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)

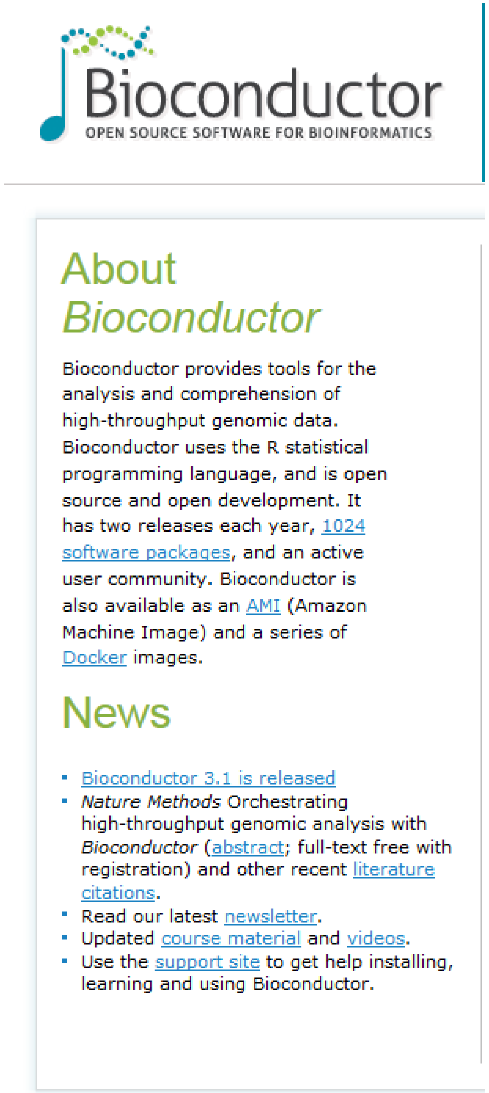
- What are the most common platforms for collecting high-throughput gene expression data?
- What are the key steps in analyzing RNA-sequencing and microarray data?
- **How can we learn about biology through analyzing gene expression data?***

***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



About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1024 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

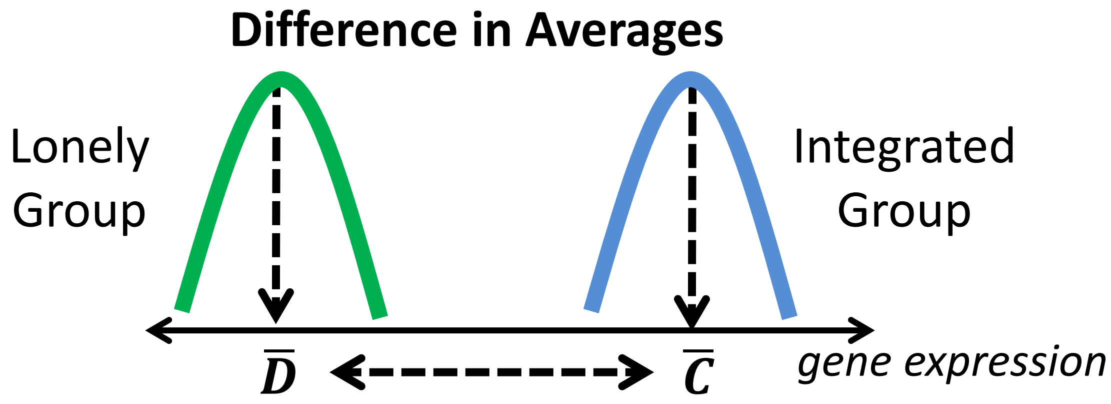
News

- Bioconductor 3.1 is released
- *Nature Methods* Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract; full-text free with registration) and other recent literature citations.
- Read our latest newsletter.
- Updated course material and videos.
- Use the support site to get help installing, learning and using Bioconductor.

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

Which Genes Are Different between Two Phenotypes?

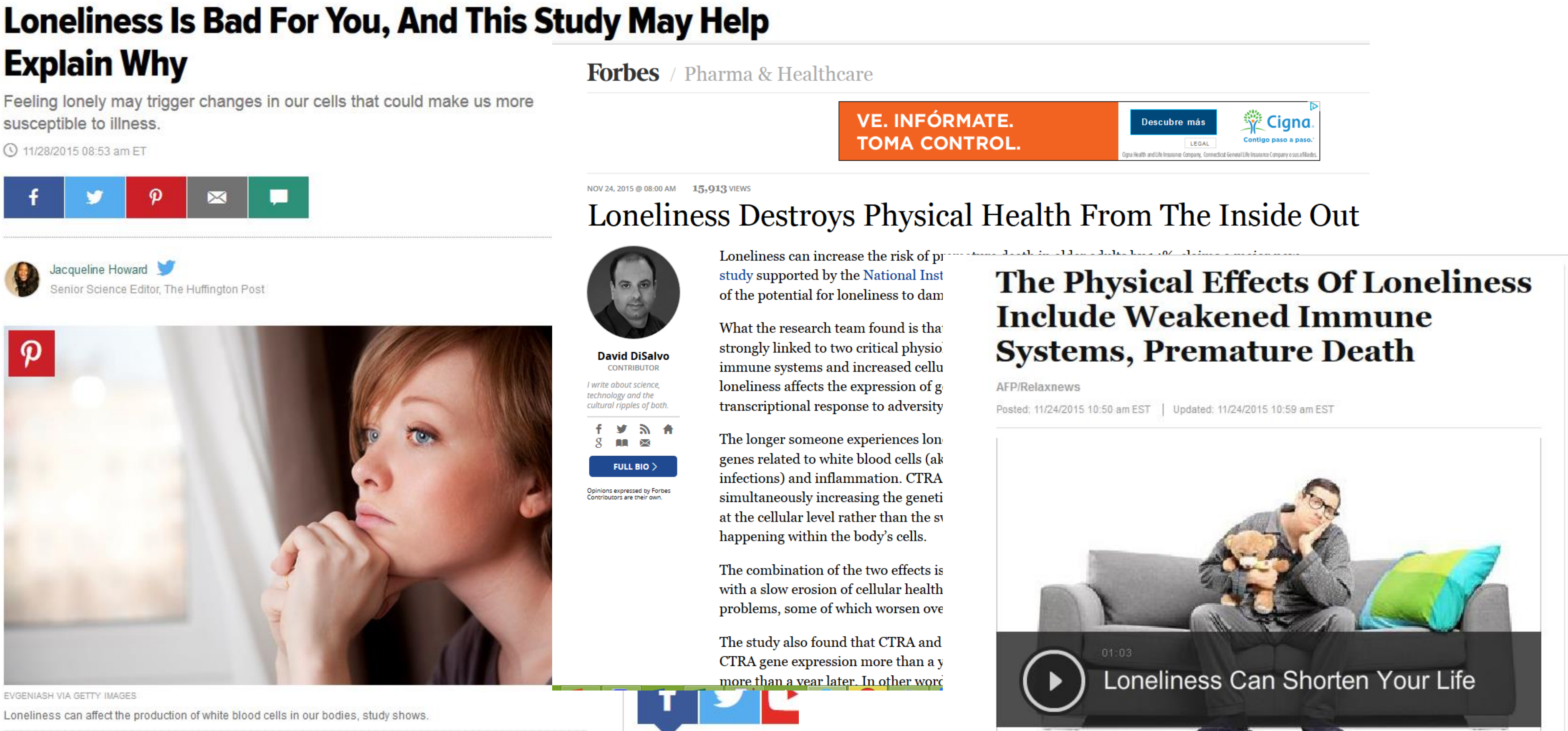
Example: the primary goal of the study was to assess differential gene expression in leukocytes between lonely and integrated people.



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 sy
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

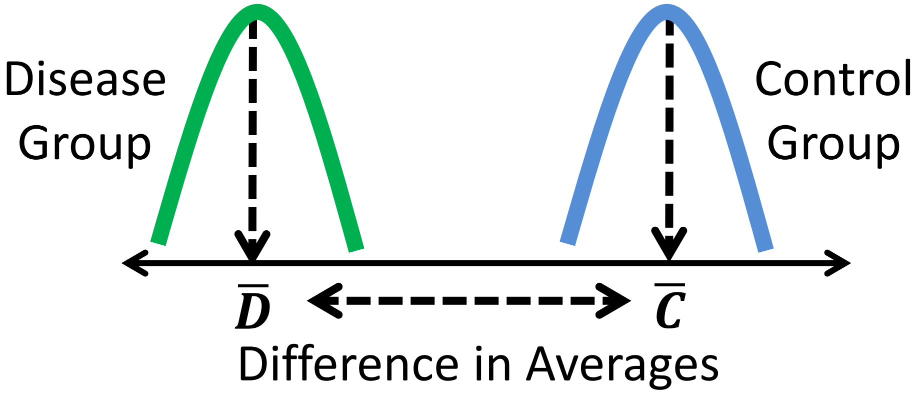
The Physical Effects Of Loneliness Include Weakened Immune Systems, Premature Death

AFP/Relaxnews

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

The T-test

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

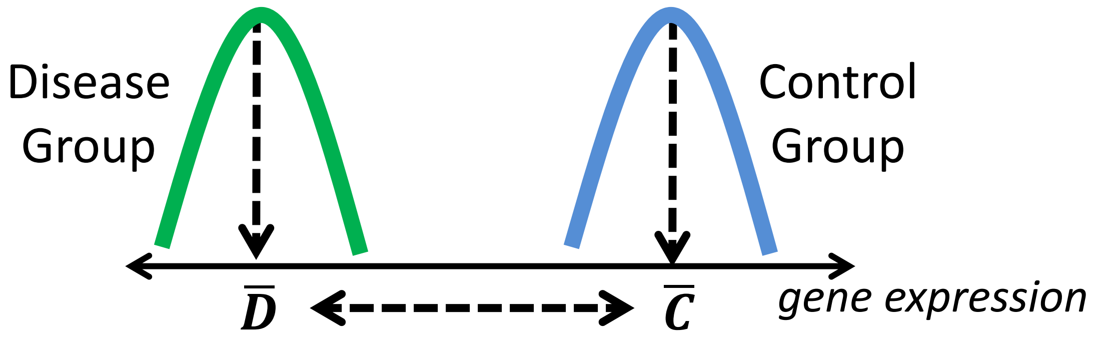


Historical Note

- 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

Difference in Averages



Disease Group

Control Group

$\overline{D}$

$\overline{C}$

gene expression

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

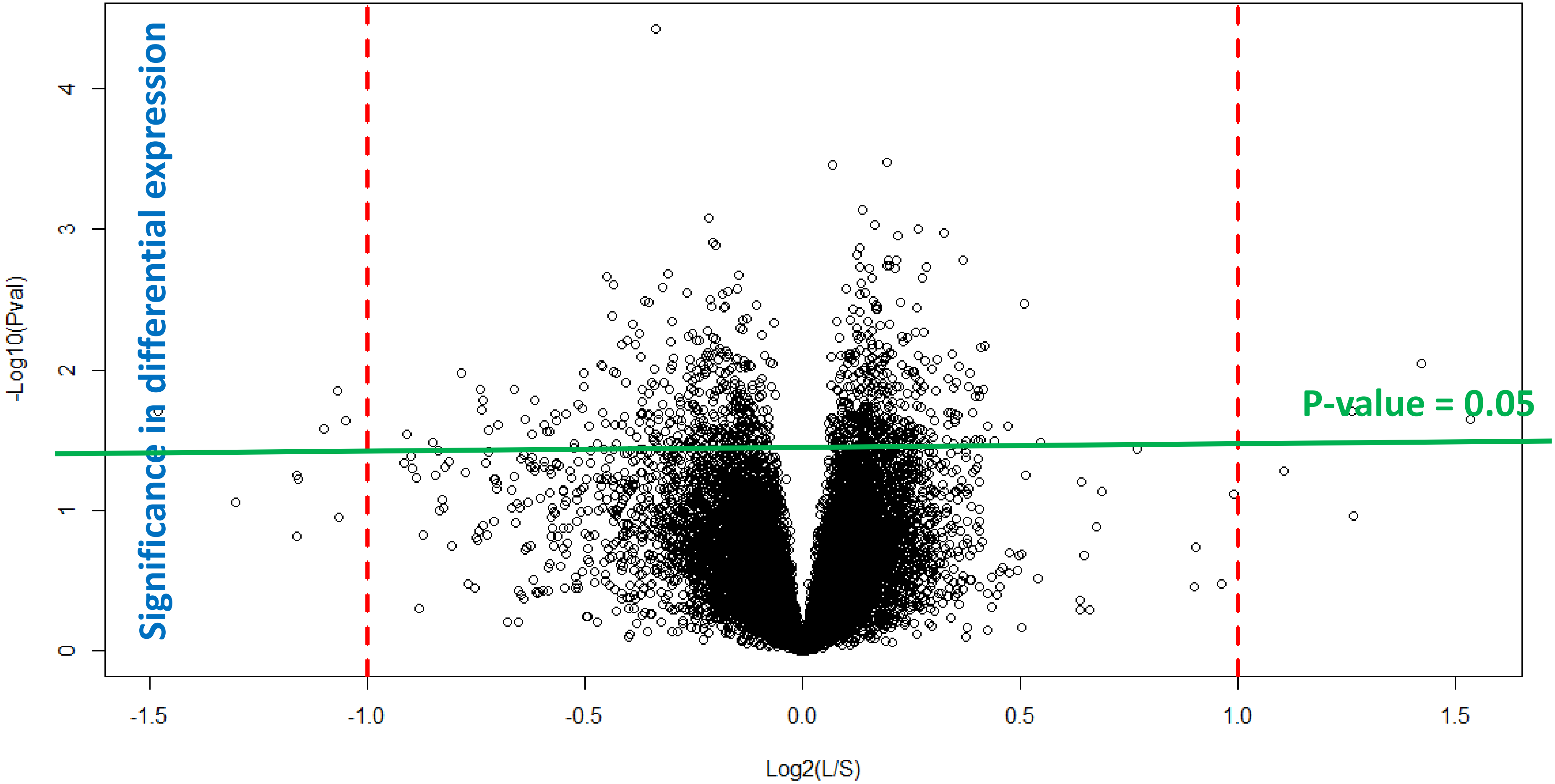
```
> runTP <- function(x,y){
                  res <- t.test(x[y=="HighLonely"],
                                x[y=="LowLonely"])
                  p <- res$p.value
                  return(p)
            }
> tpvals <- apply(edat, 1, runTP, y=lonely.index)
> length(tpvals)
> sum(tpvals < 0.05)
```

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



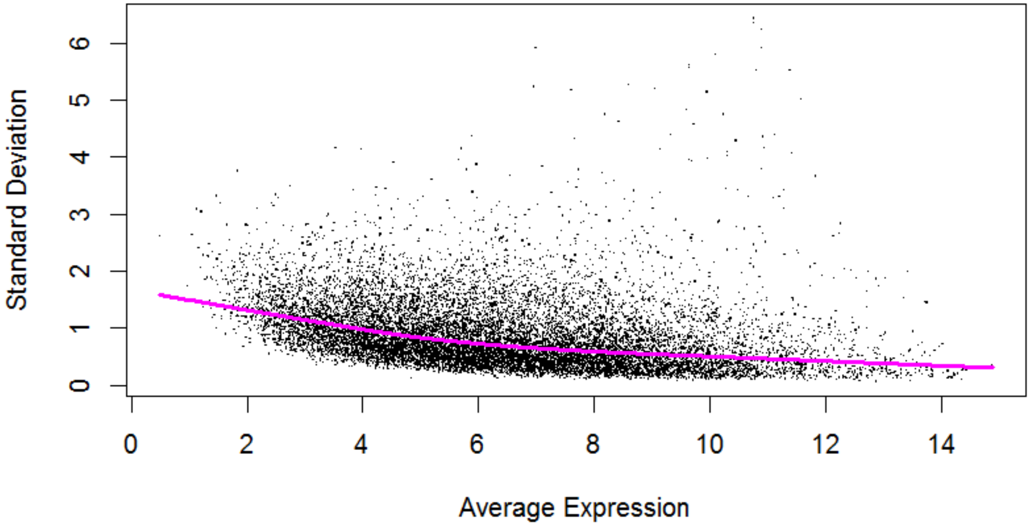
Log$_2$ Expression Fold Change shows the direction of gene expression in one condition relative to another

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:

$$\boxed{\boldsymbol{T}_{(\boldsymbol{gene})} = \frac{\overline{\boldsymbol{D}} - \overline{\boldsymbol{C}}}{\boldsymbol{f}(\boldsymbol{Var}(\boldsymbol{D},\boldsymbol{C}) + \alpha)}}$$



*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).

1. Meta-Analysis

Building evidence for consistent trends across multiples lines of data sources and experiments.

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

2. Integrating with external sources of information.

Interpreting results using pathways, gene sets or other properties of interest from the literature.

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

3. Integrating different types of genome-wide data.

Modeling related high-throughput data sets to identify multi-level regulatory events.

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`

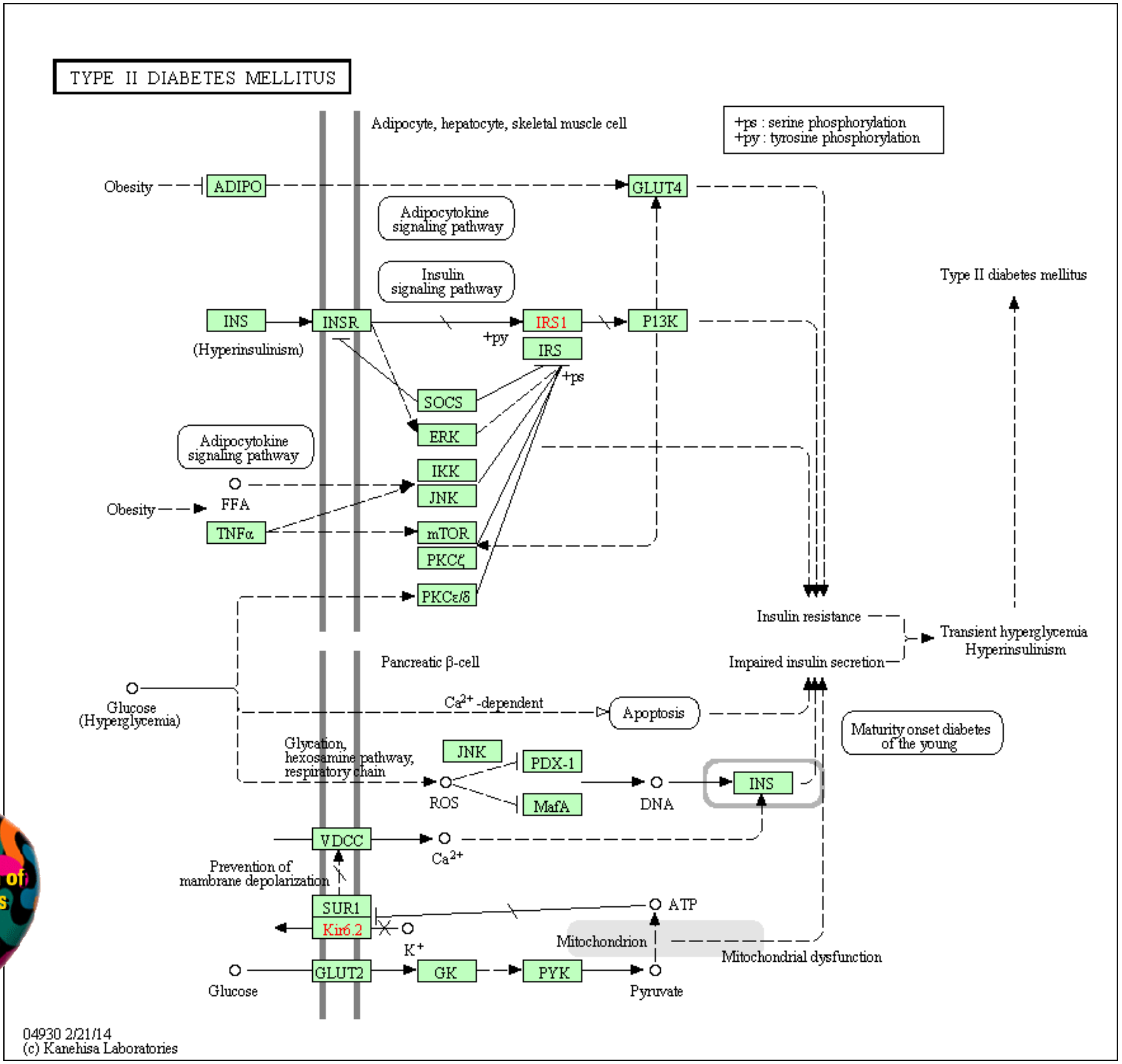
Gene Ontology

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

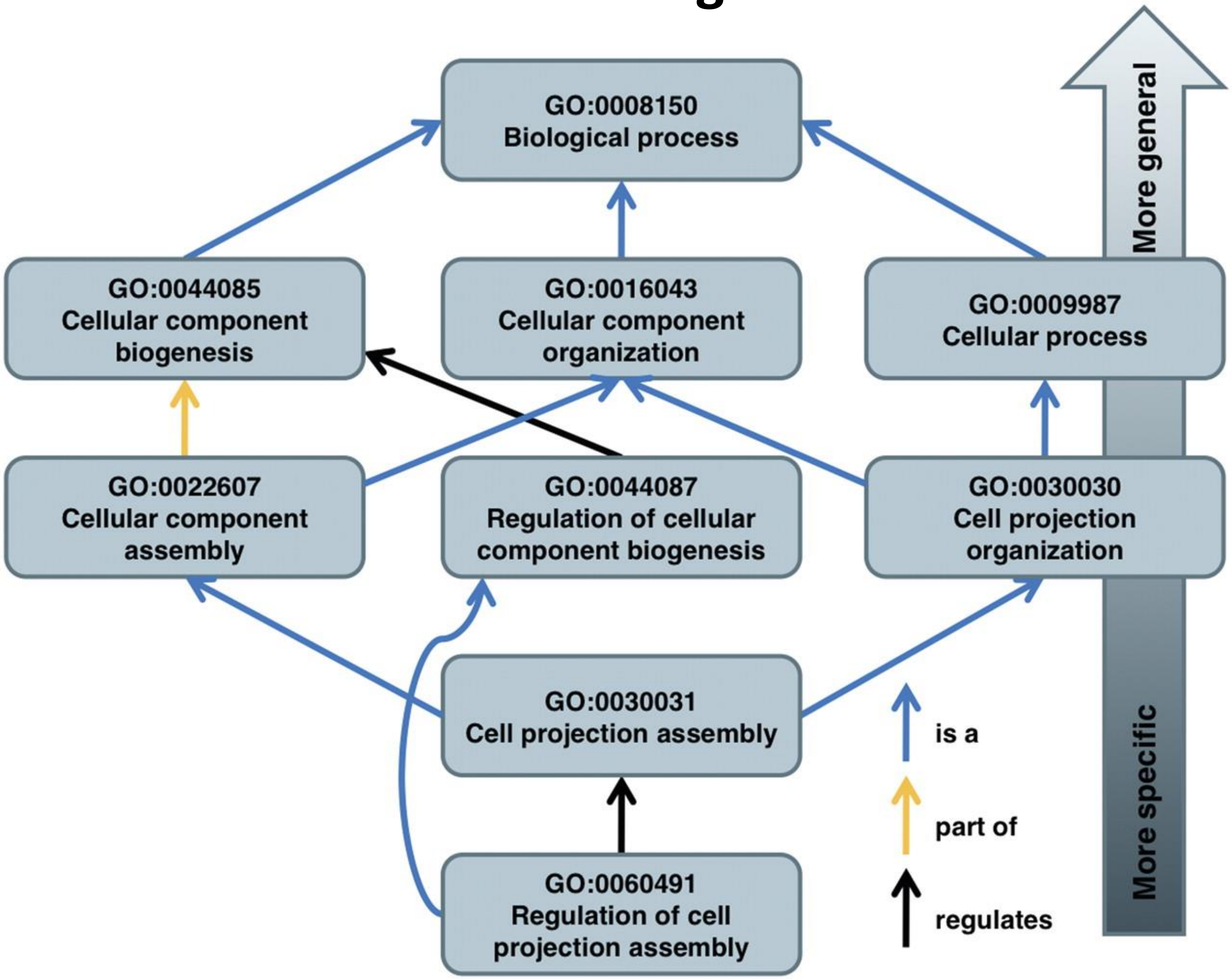
One of the most comprehensive sources of gene set information; there are 7 major groups, some of which overlap with the above resources.

KEGG Pathway: Type II Diabetes (human)


TYPE II DIABETES MELLITUS
Adipocyte, hepatocyte, skeletal muscle cell
+ps : serine phosphorylation
+py : tyrosine phosphorylation
Obesity
ADIPO
GLUT4
Adipocytokine signaling pathway
Insulin signaling pathway
Type II diabetes mellitus
INS
(Hyperinsulinism)
INSR
IRS1
+py
IRS
P13K
+ps
SOCS
ERK
Adipocytokine signaling pathway
IKK
JNK
Obesity
FFA
TNFα
mTOR
PKCζ
PKCε/δ
Insulin resistance
Transient hyperglycemia
Hyperinsulinism
Impaired insulin secretion
Pancreatic β-cell
Glucose
(Hyperglycemia)
Ca2+ -dependent
Apoptosis
Maturity onset diabetes of the young
Glycation, hexosamine pathway, respiratory chain
JNK
PDX-1
MafA
ROS
DNA
INS
VDCC
Ca2+
Prevention of mambrane depolarization
SUR1
Kir6.2
K+
ATP
Mitochondrion
Mitochondrial dysfunction
Glucose
GLUT2
GK
PYK
Pyruvate
04930 2/21/14
(c) Kanehisa Laboratories

Kyoto Encyclopedia of Genes and Genomes

Gene Ontology: a computational representation of biological knowledge

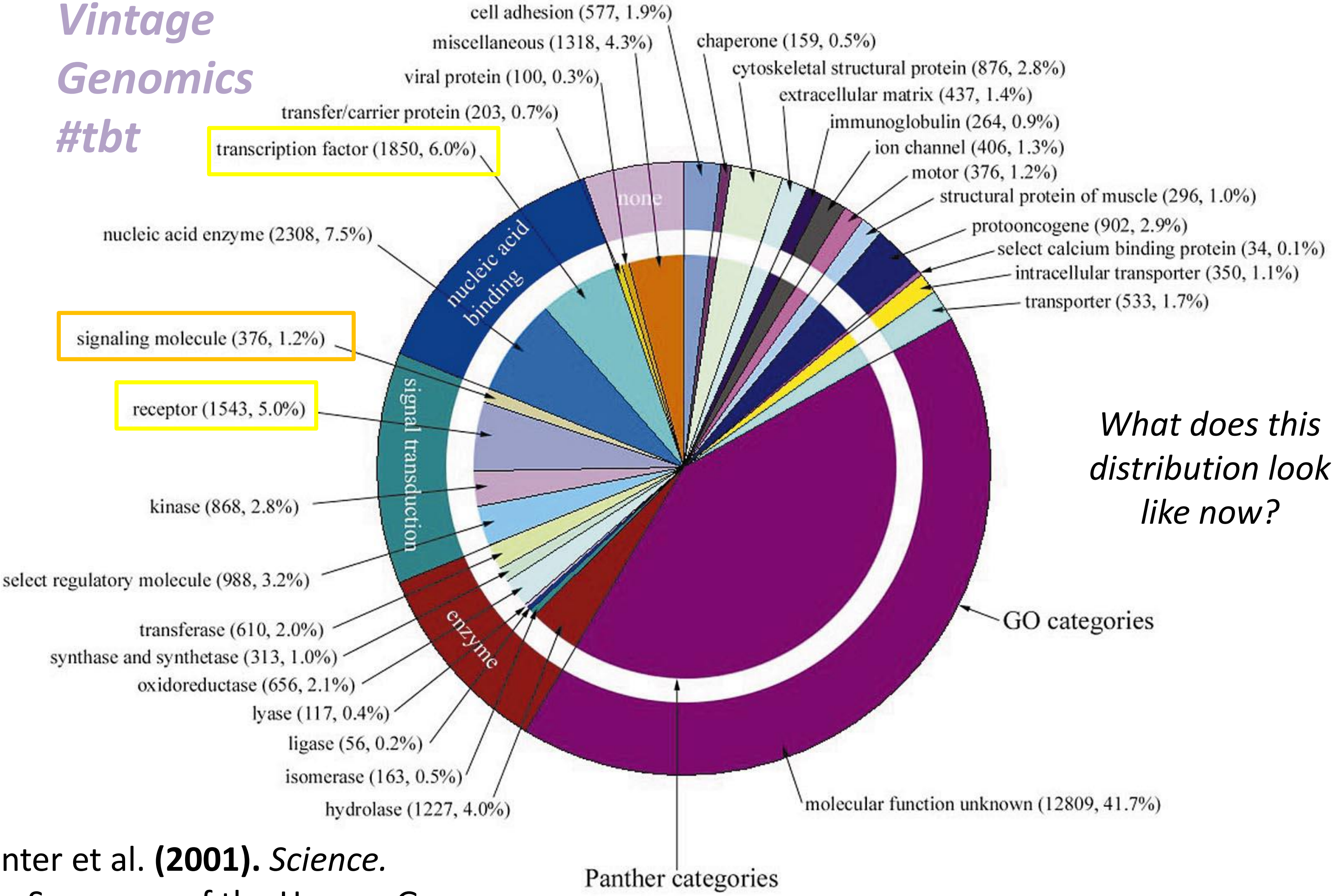

GO:0008150
Biological process
GO:0044085
Cellular component biogenesis
GO:0016043
Cellular component organization
GO:0009987
Cellular process
GO:0022607
Cellular component assembly
GO:0044087
Regulation of cellular component biogenesis
GO:0030030
Cell projection organization
GO:0030031
Cell projection assembly
GO:0060491
Regulation of cell projection assembly
is a
part of
regulates
More general
More specific

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

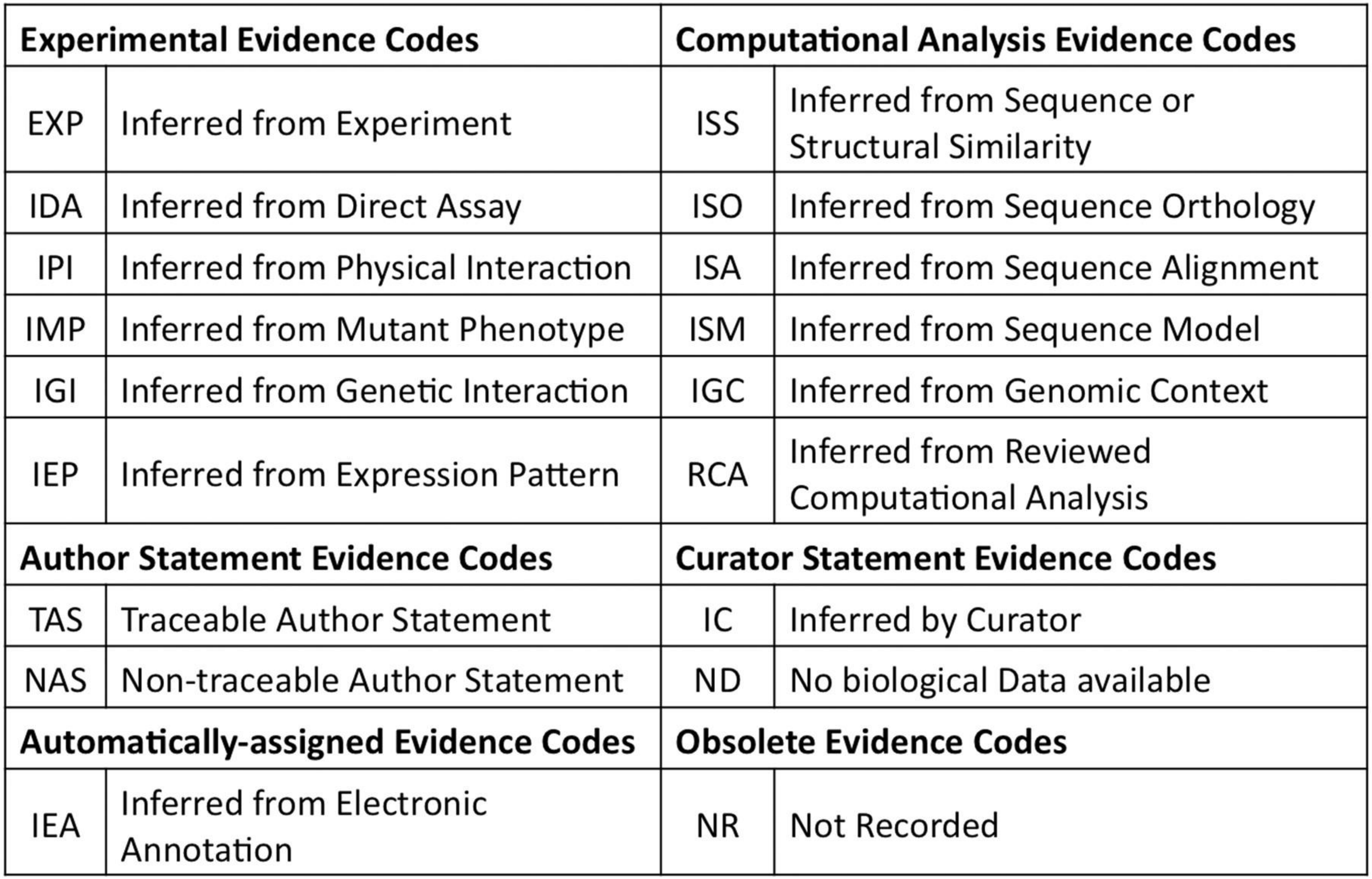


What does this distribution look like now?

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

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

ZFPM1
EXOC6
COX4I1
ECH1
ZMAT3
ECM2
PORCN
IL13RA1
RPPH1
SCRN1
TRAK1
HBEGF
WDR12
RFX1

NLGN2
GLIPR2
PLXDC2
MGP
BTF3L4
OR11L1
EGFLAM
NELFB
NR2F2
TMSB15B
SNAPC4
DKK3
STX2
HSPA1A

**What do these genes do?
Is there a significant over-representation of a certain pathway or important gene set?**

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.

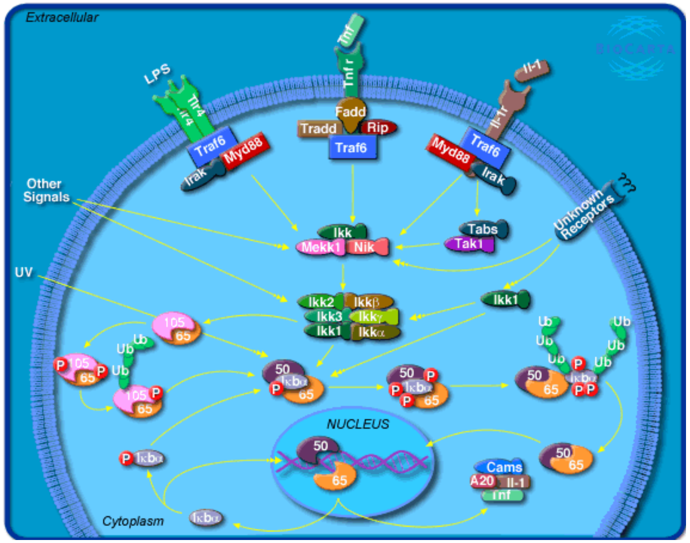
		VARIABLE A		
		YES	NO	
VARIABLE B	NO	x	y	S_Z
	YES	j	k	
		S_X	S_Y	

$$P(X=x)=\frac{\binom{S_X}{x}\binom{S_Y}{y}}{\binom{S_X+S_Y}{S_Z}}=\frac{\binom{S_X}{x}\binom{S_Y}{y}}{\binom{S_X+S_Y}{x+y}} \qquad \text{for } \max(0,S_Z-y)\le x\le \min(S_Z,S_X)$$

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

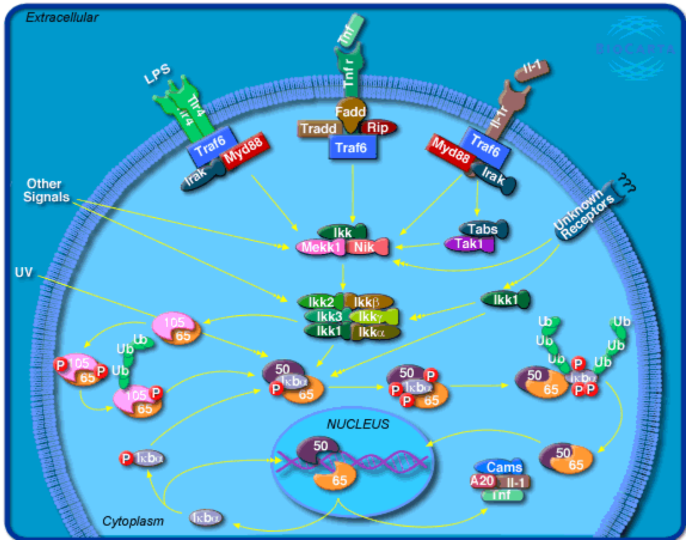
$$P(X = 70) = \frac{\binom{100}{70}\binom{120}{80}}{\binom{220}{150}} =$$

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

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

Applying Fisher’s Exact Test in R

NFκB Pathway (BioCarta)

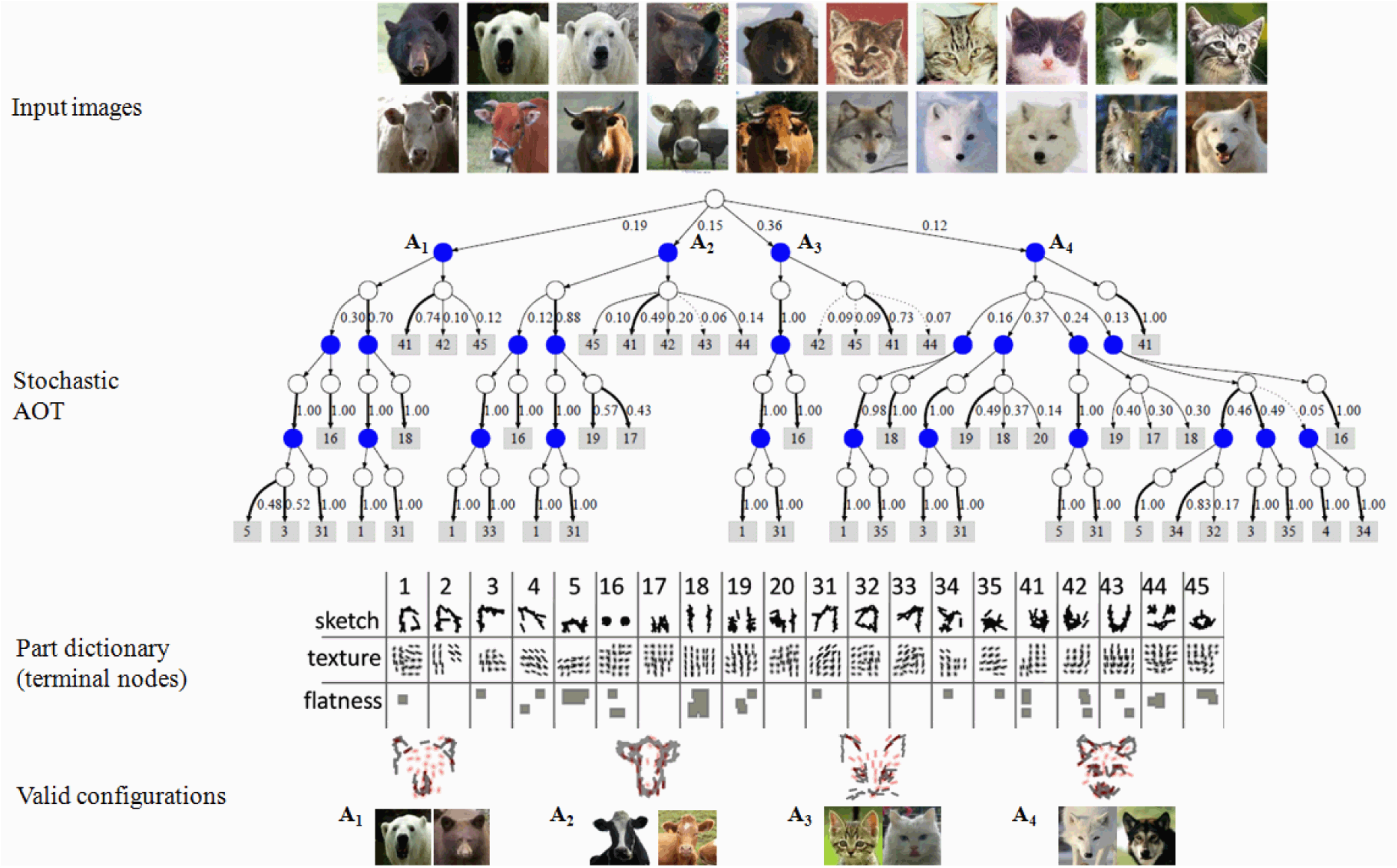


	Interesting Gene List	*Not in Interesting Gene List*	
Genes in Pathway Z	70	*y*	150
Genes Not in Pathway Z	*j*	*k*	
	100	120	

```
> 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

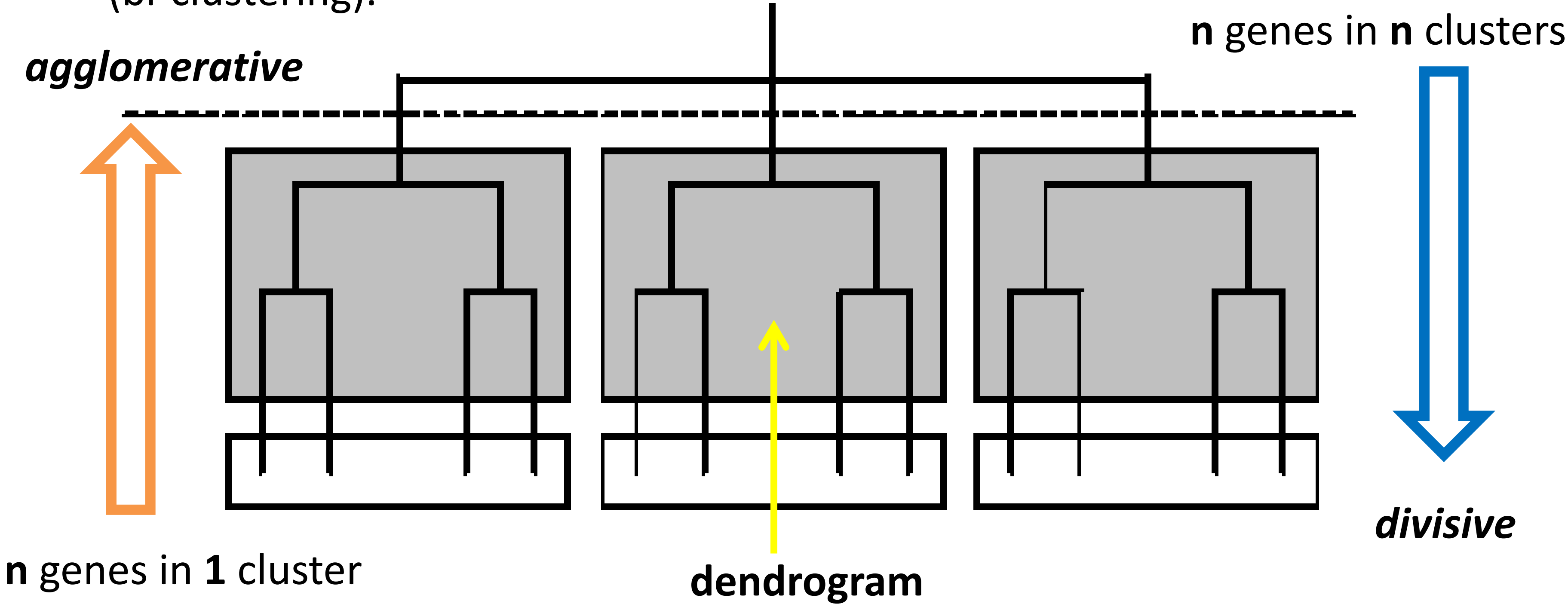
<u>Unsupervised learning</u> is the task of identifying patterns in the presence of many data variables where the number of patterns is also not known.



[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).



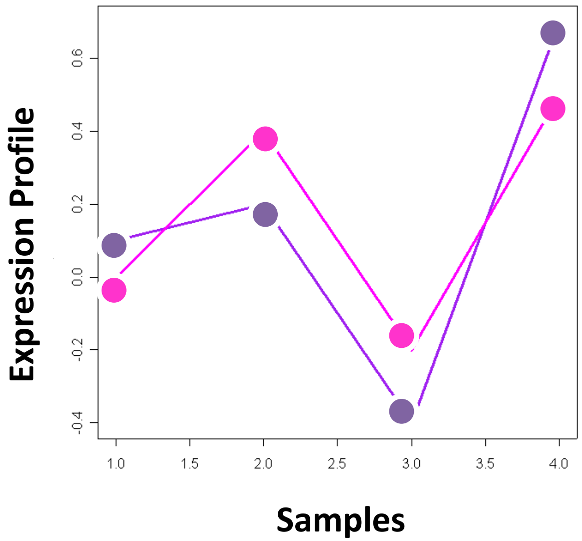
- We join nodes based on the notion of maximum **'similarity'**.
- Equivalently, we break nodes based on minimal similarity

Source: J-Express Manual

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

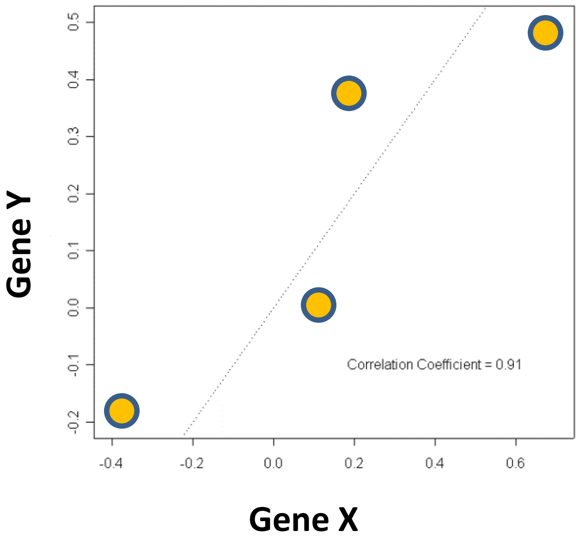
Consider expression profiles of Gene X and Gene Y: how do we score their similarity?

Euclidean Distance



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

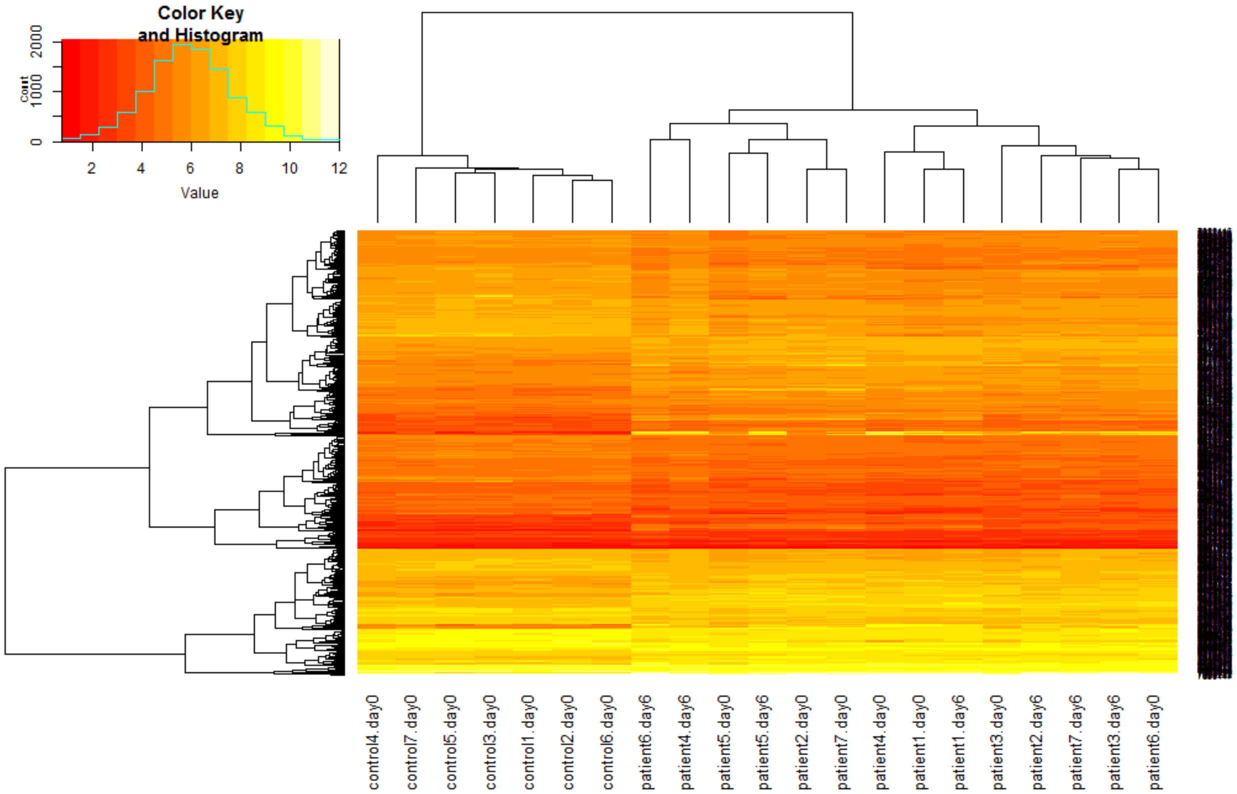
Correlation Coefficient



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

Heatmaps!

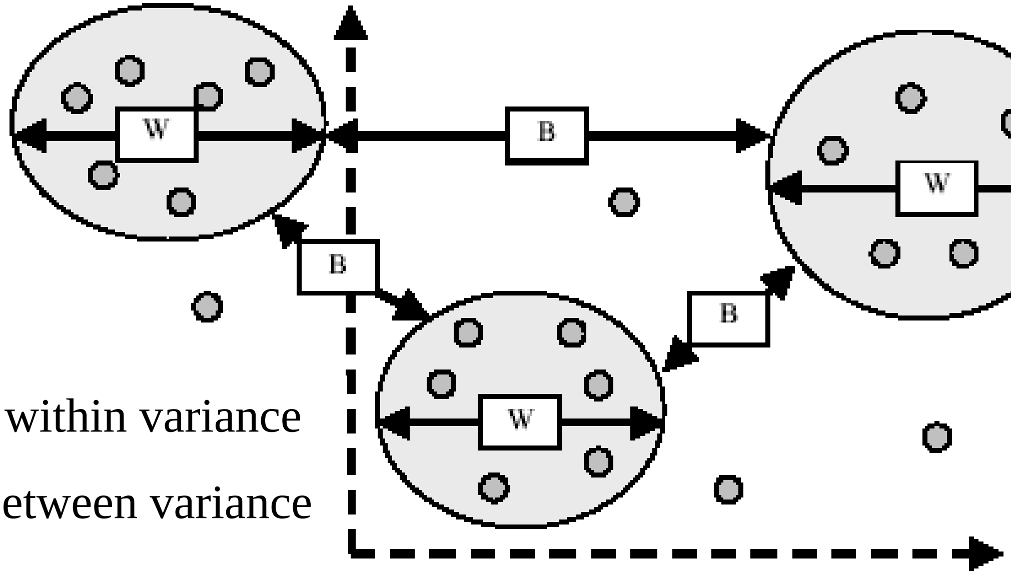
```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("gplots")
> library(gplots)
> heatmap.2(edat.sig, trace="none", margins=c(8,8))
```



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.



W – within variance

B – between variance

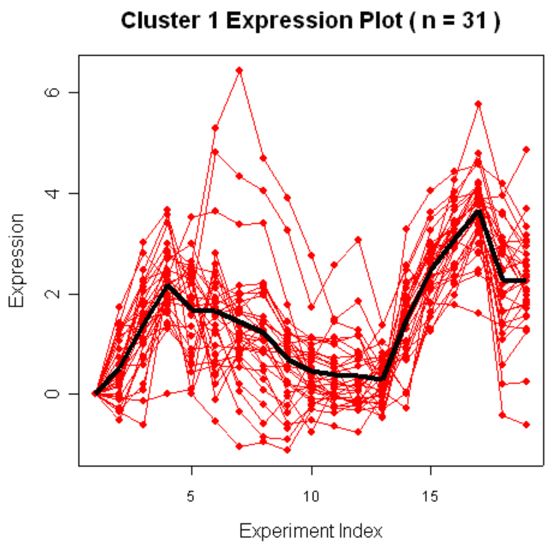
Reference: J-Express manual

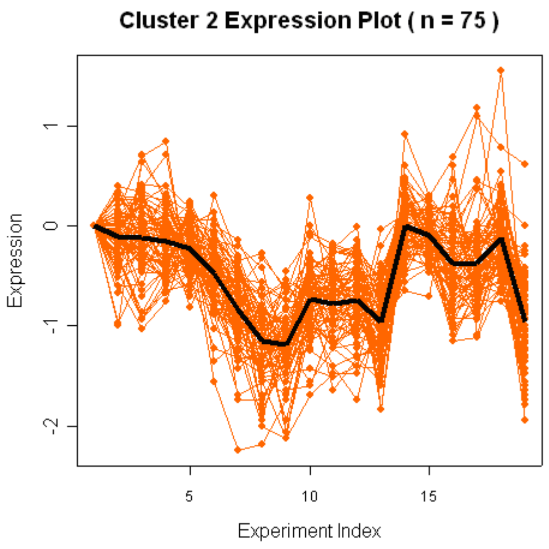
General Framework of a K-means Algorithm

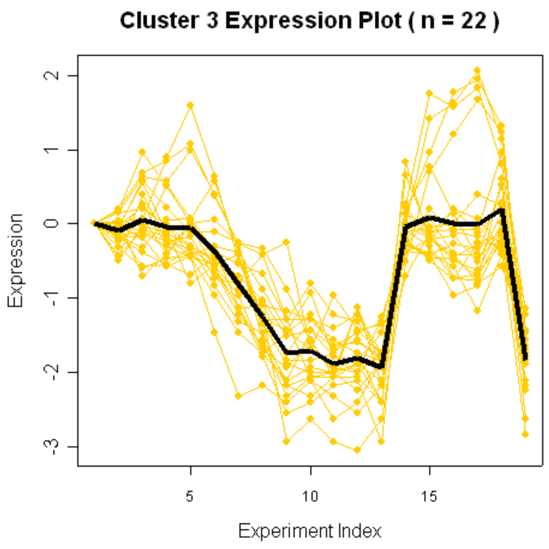
Step 0: Start with a defined number of clusters.
Step 1: Initialize clusters; usually based on agglomerative hierarchical clusters.

Means = K-means.

Medians = K-medoids, PAM


Cluster 1 Expression Plot (n = 31)

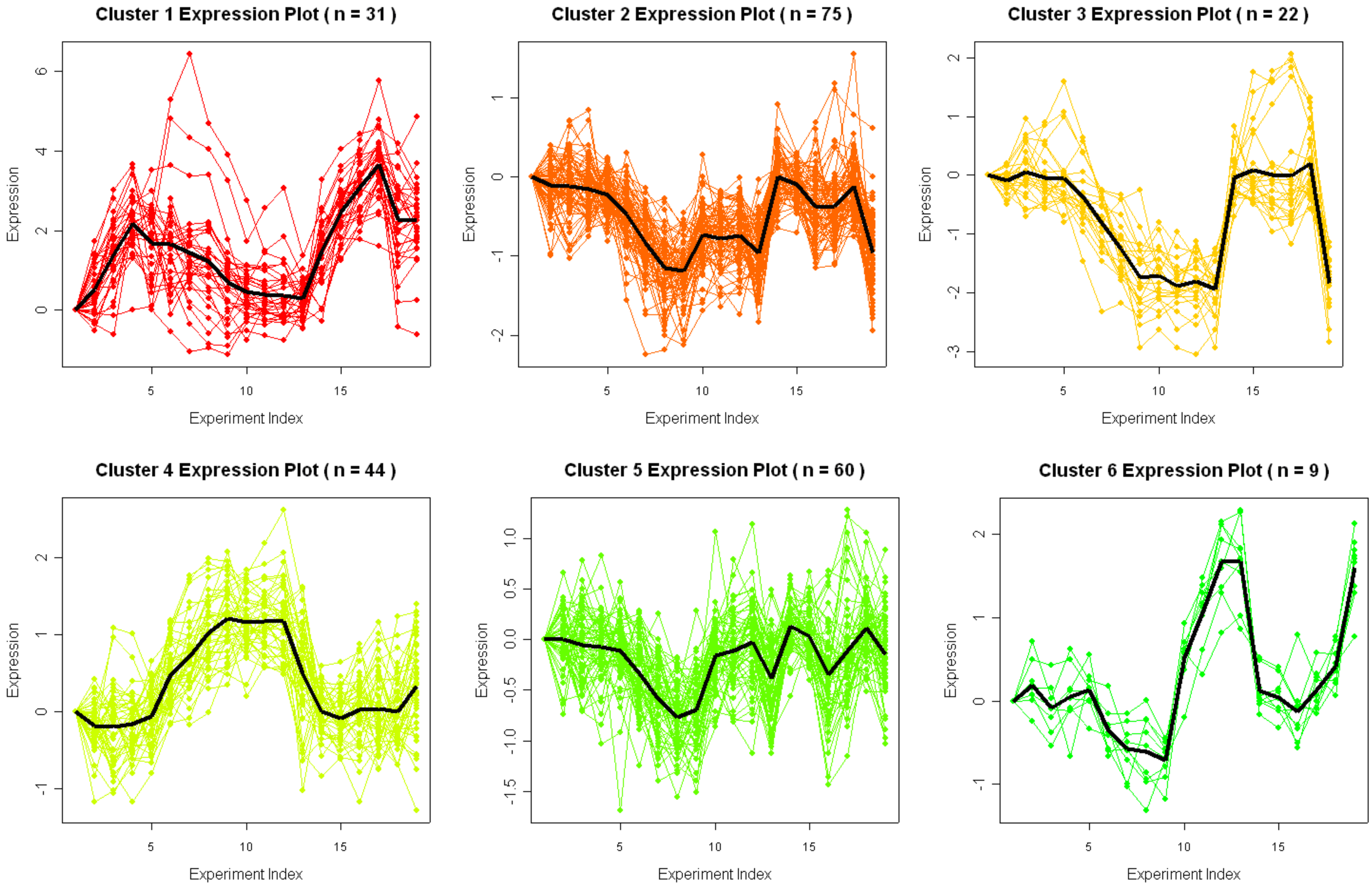

Cluster 2 Expression Plot (n = 75)


Cluster 3 Expression Plot (n = 22)

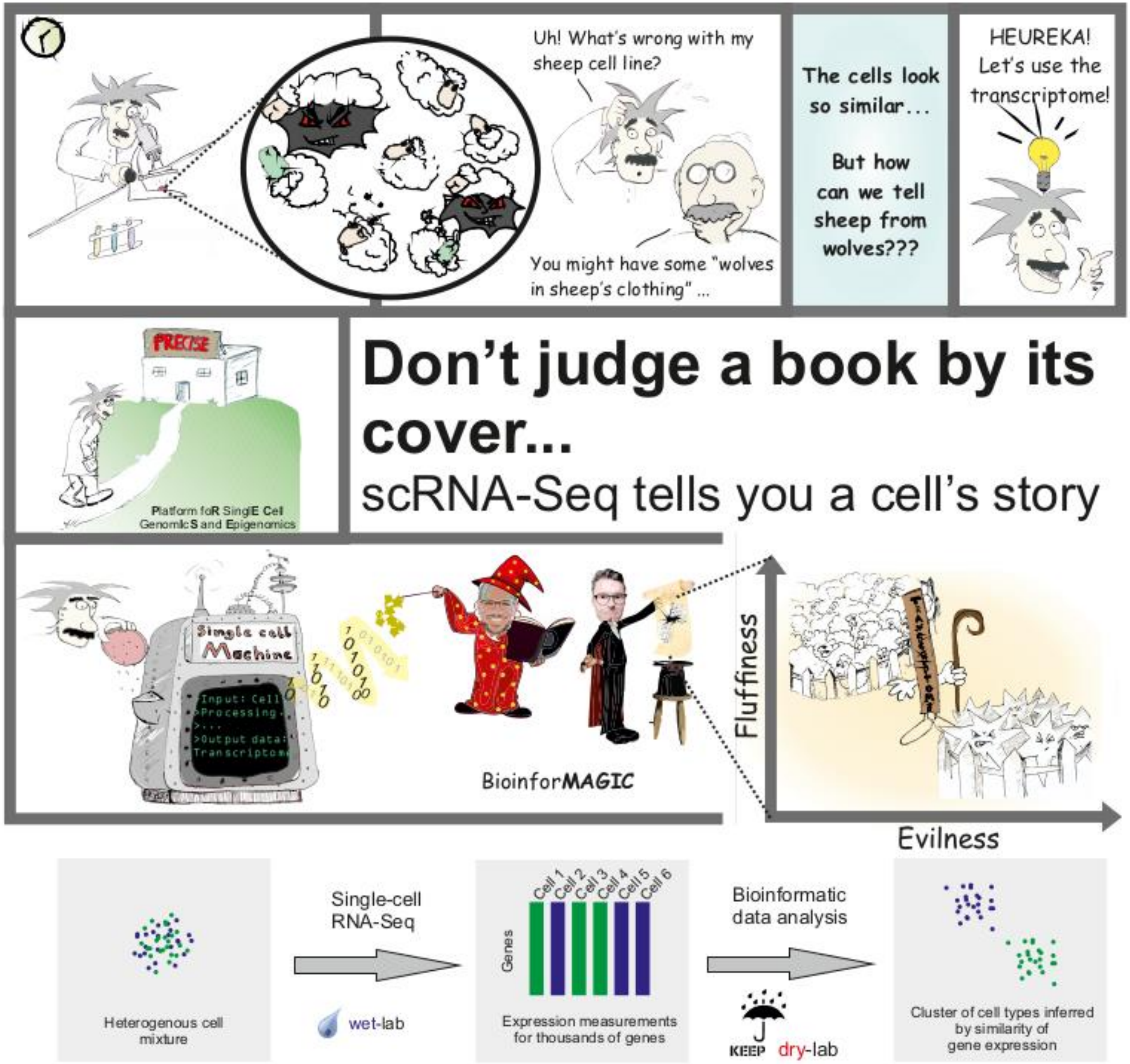
Step 2: Random sort of list, assign each gene to a cluster based on distance metric.
Step 3: Assess convergence criteria. If convergence achieved, stop. Otherwise repeat.

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.


Uh! What's wrong with my sheep cell line?
You might have some "wolves in sheep's clothing" ...
The cells look so similar...
But how can we tell sheep from wolves???
HEUREKA! Let's use the transcriptome!
PRECISE
Platform foR SinglE Cell GenomicS and Epigenomics
Don't judge a book by its cover...
scRNA-Seq tells you a cell's story
Single cell Machine
Input: Cell
>Processing
>...
>Output data:
Transcriptom
BioinforMAGIC
TRANSCRIPTOME
Fluffiness
Evilness
Heterogenous cell mixture
Single-cell RNA-Seq
wet-lab
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
Cell 6
Genes
Expression measurements for thousands of genes
Bioinformatic data analysis
KEEP dry-lab
Cluster of cell types inferred by similarity of gene expression

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

Biology occurs on many different scales

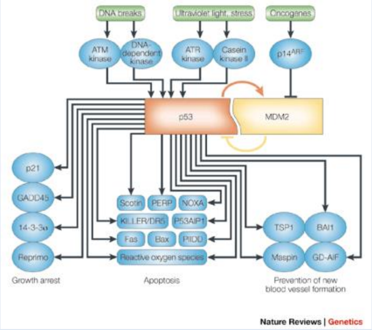
Breast Cancer: a disease-related example

Multi-cellular organization

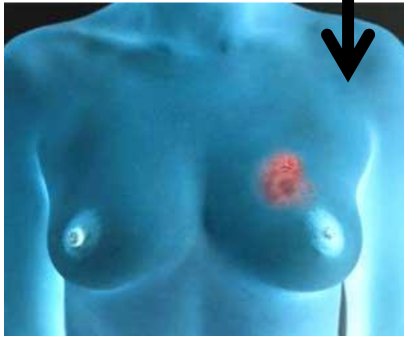
Behavior/physiology of cells:
Staging – TNM classification; 0 to 4

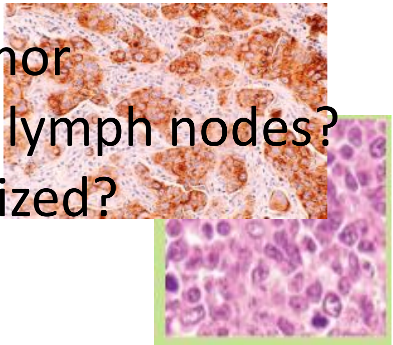
Single cell features:
Histopathology & Grade

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



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

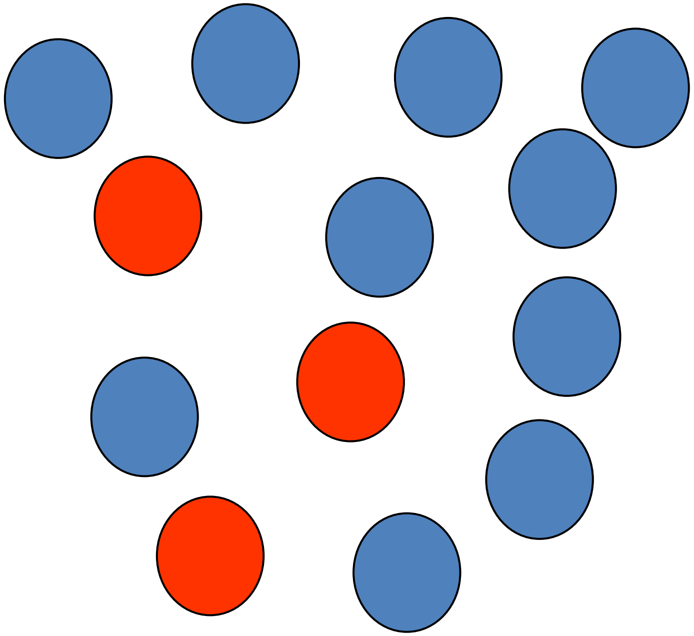




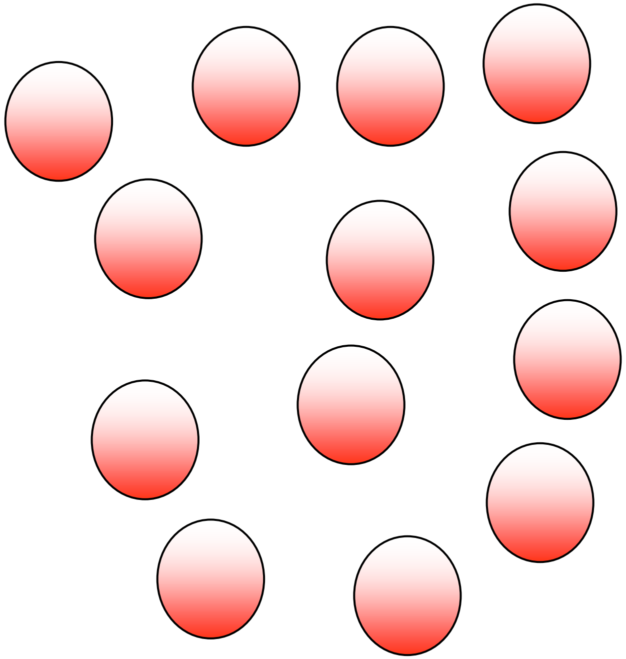
T = size of tumor
N = spread to lymph nodes?
M = metastasized?

Genome-wide
Microarrays & Sequencing

Cell populations are inherently heterogeneous



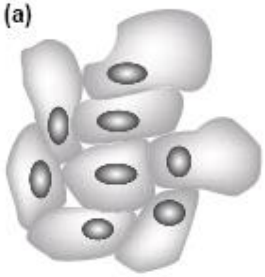
Ensemble methods survey the "average" transcriptome: microarrays, qPCR, RNA-seq



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

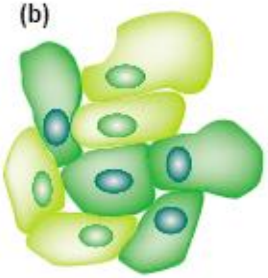
Conceptualizing gene expression in single cells

1980s:
Before single cell assays were invented:



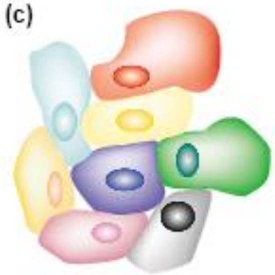
Cells were thought to be identical.

In situ hybridization in 1989 gave snapshots of individual nuclei.



Genes are either "on" or "off".

Single-cell gene expression profiling in 2001.

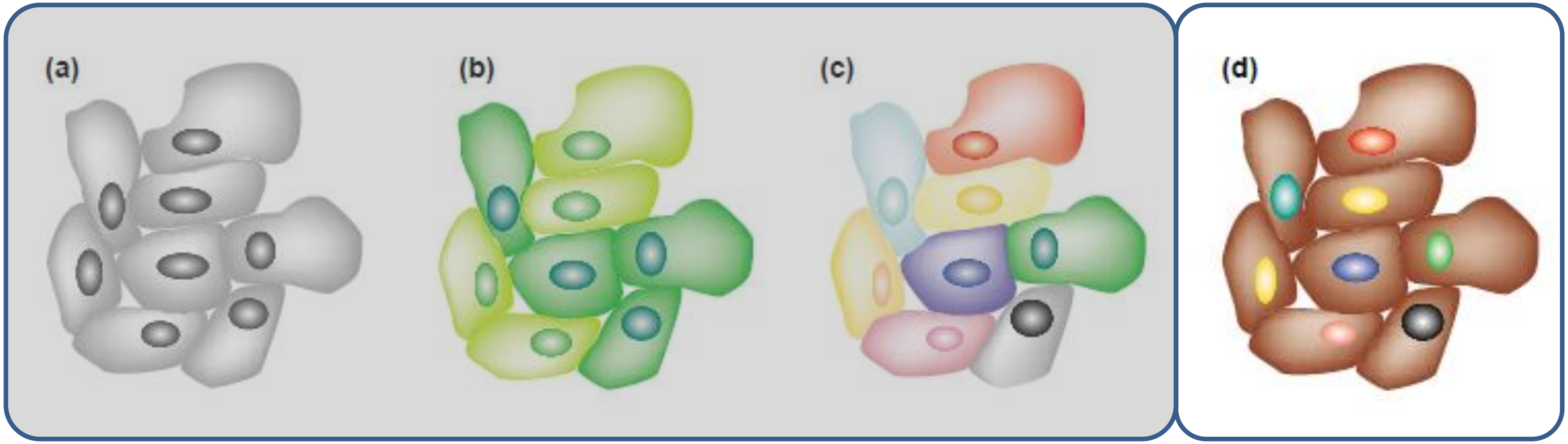


Cells express genes heterogeneously around a distribution of levels.

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

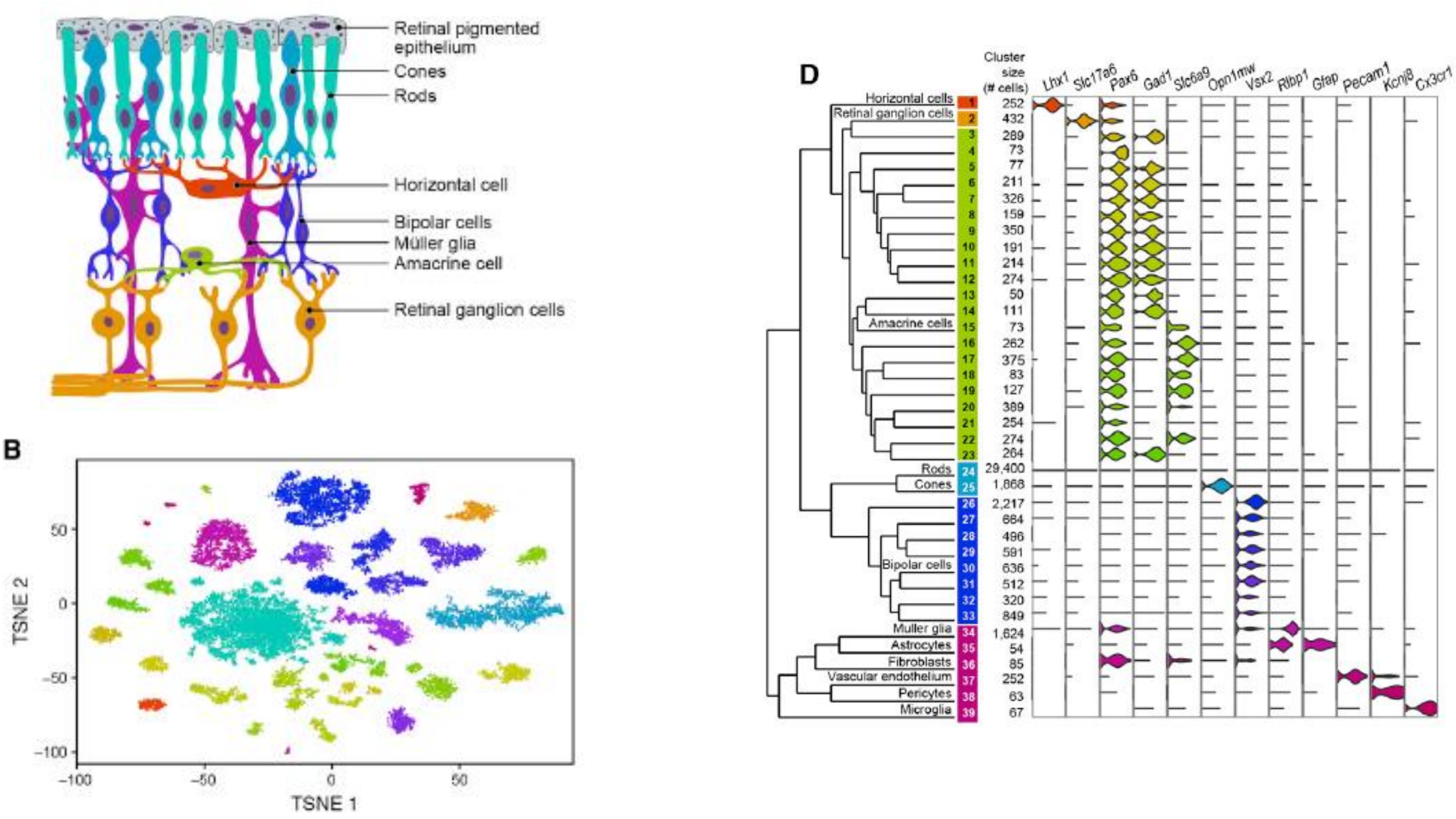
How are cells able to tolerate gene expression variability and maintain similar physiological function?



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

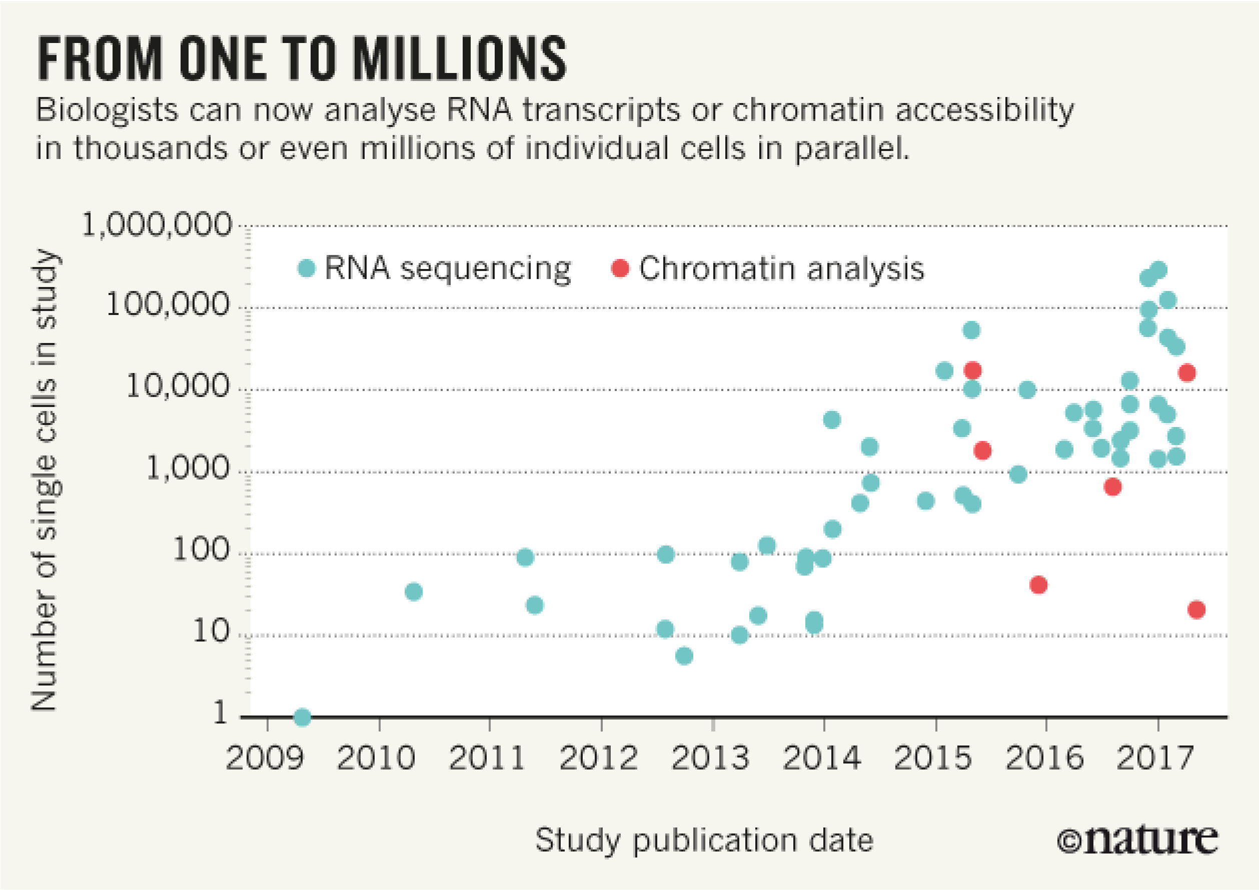
Global patterns of transcriptional regulation of cell type diversity are within reach!

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

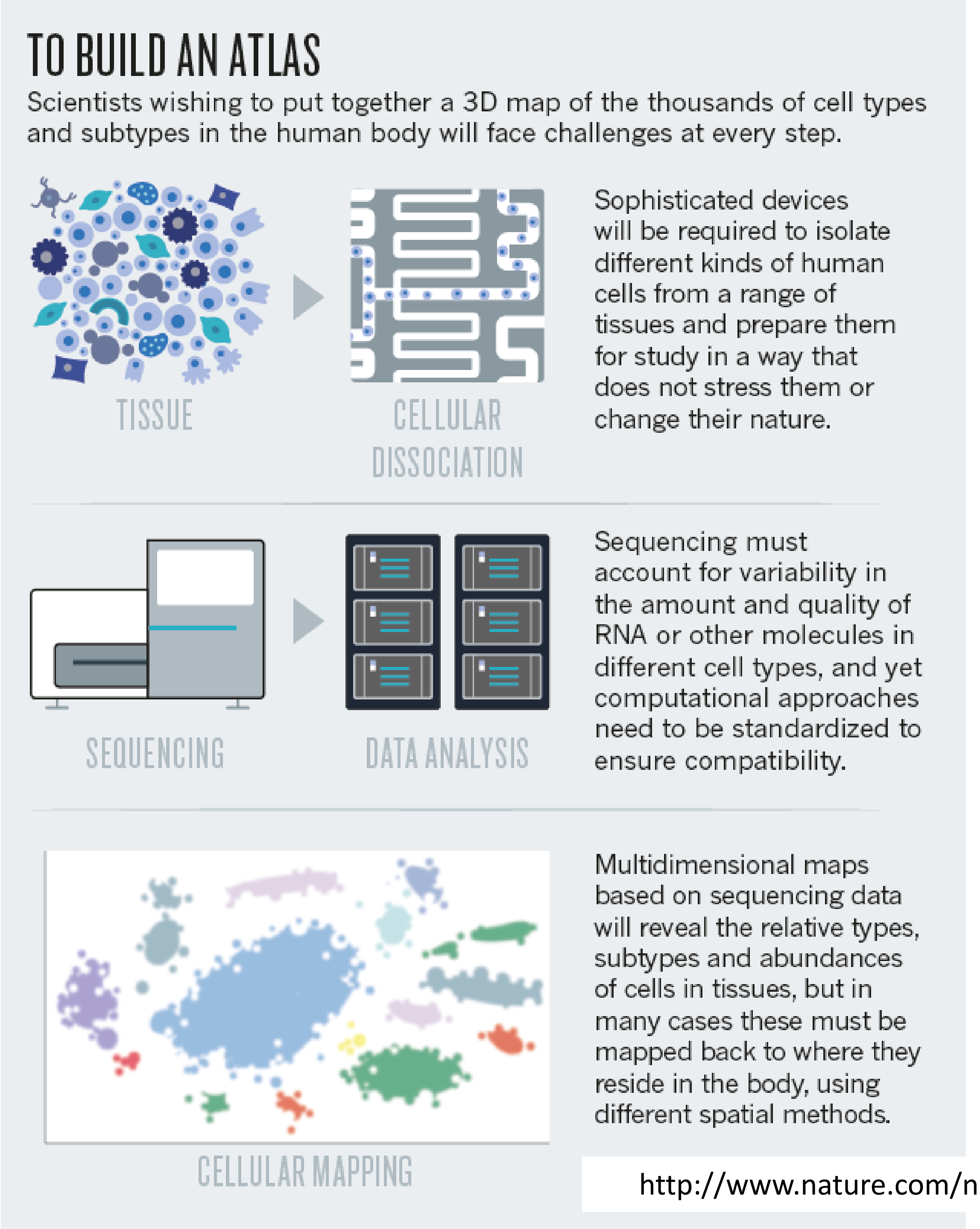


Macosko et al. **(2015).** *Cell.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.

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.

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.

CELLULAR MAPPING

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