Bioinformatics, statistics & transcriptomics

Yi Jin **LIEW**
Aranda Lab

080218

Outline

- Morning session
  - 9.00—10.00
    - Intro to bioinformatics
    - Why use Linux?
    - Why you shouldn’t fear the command line
  - 10.20—12.10
    - Statistics in transcriptomics
    - Q & A
- Lunch!
- Afternoon session (1330—1700)

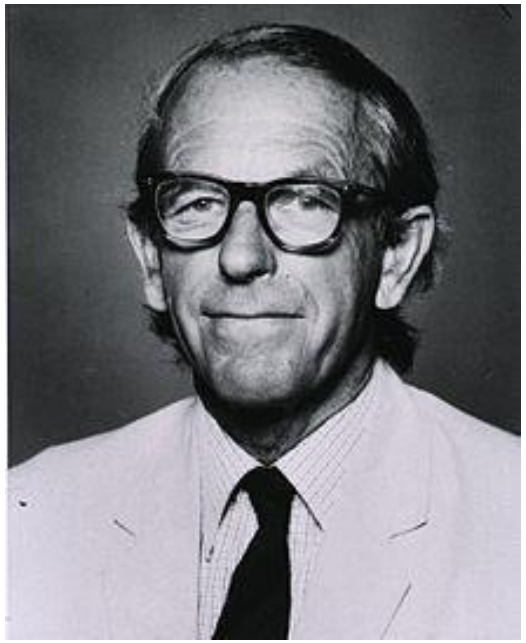
Outline

- Afternoon session (make sure you have a functioning terminal!)
  - Process raw sequencing reads: 1 hour
  - Read mapping and quantification with kallisto: 30 mins
  - Intermission II: 10 mins
  - Normalisation with sleuth: 30 mins
  - Functional enrichment with topGO: 50 mins

Part 0:
… bioinformatics?

Bioinformatics (= computational biology)

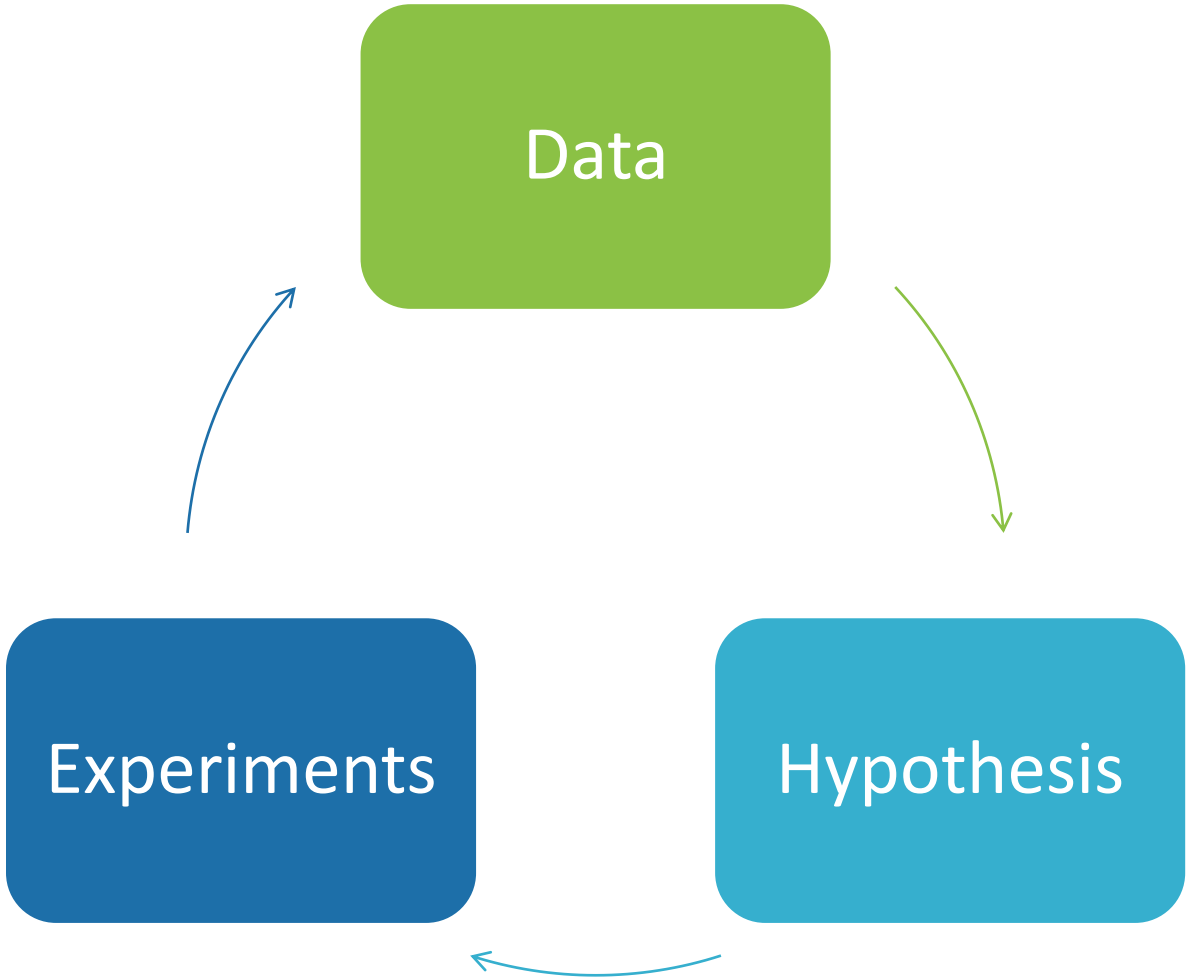
- Using computers to understand biological data
- Earliest bioinformatics problems were:
  - Aligning protein sequences (protein sequencing pioneered 1950s)
  - Studying bacteriophage genomes (DNA sequencing pioneered 1970s), reinforced concepts of codons, open reading frames etc.



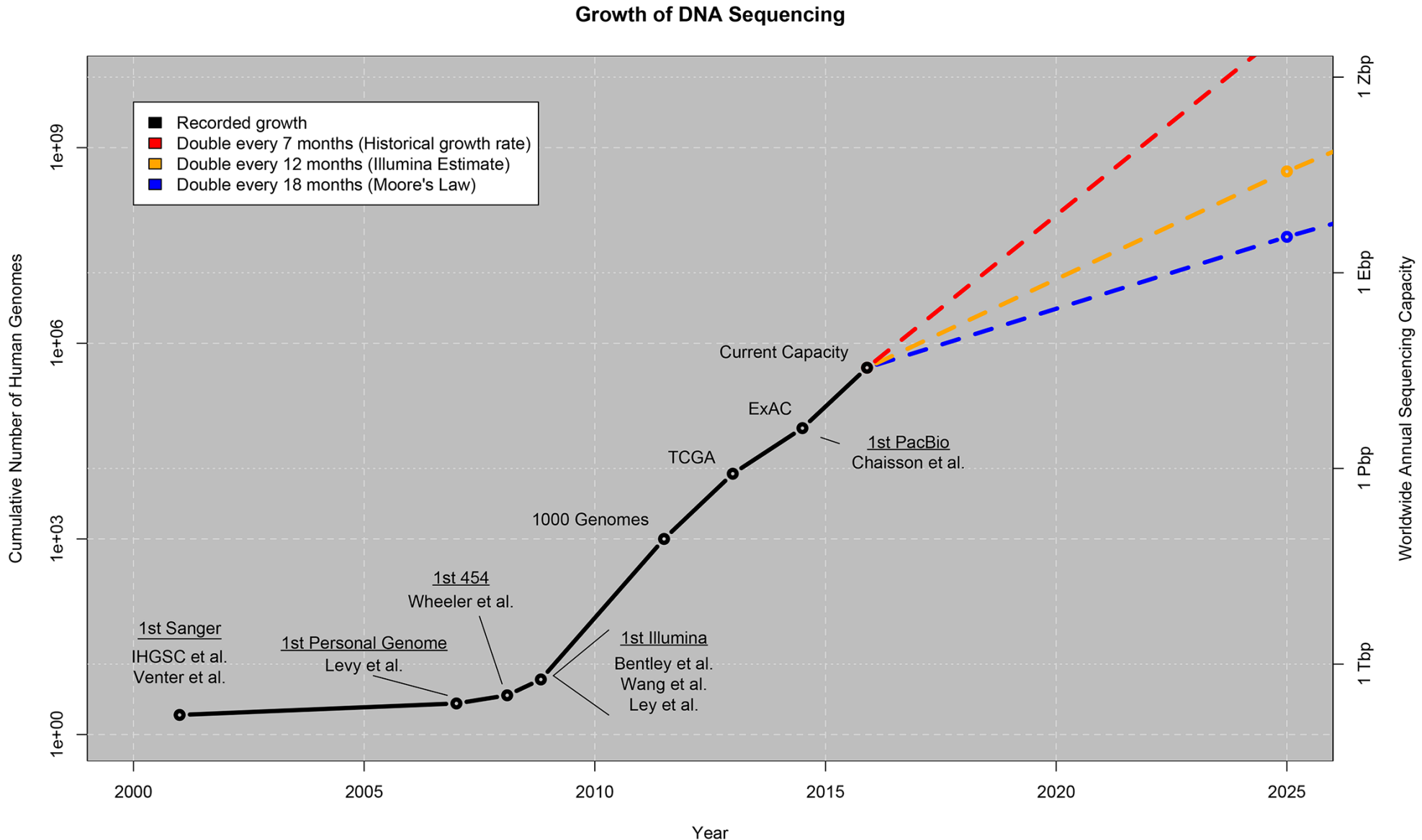
Fred Sanger
(1918 – 2013)

Bioinformatics (= computational biology)

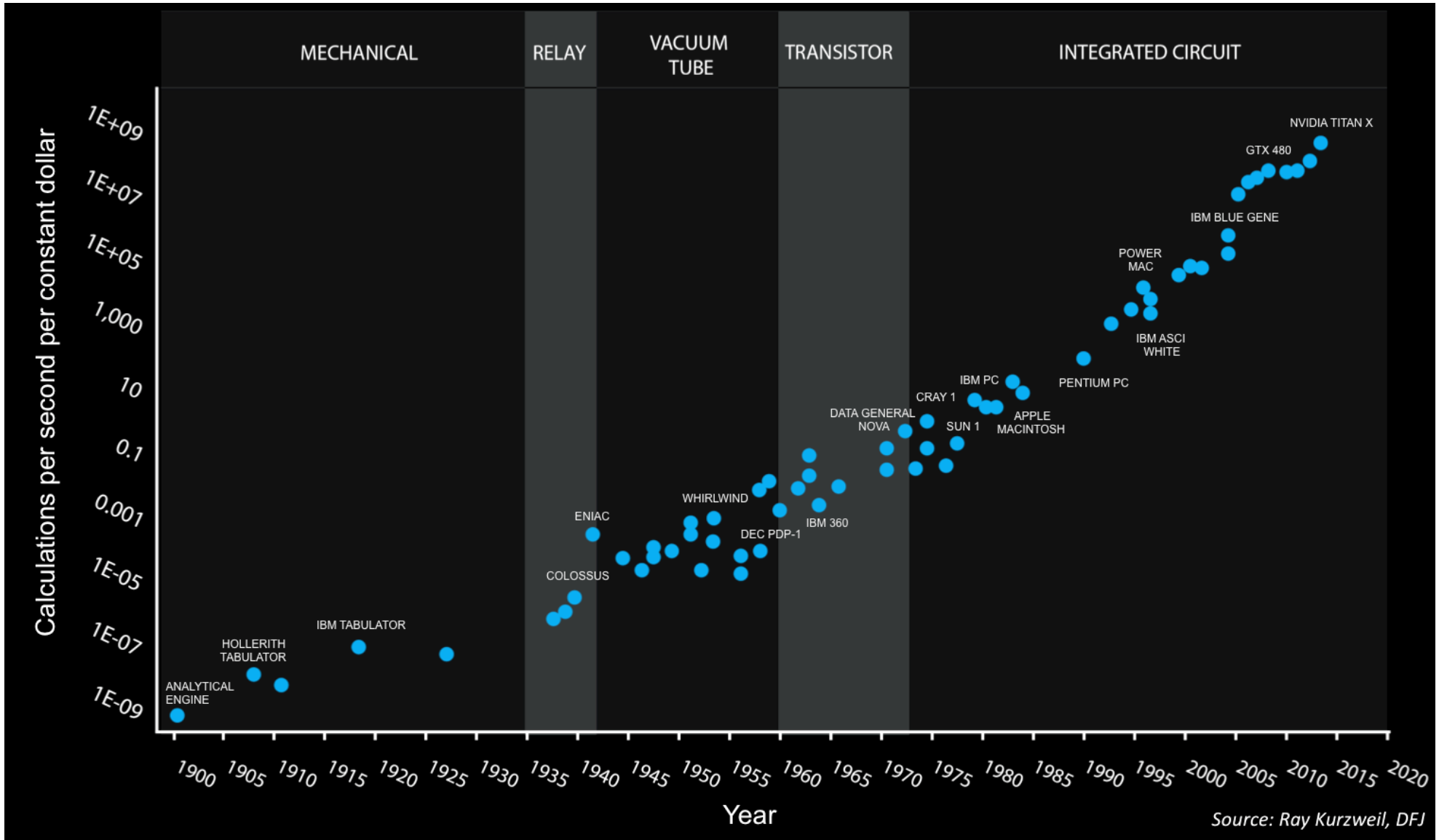
- Data-driven hypothesis



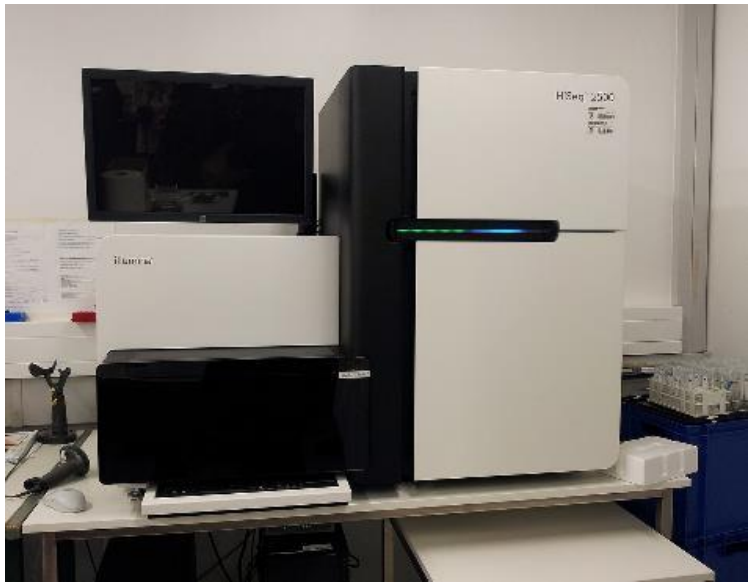
Explosive growth in data

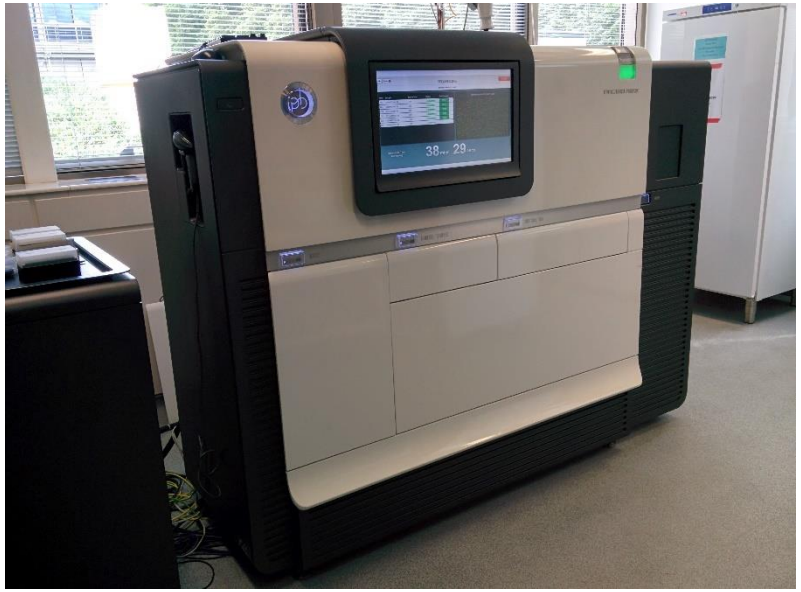


Exponential growth in computing power



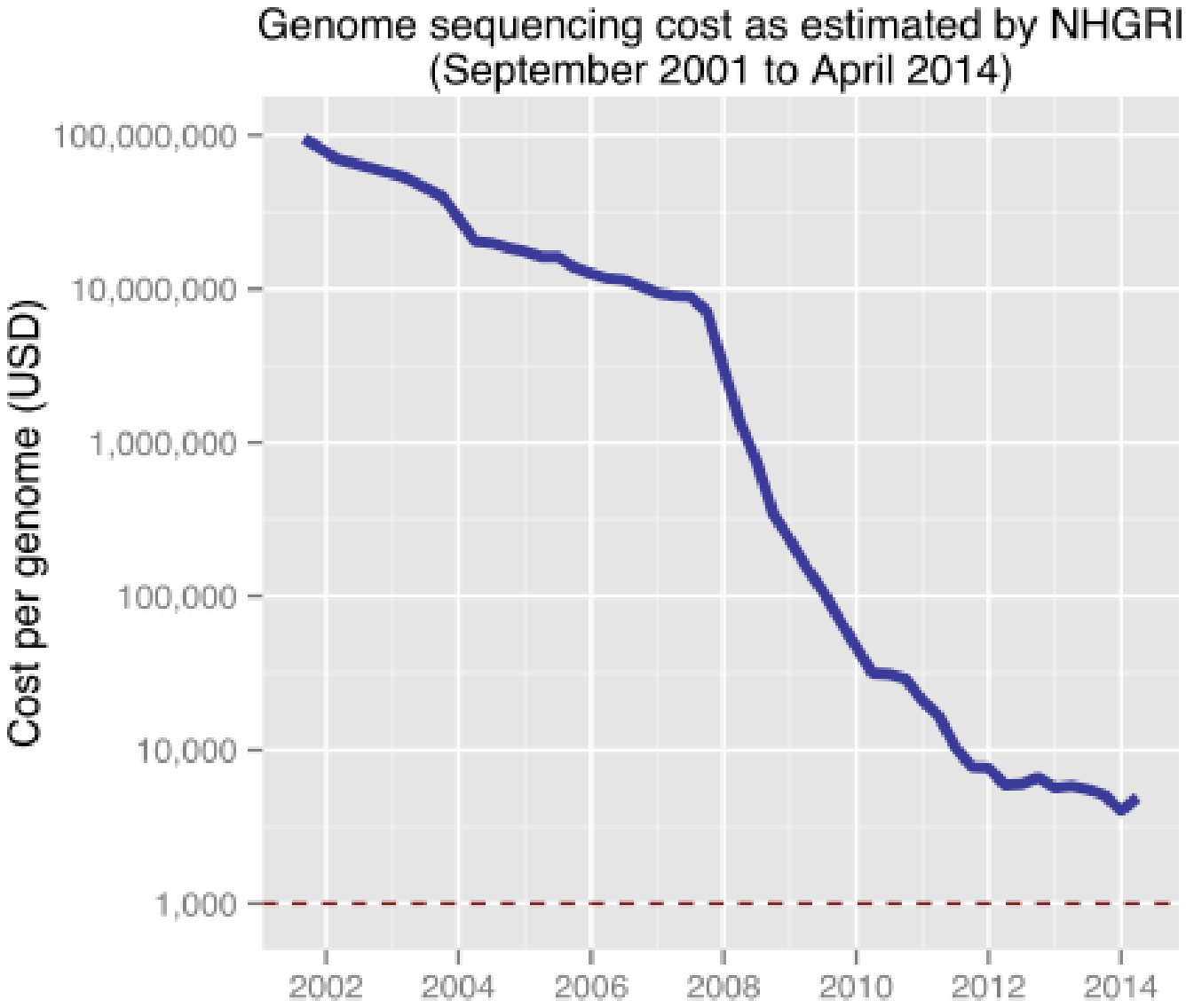
New sequencing technologies



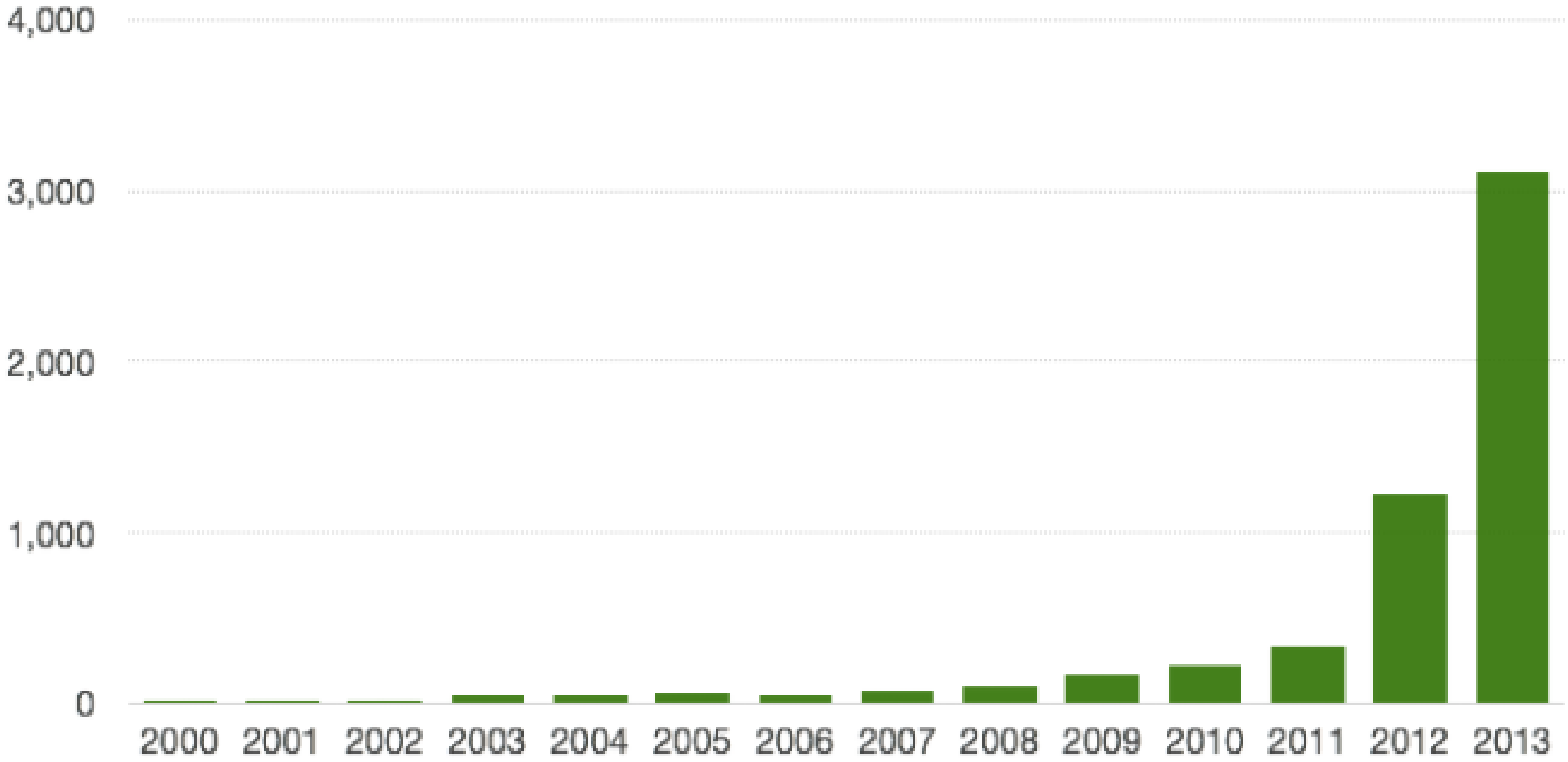


1. Sanger sequencer
2. Next-generation sequencing (Illumina / SOLiD / 454)
3. Single molecule sequencing (Pacific Bio / Ion Torrent)

New sequencing technologies



Explosive growth in papers



Growth of bioinformatics papers on Google Scholar that mention "big data".

Modern examples of bioinformatics

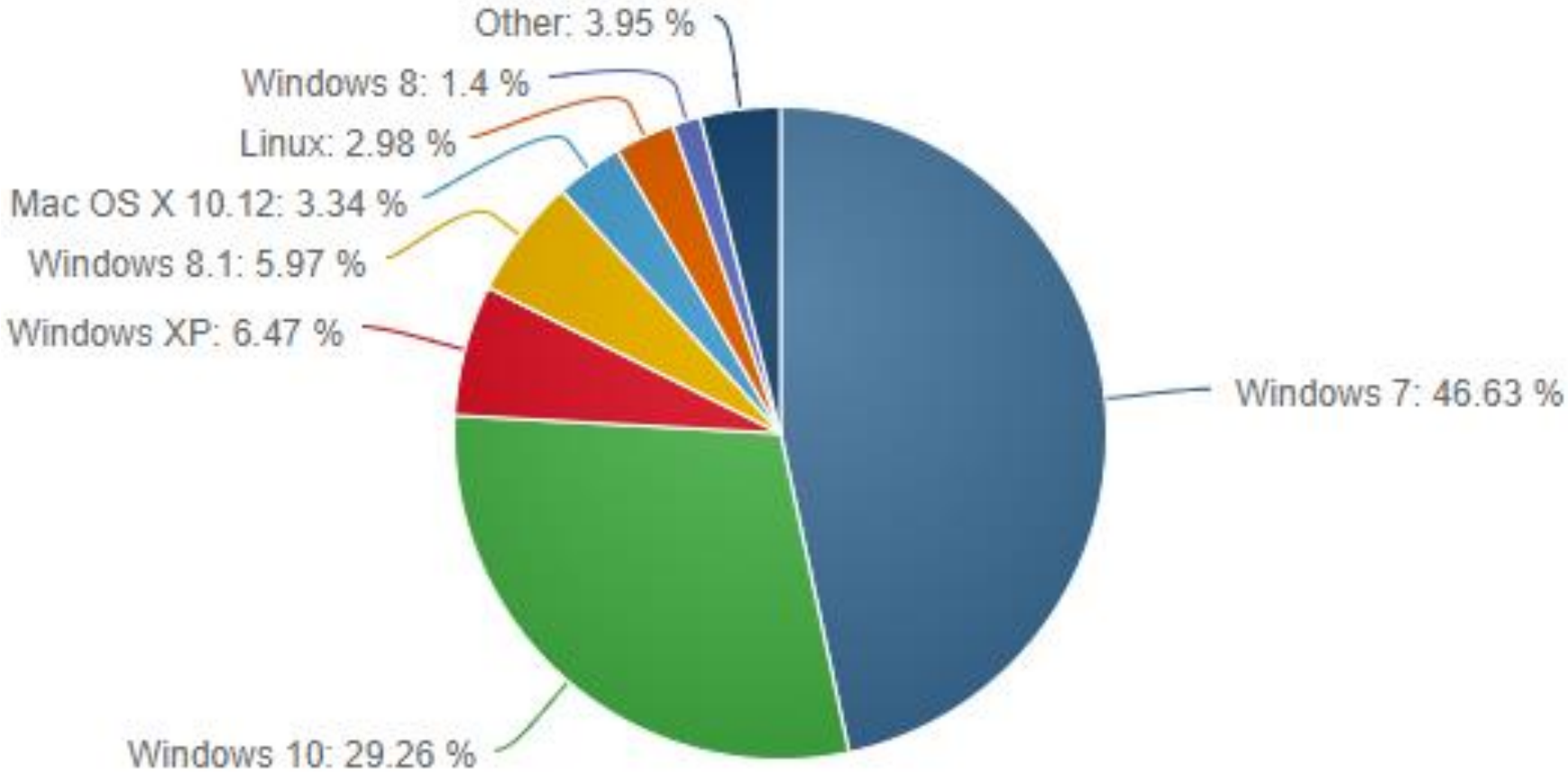
- Sequence analysis
  - Sequence searches: infer function of unknown DNA sequence
  - Comparative genomics: infer evolutionary trees from conserved proteins
  - Evolutionary biology: detect gene duplication / horizontal gene transfer
  - Mutational analysis: detect predisposition to diseases via SNP patterns
- Expression studies
  - Microarrays / RNA-seq: detect upregulated or downregulated genes
  - Protein mass spectrometry: deduce function, quantify expression
- Structural studies
  - Protein X-ray crystallography: calculate most likely structure of enzymes

Modern examples of bioinformatics

- Systems biology
  - “Interactome”: deduce key proteins from map of protein interactions
  - Pathway analysis: deduce presence/absence of conserved pathways in new genomes
- Image analysis
  - Track movement of cells, flies, fish, humans…
- Data mining
  - IBM’s “Watson” supercomputer chomps thru medical literature, helps provide diagnosis and detect whether drug combinations have bad side effects

Part I:
Why Linux?

Popular Operating Systems



Bioinformaticians are the 3%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Openness: anyone can read source code of bioinformatics software

GitHub, Inc. [US] https://github.com/pachterlab/kallisto/blob/master/src/main.cpp

```
  std::string ret(asctime(timeinfo));

  // chomp off the newline
  return ret.substr(0, ret.size() - 1);
}

int main(int argc, char *argv[]) {
  std::cout.sync_with_stdio(false);
  setvbuf(stdout, NULL, _IOFBF, 1048576);


  if (argc < 2) {
    usage();
    exit(1);
  } else {
    auto start_time(get_local_time());
    ProgramOptions opt;
```

Bioinformaticians are the 3%

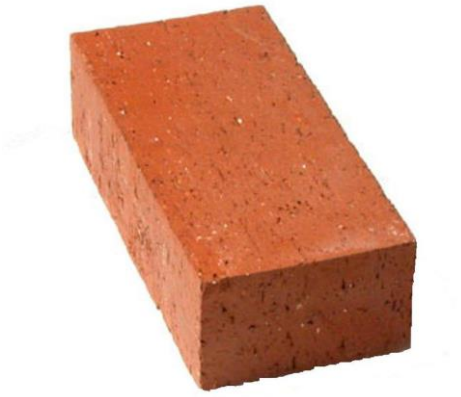
- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Server architecture: most large servers in universities run Linux

Top 10 positions of the 50th TOP500 in November 2017[15]

Rank	Rmax Rpeak (PFLOPS)	Name	Model	Processor	Interconnect	Vendor	Site country, year	Operating system
1	93.015 125.436	*Sunway TaihuLight*	Sunway MPP	SW26010	Sunway[16]	NRCPC	National Supercomputing Center in Wuxi China, 2016[16]	Linux (Raise)
2	33.863 54.902	*Tianhe-2*	TH-IVB-FEP	Xeon E5–2692, Xeon Phi 31S1P	TH Express-2	NUDT	National Supercomputing Center in Guangzhou China, 2013	Linux (Kylin)
3	19.590 25.326	*Piz Daint*	Cray XC50	Xeon E5-2690v3, Tesla P100	Aries	Cray	Swiss National Supercomputing Centre Switzerland, 2016	Linux (CLE)
4	19.136 28.192	*Gyoukou*	ZettaScaler-2.2 HPC system	Xeon D-1571, PEZY-SC2	Infiniband EDR	ExaScaler	Japan Agency for Marine-Earth Science and Technology Japan, 2017	Linux (CentOS)
5	17.590 27.113	*Titan*	Cray XK7	Opteron 6274, Tesla K20X	Gemini	Cray	Oak Ridge National Laboratory United States, 2012	Linux (CLE, SLES based)
6	17.173 20.133	*Sequoia*	Blue Gene/Q	A2	Custom	IBM	Lawrence Livermore National Laboratory United States, 2013	Linux (RHEL and CNK)
7	14.137 43.902	*Trinity*	Cray XC40	Xeon E5–2698v3, Xeon Phi	Aries	Cray	Los Alamos National Laboratory United States, 2015	Linux (CLE)
8	14.015 27.881	*Cori*	Cray XC40	Xeon Phi 7250	Aries	Cray	National Energy Research Scientific Computing Center United States, 2016	Linux (CLE)
9	13.555 24.914	*Oakforest-PACS*	Fujitsu	Xeon Phi 7250	Intel Omni-Path	Fujitsu	Kashiwa, Joint Center for Advanced High Performance Computing Japan, 2016	Linux
10	10.510 11.280	*K computer*	Fujitsu	SPARC64 VIIIfx	Tofu	Fujitsu	Riken, Advanced Institute for Computational Science (AICS) Japan, 2011	Linux

Bioinformaticians are the 3%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Ease at manipulating large files: nowadays, files > 1 GB are extremely common. Handling large files in Windows/OS X is extremely clunky!
  - Programming philosophy: **modular** vs **monolithic**



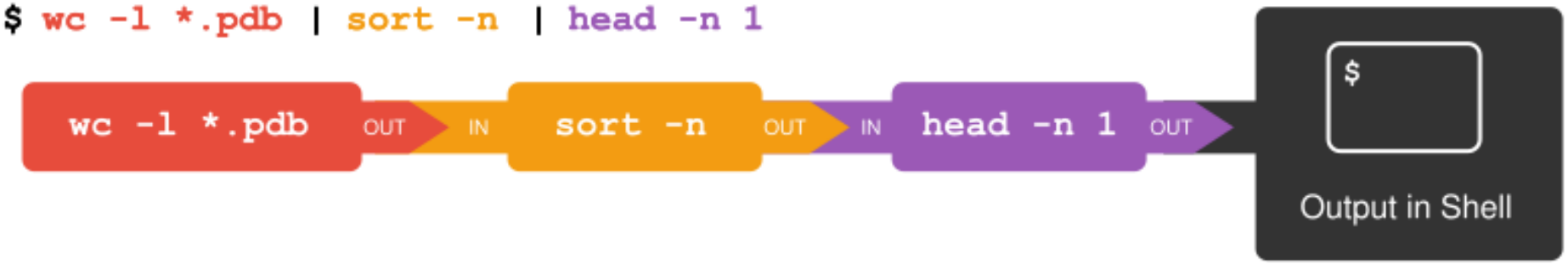
Linux

Windows/OS X

Bioinformaticians are the 3%

- Bioinformatics programs are **PREDOMINANTLY** written for Linux
- Why?
  - Text input/output, not silly proprietary filetypes (try opening a Word document in TextEdit/Notepad)
  - Piping: output of a tool can be “piped” as the input into another tool

```
$ wc -l *.pdb | sort -n | head -n 1
```

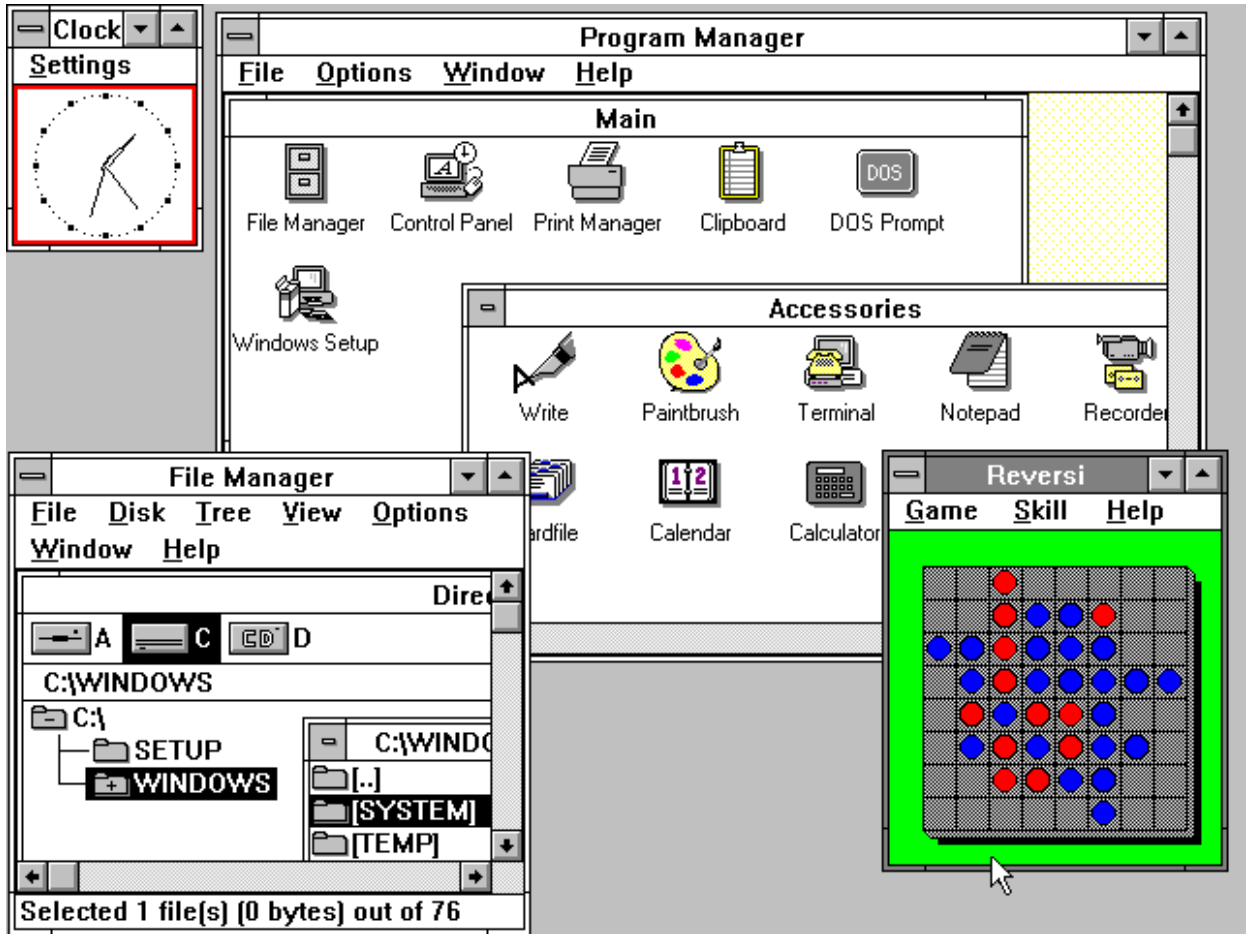


When bioinformaticians put several programs together to produce the desired output, they say they’ve built a “**pipeline**”

Part II:
Command-line-o-philia

Command-line vs. GUI

- GUIs used to be launched from the command-line
- Command-lines are now launched from GUIs



Terminal — ttyp1

```
Last login: Tue May  6 20:36:52 on ttyp1
Welcome to Darwin!
[thedj:~] pjotr%
```

chris@ubuntu: ~

```
chris@ubuntu:~$ ping google.com
PING google.com (216.58.216.142) 56(84) bytes of data.
64 bytes from sea15s01-in-f14.1e100.net (216.58.216.142): icmp_seq=1 ttl
=35.8 ms
64 bytes from sea15s01-in-f14.1e100.net (216.58.216.142): icmp_seq=2 ttl
=51.5 ms
^Z
[1]+  Stopped                 ping google.com
chris@ubuntu:~$ fg ping
ping google.com
64 bytes from sea15s01-in-f14.1e100.net (216.58.216.142): icmp_seq=3 ttl
=38.0 ms
64 bytes from sea15s01-in-f14.1e100.net (216.58.216.142): icmp_seq=4 ttl
=37.0 ms
```

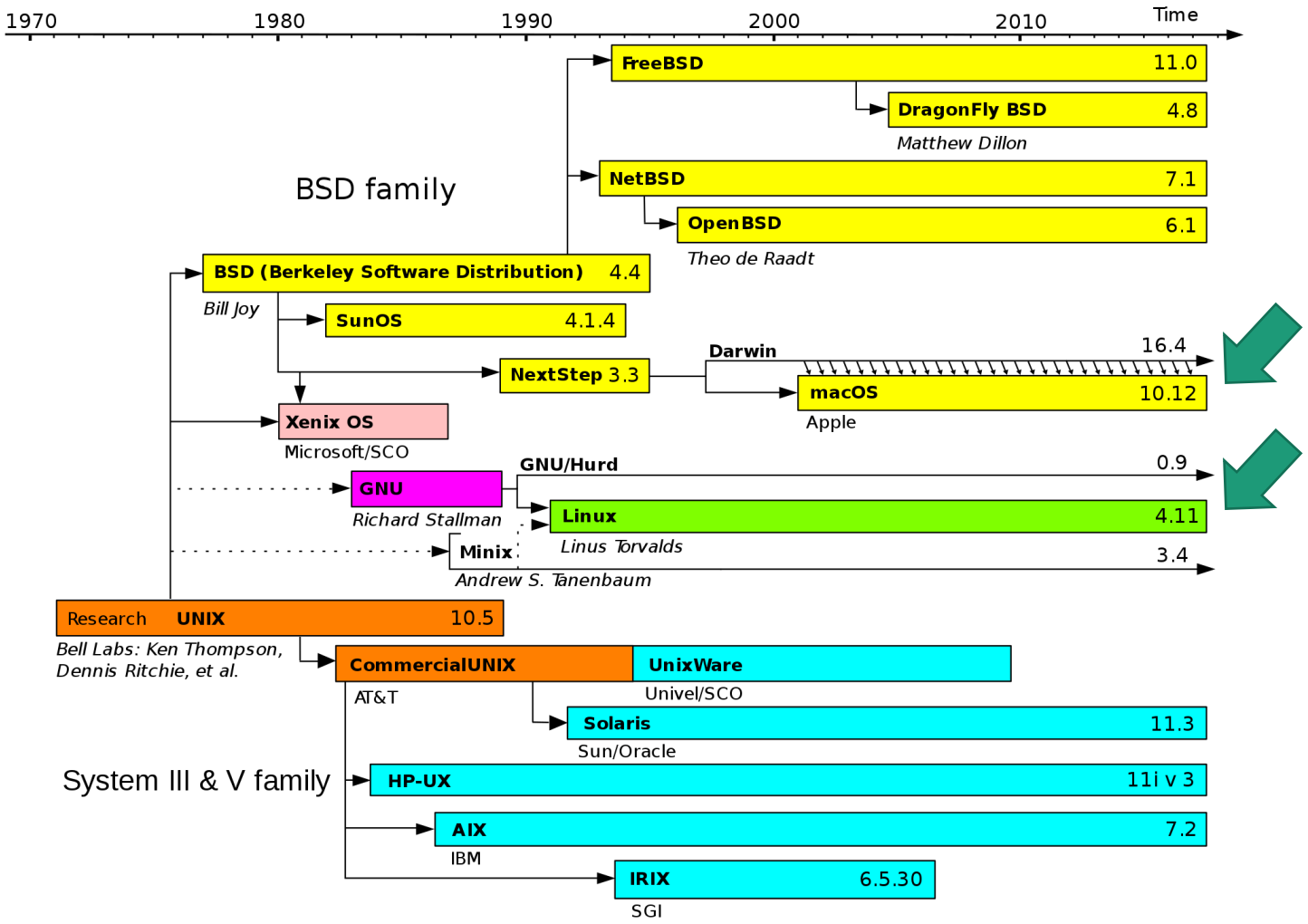
Windows PowerShell

```
PS C:\> Get-AzureRmADServicePrincipal -SearchString "AzureKeysRollerDaemon"

DisplayName                    Type                           ObjectId
-----------                    ----                           --------
AzureKeysRollerDaemon                                         672f1afa-526a-4ef6-819c-975c7cd79022


PS C:\> New-AzureRmRoleAssignment -ObjectId 672f1afa-526a-4ef6-819c-975c7cd79022 -RoleDefinitionName
 Contributor -Scope /subscriptions/c276fc76-9cd4-44c9-99a7-4fd71546436e


RoleAssignmentId   : /subscriptions/c276fc76-9cd4-44c9-99a7-4fd71546436e/providers/Microsoft.Author
                     ization/roleAssignments/9182ae32-4219-4c26-a80b-891ba649ae36
Scope              : /subscriptions/c276fc76-9cd4-44c9-99a7-4fd71546436e
DisplayName        : AzureKeysRollerDaemon
SignInName         :
RoleDefinitionName : Contributor
RoleDefinitionId   : b24988ac-6180-42a0-ab88-20f7382dd24c
ObjectId           : 672f1afa-526a-4ef6-819c-975c7cd79022
ObjectType         : ServicePrincipal



PS C:\>
```

Command Prompt

```
 Directory of C:\Users\Jon\AppData\Local\BrawlBox

04/28/2016  11:20 AM    <DIR>          .
04/28/2016  11:20 AM    <DIR>          ..
04/28/2016  11:20 AM    <DIR>          BrawlBox.exe_Url_nlhqq3eglxy4so4t2xnh5x0n
o5j5q4iv
               0 File(s)              0 bytes

 Directory of C:\Users\Jon\AppData\Local\BrawlBox\BrawlBox.exe_Url_nlhqq3eglxy4s
o4t2xnh5x0no5j5q4iv

04/28/2016  11:20 AM    <DIR>          .
04/28/2016  11:20 AM    <DIR>          ..
04/28/2016  11:20 AM    <DIR>          0.71.5111.26120
               0 File(s)              0 bytes

 Directory of C:\Users\Jon\AppData\Local\BrawlBox\BrawlBox.exe_Url_nlhqq3eglxy4s
o4t2xnh5x0no5j5q4iv\0.71.5111.26120

04/28/2016  11:20 AM    <DIR>          .
04/28/2016  11:20 AM    <DIR>          ..
04/28/2016  11:20 AM             1,111 user.config
               1 File(s)          1,111 bytes

 Directory of C:\Users\Jon\AppData\Local\Broadcom

04/18/2015  09:19 AM    <DIR>          .
04/18/2015  09:19 AM    <DIR>          ..
04/18/2015  09:19 AM    <DIR>          Bluetooth Software
               0 File(s)              0 bytes

 Directory of C:\Users\Jon\AppData\Local\Broadcom\Bluetooth Software

04/18/2015  09:19 AM    <DIR>          .
04/18/2015  09:19 AM    <DIR>          ..
04/18/2015  09:19 AM    <DIR>          sync
               0 File(s)              0 bytes

 Directory of C:\Users\Jon\AppData\Local\Broadcom\Bluetooth Software\sync
```

Command-lines are different!



- OS X and Linux have similar command lines
- Windows command lines are very different
  - **cmd**: more UNIX-like
  - **PowerShell**: manipulates objects instead of text

Mini-demo

- Commands covered:

Command	Function
ls	List files
cd	Change directory
cat	Concatenate files, print to standard output
less	View contents of a file
nano	Edit contents of a file

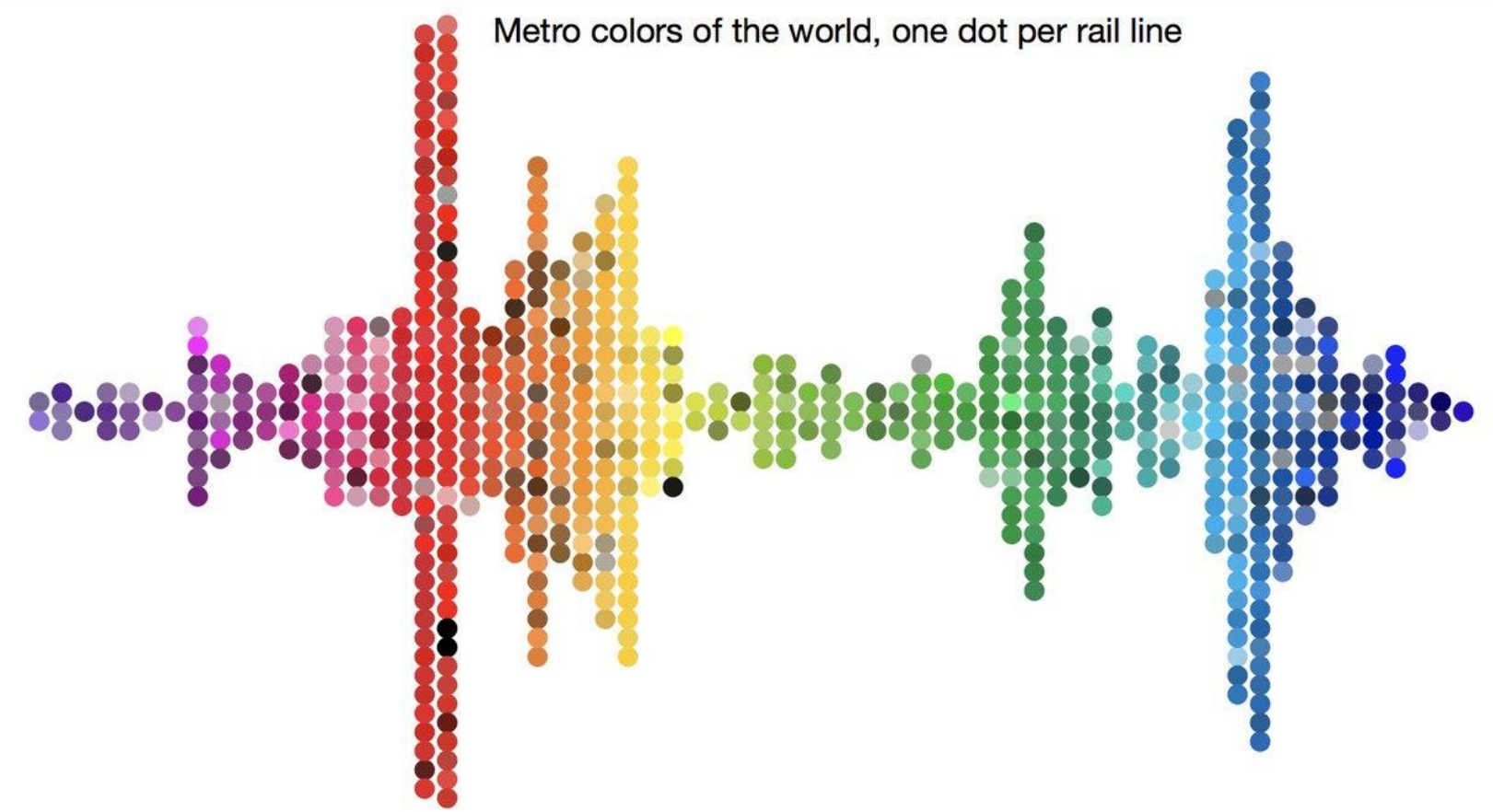
Intermission I: Questions?

CGP Grey - How Machines Learn
https://www.youtube.com/watch?v=R9OHn5ZF4Uo

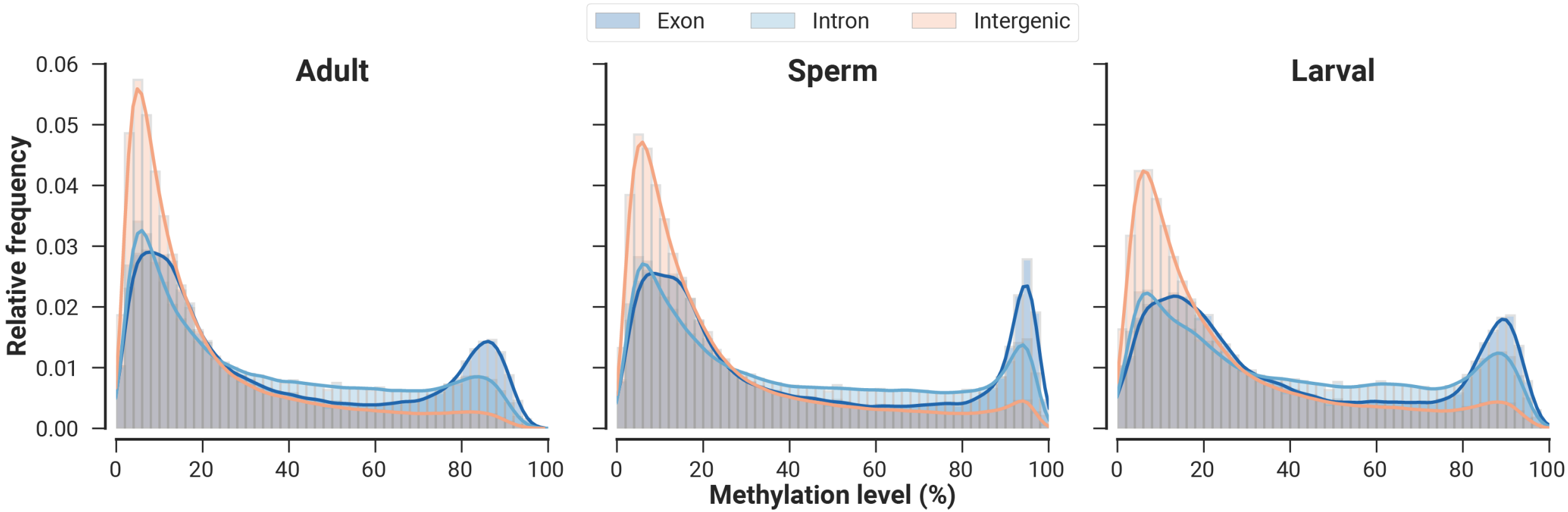
Part III:
Statistics in transcriptomics

Statistics: why learn it?

- Allows one to **quantify** whether data is interesting!



Understand your data!



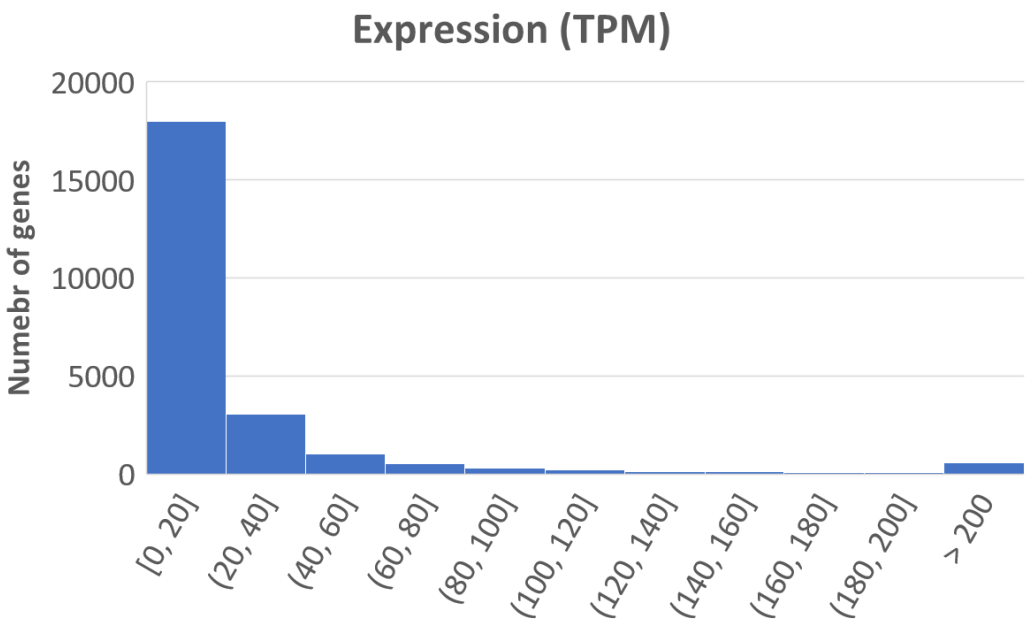
Is this **parametric**, or **non-parametric**?

Why use normal data?

- Allows the use of parametric tests:
  - Student's *t*-test
  - ANOVAs
  - Pearson correlation (r^2)
  - z-scores
- While not compulsory,
  - PCAs are more meaningful
  - Prettier graphs
- Much much easier to imagine normally-distributed data!

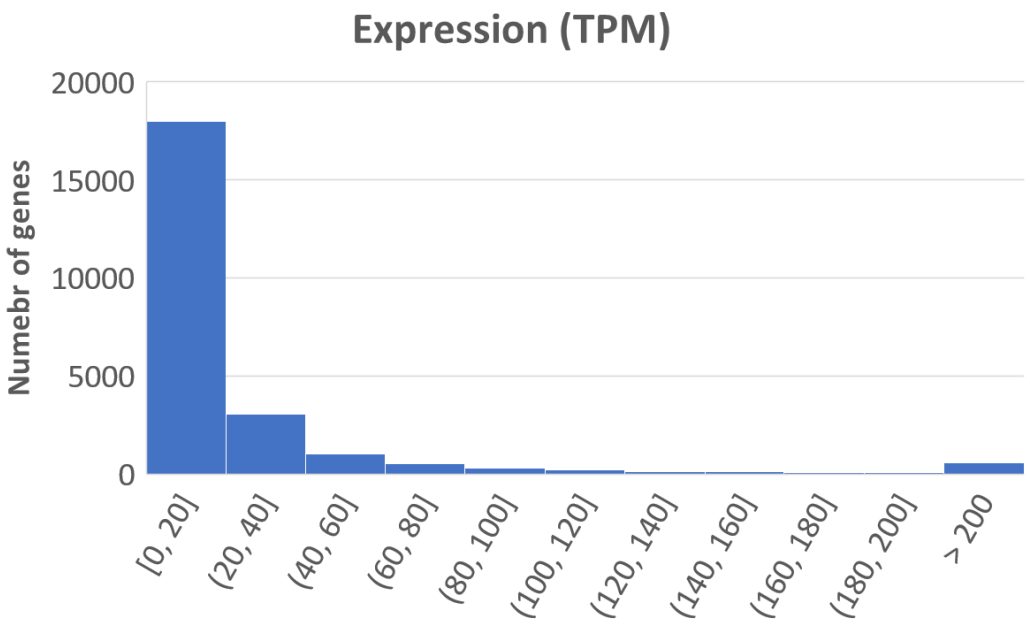
Do I need to transform my data?

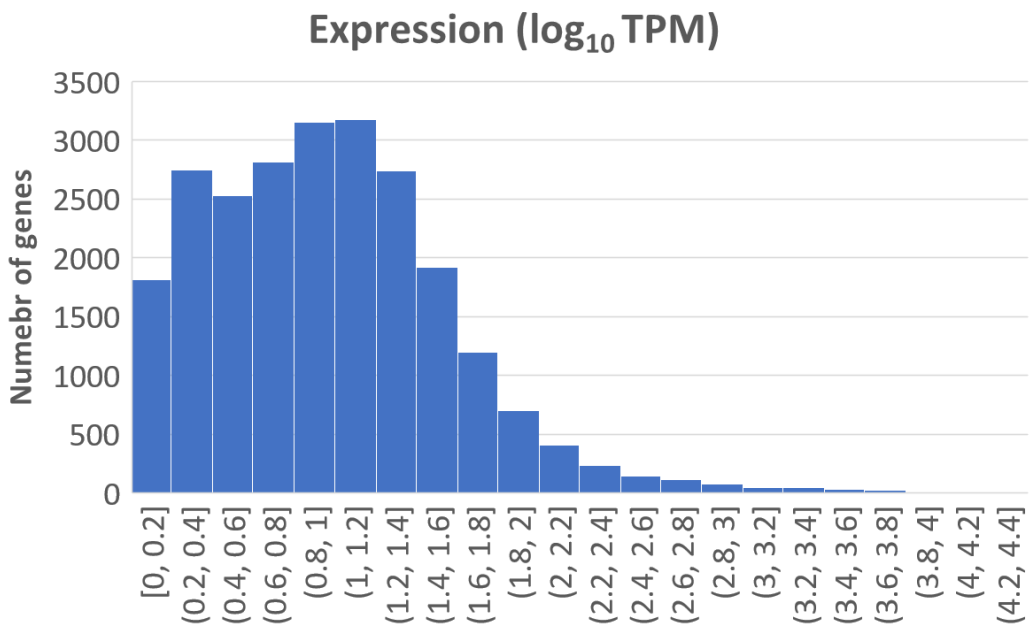
- Especially for expression data, raw values tend to not be normal



Do I need to transform my data?

- After log-transformation:





- This implies original data was log-normal!

“Differential”

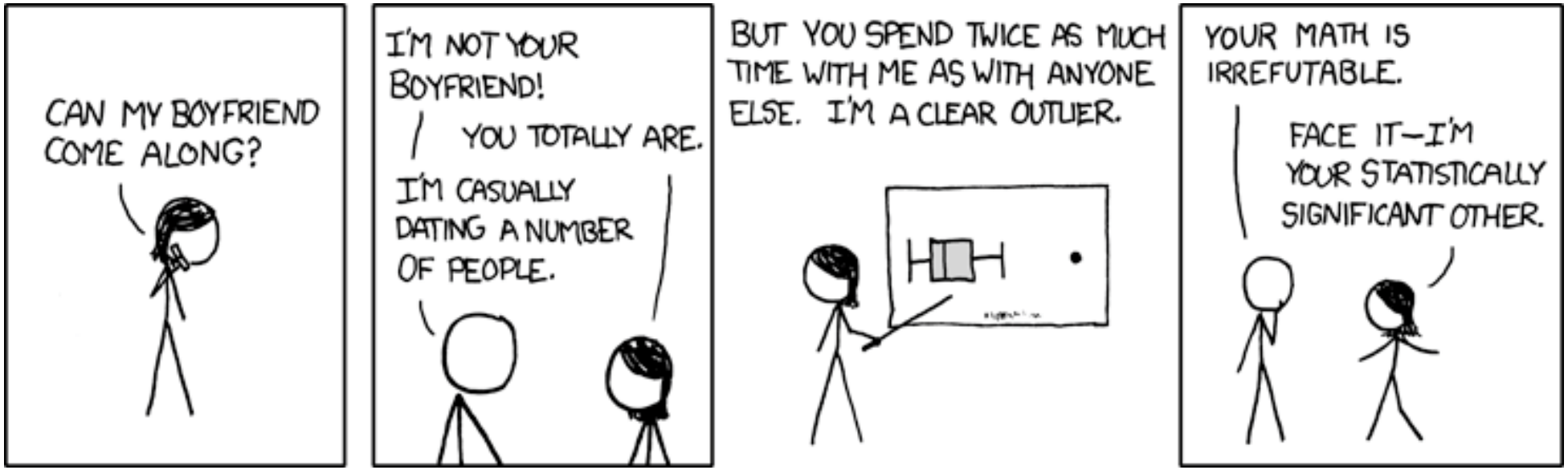
- If a gene is said to be “**differentially** expressed under **heat stress** relative to **control**”, what does it mean? What about **numerically**?
  - Also think about the corollary: how do you know the gene is **NOT** differentially expressed?

Short-Term Acute Exposure Of Healthy Humans To Particulate Matter Induces **Differential Gene Expression** In Lung Immune Cells
AY Meliton, T Cho, RB Hamanaka… - C103. OUTDOOR AIR …, 2017 - Am Thoracic Soc
Rationale: mortality. Current levels of PM in American cities are estimated to be responsible for 50,000 to 60,000 excess deaths per year in the US; however, the mechanisms by which PM causes adverse health effects are not completely understood. Experimental exposure of

[HTML] **Differential gene expression** in brain and peripheral tissues in depression across the life span: A review of replicated findings
LG Ciobanu, PS Sachdev, JN Trollor… - Neuroscience & …, 2016 - Elsevier
Abstract There is a growing body of research investigating the **gene expression** signature of depression at the genome-wide level, with potential for the discovery of novel pathophysiological mechanisms of depression. However, heterogeneity of depression,
Cited by 2 Related articles All 7 versions Web of Science: 2

“significant”

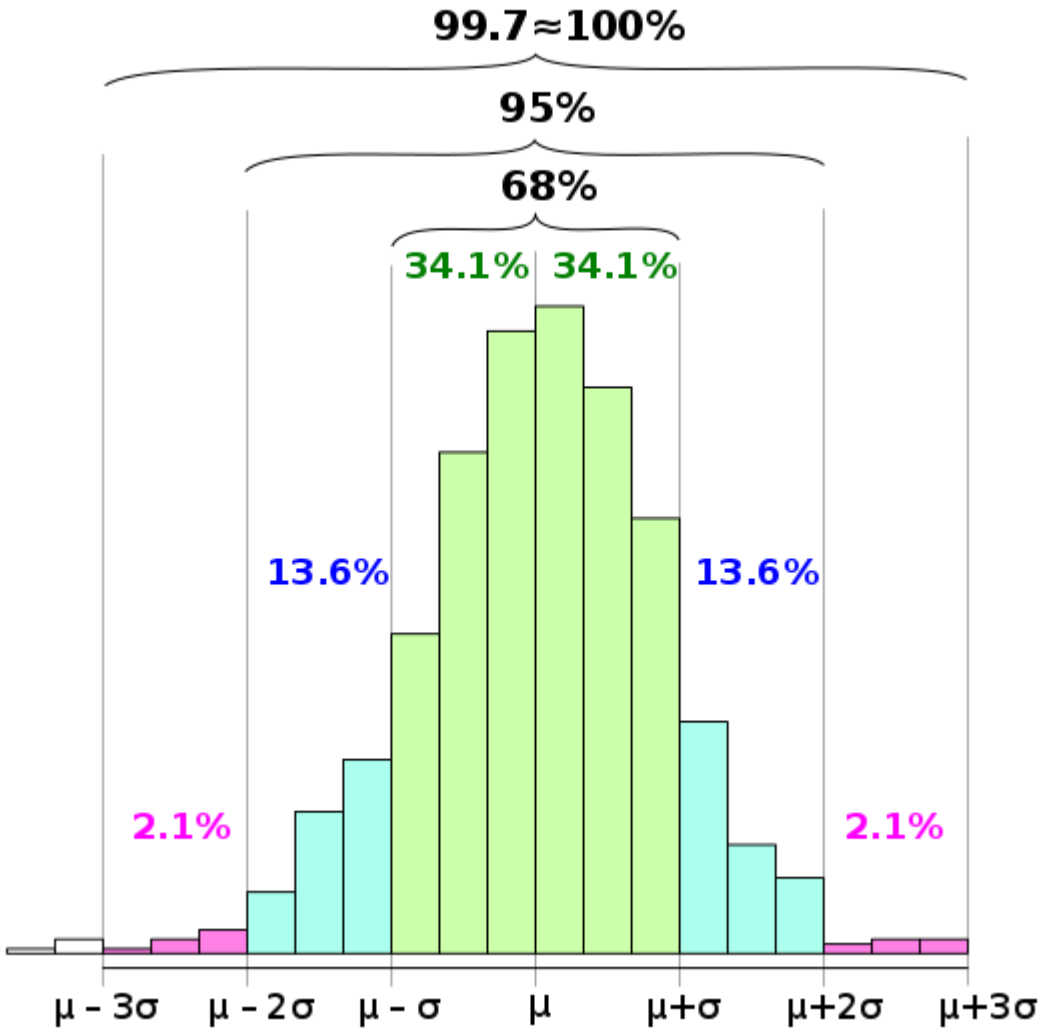
- “p < 0.05”: what do the *p* values **mean**?
  - Can you think of situations where heat stress values are very different from control, yet not considered statistically significant?



68-95-99 rule

- $p < 0.05$ implies data point is 2 SD outside of the mean

- e.g. IQ scores: 100 ± 15 (s.d.)
  - (at what IQ is one significantly smarter than the general pop?)



Student's *t*-test (two-sample, two-tailed)

- Tests whether observations have a **significantly different** mean from an expected baseline
  - e.g. IQ of sampled KAUSTians (n = 10) are 120 ± 10 (s.d.)
  - Is our IQ significantly different to that of a random sample (n = 10) from the population?

QuickCalcs

1. Select category 2. Choose calculator 3. Enter data **4. View results**

Unpaired *t* test results

P value and statistical significance:
The two-tailed P value equals 0.0025
By conventional criteria, this difference is considered to be very statistically significant.

Confidence interval:
The mean of General pop minus KAUSTians equals -20.00
95% confidence interval of this difference: From -31.98 to -8.02

Intermediate values used in calculations:
t = 3.5082
df = 18
standard error of difference = 5.701

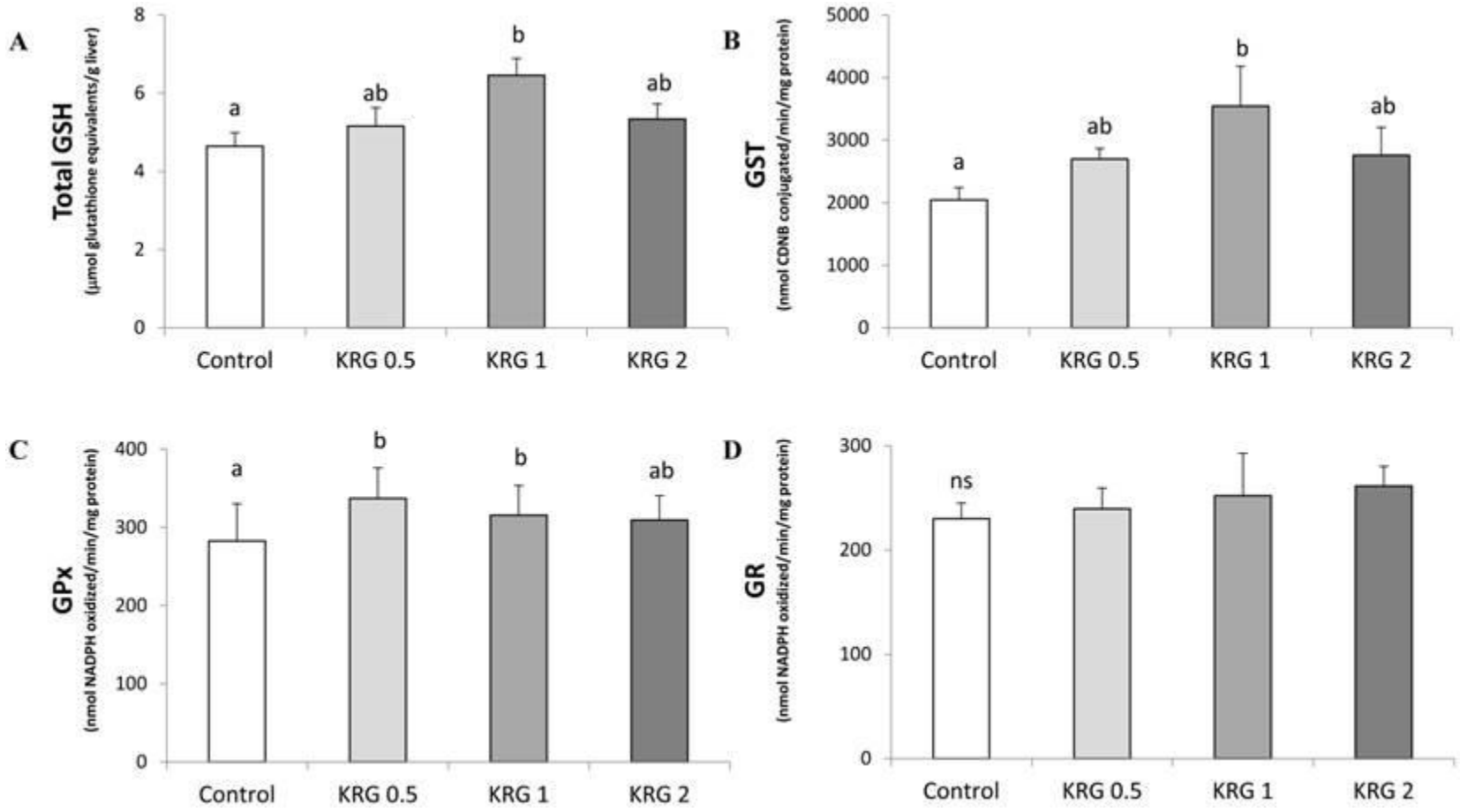
Learn more:
GraphPad's web site includes portions of the manual for GraphPad Prism that can help you learn statistics. First, review the meaning of P values and confidence intervals . Then learn how to interpret results from an unpaired or paired *t* test. These links include GraphPad's popular *analysis checklists* .

Review your data:

Group	General pop	KAUSTians
Mean	100.00	120.00
SD	15.00	10.00
SEM	4.74	3.16
N	10	10

Student's *t*-test

- Very commonly done in scientific papers!



Student's *t*-test

- Alternative tests with the same idea
  - Testing something, but with preconceived idea of which direction the effect should be in? One-tailed *t*-test
  - Testing something, but against a fixed value instead of another distribution? One-sample *t*-test
  - More than two groups? ANOVA
  - Non-parametric data? Mann-Whitney U

Fisher’s exact test

- Is this observation **statistically significant**?

Fisher’s exact test

- … depends on which bag of M&Ms you used!

$p < 0.05$

$p > 0.05$

Fisher's exact test

- For the heck of it, let's calculate some *p* values:
  - M&M website says that pack has ~50 candies
  - There are 5 colours, R G Y B O
  - We expect 10 green per packet
  - Cookie has 10 green M&Ms
- Set up Fisher's exact table

	Green M&M	Non-green M&M
Cookie	10	0
Not-on-cookie	0	40

Fisher’s exact test

https://www.graphpad.com/quickcalcs/contingency1/

GraphPad Software

Scientific Software | Data Analysi

QuickCalcs

1. Select category | 2. Choose calculator | **3. Enter data** | 4. View results

Analyze a 2x2 contingency table

Enter your data

Enter the number of subjects actually observed. Don't enter proportions, percentages or means.
Learn how to create a contingency table.

	green	non-green
cookie	10	0
not-on-cookie	0	40

https://www.graphpad.com/quickcalcs/contingency2/

GraphPad Software

Scientific Software | Data Analysi

QuickCalcs

1. Select category | 2. Choose calculator | 3. Enter data | **4. View results**

Analyze a 2x2 contingency table

	green	non-green	Total
cookie	10	0	10
not-on-cookie	0	40	40
Total	10	40	50

Fisher's exact test

The two-tailed P value is less than 0.0001
The association between rows (groups) and columns (outcomes) is considered to be extremely statistically significant.

Learn how to interpret the P value.

Fisher’s exact test

- If we want precise *p* values, we can use R:

R Console

```
> fishers_matrix <- matrix(c(10,0,0,40), nrow=2)
> fishers_matrix
     [,1] [,2]
[1,]   10    0
[2,]    0   40
> fisher.test(fishers_matrix)

        Fisher's Exact Test for Count Data

data:  fishers_matrix
p-value = 9.735e-11
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 47.75873      Inf
sample estimates:
odds ratio
       Inf

>
```

Fisher's exact test

- Let's look at a more biological example
  - Perform a heat stress experiment
  - Obtain genes that were upregulated under stress
  - Check GO terms associated with these genes

	Genes with GO:0016209	Genes without GO:0016209
Upregulated	10	0
Not upregulated	0	40

  - p = ?

… GO:0016209?

http://amigo.geneontology.org/amigo/term/GO:0016209

AmiGO 2 | Home | Search | Browse | Tools & Resources | Help | Feedback | About | AmiGO 1.8 | Quick search | Search

antioxidant activity

Term Information

Data health

Accession GO:0016209
Name antioxidant activity
Ontology molecular_function
Synonyms None
Alternate IDs None
Definition Inhibition of the reactions brought about by dioxygen (O2) or peroxides. Usually the antioxidant is effective because it can itself be more easily oxidized than the substance protected. The term is often applied to components that can trap free radicals, thereby breaking the chain reaction that normally leads to extensive biological damage. *Source:* ISBN:0198506732
Comment None
History See term history for GO:0016209 at QuickGO
Subset goslim_metagenomics
gosubset_prok
goslim_pir
Related Link to all **genes and gene products** annotated to antioxidant activity.
Link to all direct and indirect **annotations** to antioxidant activity.
Link to all direct and indirect **annotations download** (limited to first 10,000) for antioxidant activity.

… GO:0016209?

http://amigo.geneontology.org/amigo/term/GO:0016209

AmiGO 2 Home Search

BIOLOGIA PLANTARUM 43 (2): 245-251, 2000

Increased antioxidant activity under elevated temperatures: a mechanism of heat stress tolerance in wheat genotypes

R.K. SAIRAM, G.C. SRIVASTAVA and D.C. SAXENA

Division of Plant Physiology, Indian Agricultural Research Institute, New Delhi - 110012, India

…y oxidized than the substance protected. The …ive biological damage. *Source:* ISBN:0198506732

… antioxidant activity.

…ions to antioxidant activity.

…direct **annotations download** (limited to first 10,000) for antioxidant activity.

Fisher’s exact test

- In academic-ese, what I showed is termed **“functional enrichment”**
- Also known as **“GO term enrichment analysis”**

The effect of mechanical **stress** on the proliferation, adipogenic differentiation and gene expression of human adipose-derived stem cells

NE Paul, B Denecke, BS Kim, A Dreser… - Journal of tissue …, 2017 - Wiley Online Library

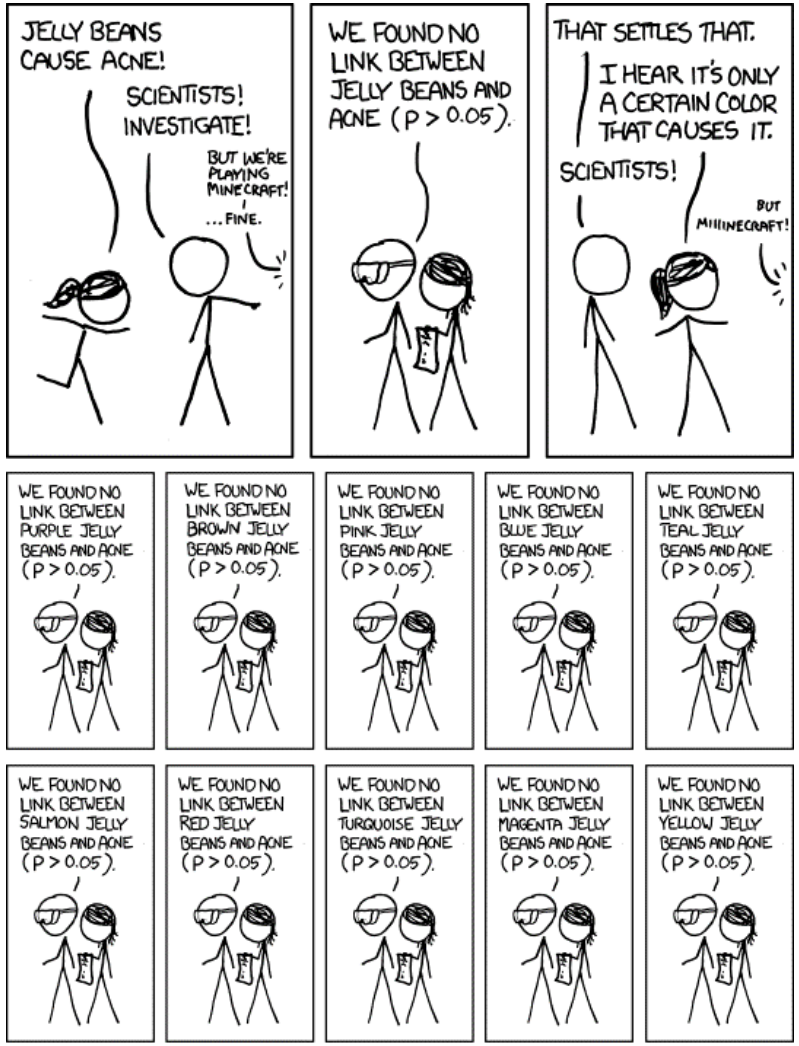
… These data suggest that the impact of mechanical **stress** on gene expression persisted over the treatment period of 10 days. **Functional enrichment** analyses were performed with genes that are at least 1.5-fold regulated (adjusted $p < 0.01$) by stretching compared with a static …

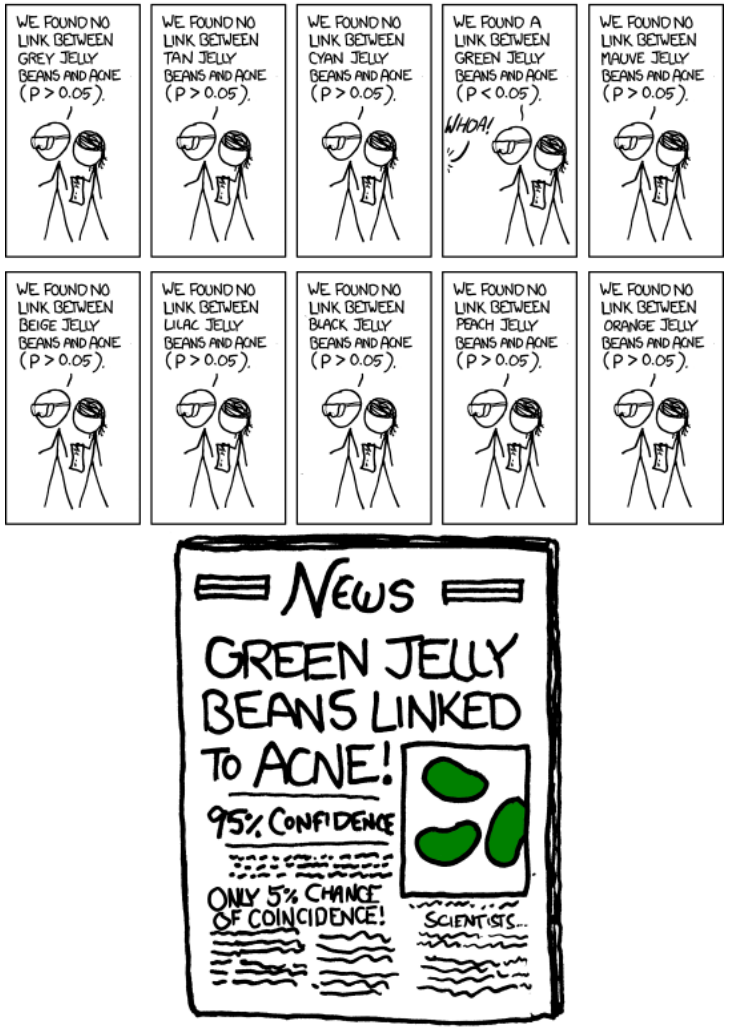
☆ ❝❝ Cited by 5 All 3 versions

Fisher’s exact test

- Back in plain English: Fisher’s exact tests calculates the probability of **observing** something, given that the overall distribution (i.e. universe) has a **quantifiable** distribution
- Alternative tests with the same idea
  - More than 2x2 table? Chi-squared
  - 2x2, but values are absurdly large? Chi-squared
  - Proportions of universe known, but not exact values? Binomial test

Multiple testing correction





Multiple testing correction

- Benjamini-Hochberg (1995)
  - How to correct *p* value
- Benjamini-Yekutieli (2001)
  - How to make sure corrected list have the same ranking as the uncorrected one

[PDF] Controlling the false discovery rate: a practical and powerful approach to multiple testing
Y Benjamini, Y Hochberg - Journal of the royal statistical society. Series B ..., 1995 - JSTOR
The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected
Cited by 44085 Related articles All 50 versions Web of Science: 30104

More powerful procedures for multiple significance testing
Y Hochberg, Y Benjamini - Statistics in medicine, 1990 - Wiley Online Library
Abstract The problem of multiple comparisons is discussed in the context of medical research. The need for more powerful procedures than classical multiple comparison procedures is indicated. To this end some new, general and simple procedures are
Cited by 1682 Related articles All 4 versions Web of Science: 1267

[PDF] The control of the false discovery rate in multiple testing under dependency
Y Benjamini, D Yekutieli - Annals of statistics, 2001 - JSTOR
Benjamini and Hochberg suggest that the false discovery rate may be the appropriate error rate to control in many applied multiple testing problems. A simple procedure was given there as an FDR controlling procedure for independent test statistics and was shown to be
Cited by 5452 Related articles All 20 versions Web of Science: 3466

Multiple testing correction

- **R:** p.adjust()
- Manually:

Gene	Uncorrected *p*	Correction factor	Corrected *p*
F	**0.0001**	x 6 / 1	**0.0006**
E	**0.001**	x 6 / 2	**0.003**
C	**0.01**	x 6 / 3	**0.02**
A	**0.04**	x 6 / 4	0.06
B	0.08	x 6 / 5	0.096
D	0.2	x 6 / 6	0.2

End of morning session: Questions?