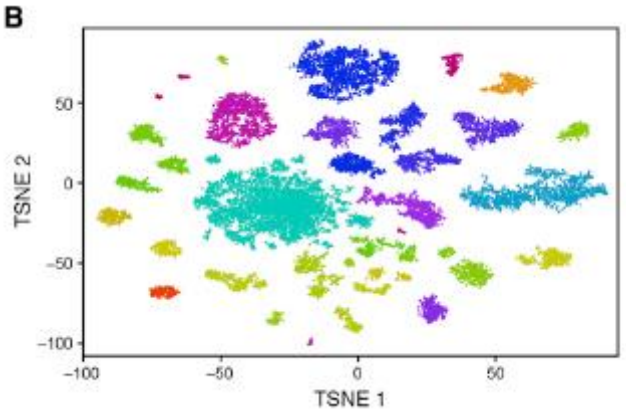


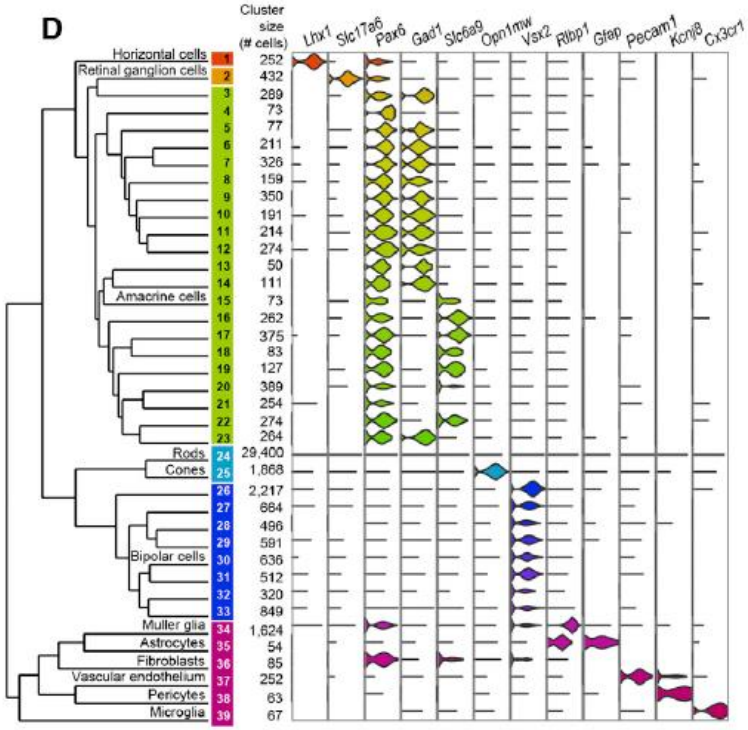
Are fluctuations dampened out at the protein level, over time, via different network configurations?

Global patterns of transcriptional regulation of cell type diversity are within reach!

Analysis of 44,80 cells via Drop-seq identified 39 cell populations in the mouse retina.

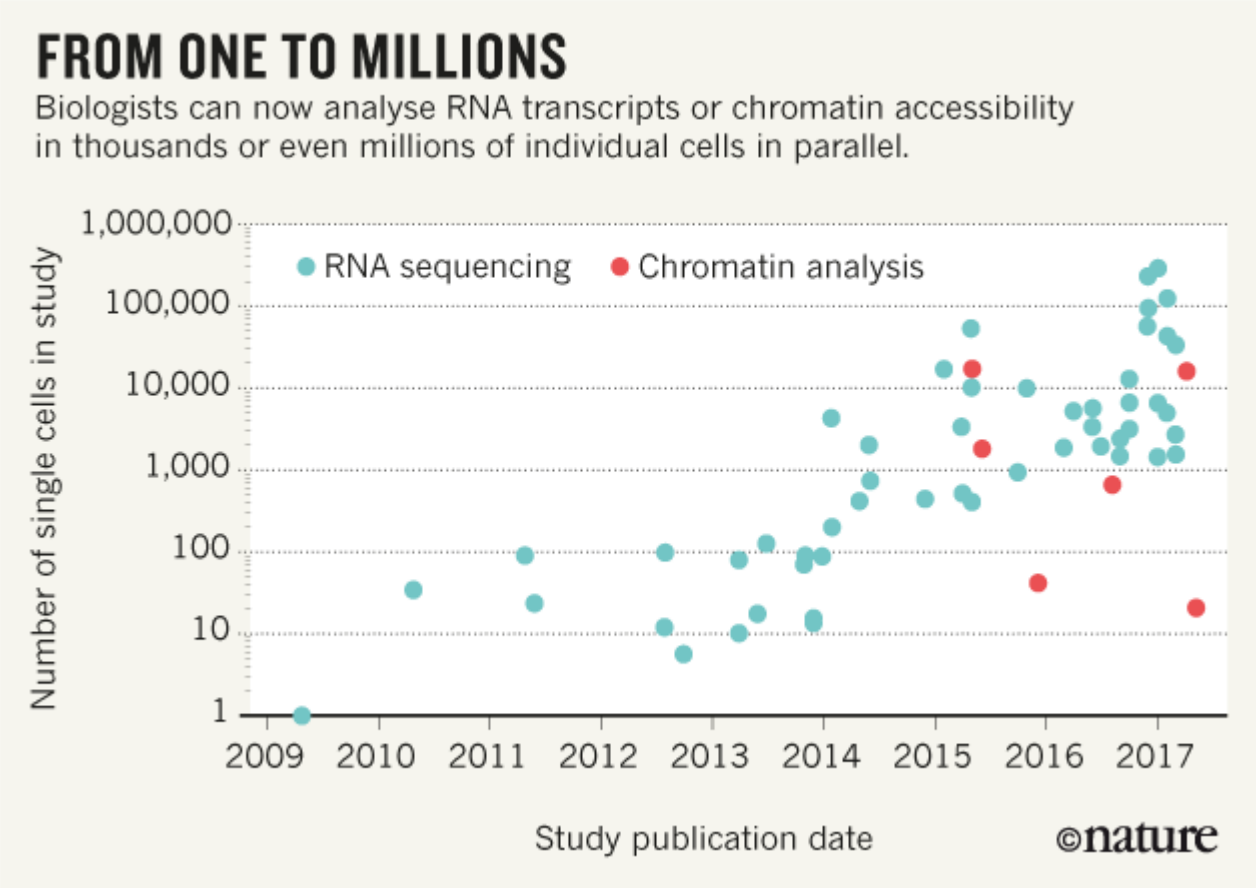






Macosko et al. **(2015).** *Cell.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.

Single cell RNA-seq throughput has exploded!



The Human Cell Atlas: from vision to reality : Nature News & Comment. 26 Oct 2017. Vol 550.

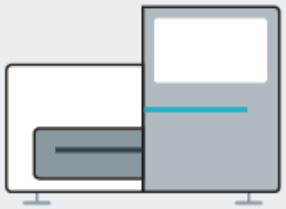
TO BUILD AN ATLAS

Scientists wishing to put together a 3D map of the thousands of cell types and subtypes in the human body will face challenges at every step.

TISSUE

CELLULAR DISSOCIATION

Sophisticated devices will be required to isolate different kinds of human cells from a range of tissues and prepare them for study in a way that does not stress them or change their nature.



SEQUENCING

DATA ANALYSIS

Sequencing must account for variability in the amount and quality of RNA or other molecules in different cell types, and yet computational approaches need to be standardized to ensure compatibility.

CELLULAR MAPPING

Multidimensional maps based on sequencing data will reveal the relative types, subtypes and abundances of cells in tissues, but in many cases these must be mapped back to where they reside in the body, using different spatial methods.

The **Human Cell Atlas** is currently the latest big data international consortium for RNA-sequencing.

https://www.humancellatlas.org/

The goal is to create a reference map for all human cells in the body – at single cell resolution.

This creates new challenges in technology, data analysis, and storage.

This is a great example of advances of next-generation sequencing are giving us new ways to do (exquisite) cell biology!

http://www.nature.com/news/how-to-build-a-human-cell-atlas-1.22239

Lecture Summary

- RNA-sequencing and microarrays are generally used for high-throughput gene expression data, with the former looking to eclipse the latter.
- Pre-processing of RNA-seq data requires alignment of reads, transcript identification and quantification.
- **Different statistical approaches can be used to identify changes and patterns in expression data.**
- **Bioinformatic tools based ontologies and pathways can be used to identify biological themes in the data.**
- **There is no MAGIC in bioinformatics methods!** ***(Only straightforward math and programming code).***

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Australian Institute for Bioengineering and Nanotechnology

For any questions, clarification or inspiring ideas please get in touch via email!

Research internships, Hons & Masters projects, PhD applications available with my lab.

Jess Mar, PhD

Australian Institute for Bioengineering & Nanotechnology Level 4 West

jmar@uq.edu.au

https://aibn.uq.edu.au/mar