Best pratice in applied machine learning

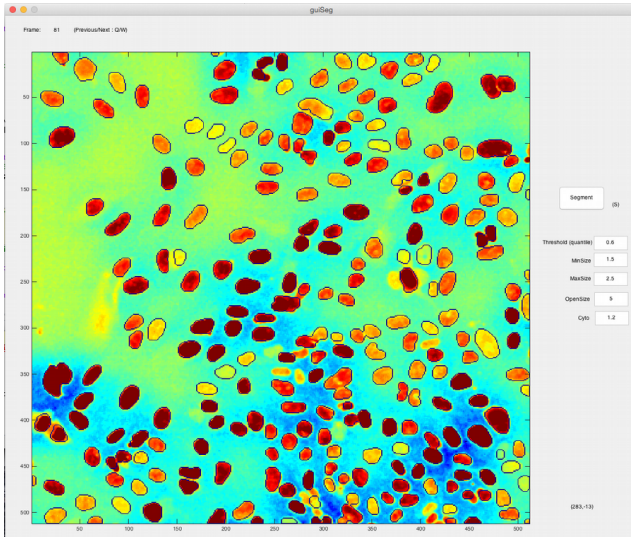
Eric Paquet
Computational Systems Biology
EPFL
November 21th, 2017

Who am I?

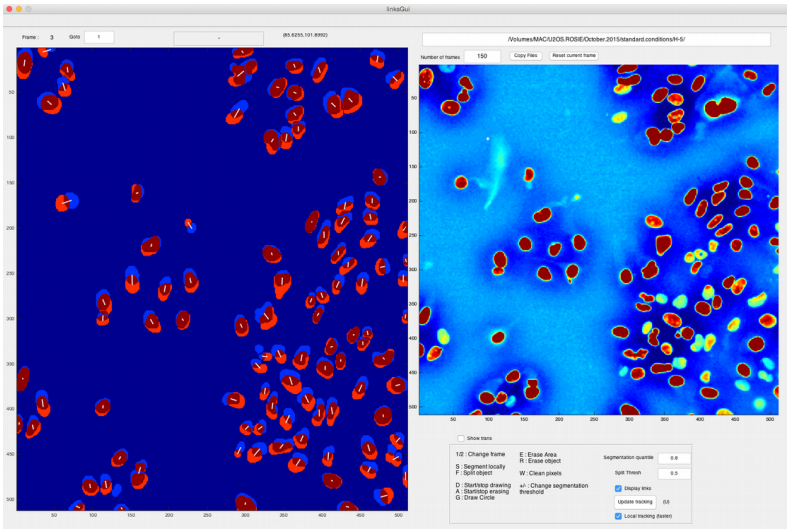
- Post-Doc at EPFL in the Computational Systems Biology group (Naef's lab)
  - Currently involve in projects tracking individual cells over long period of time using live-cell imaging data to study protein dynamics

Our tracking pipeline

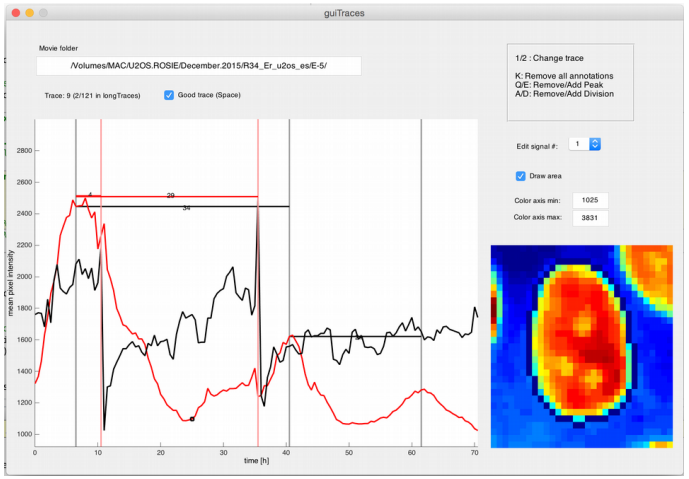
1- Segmentation



2- Tracking



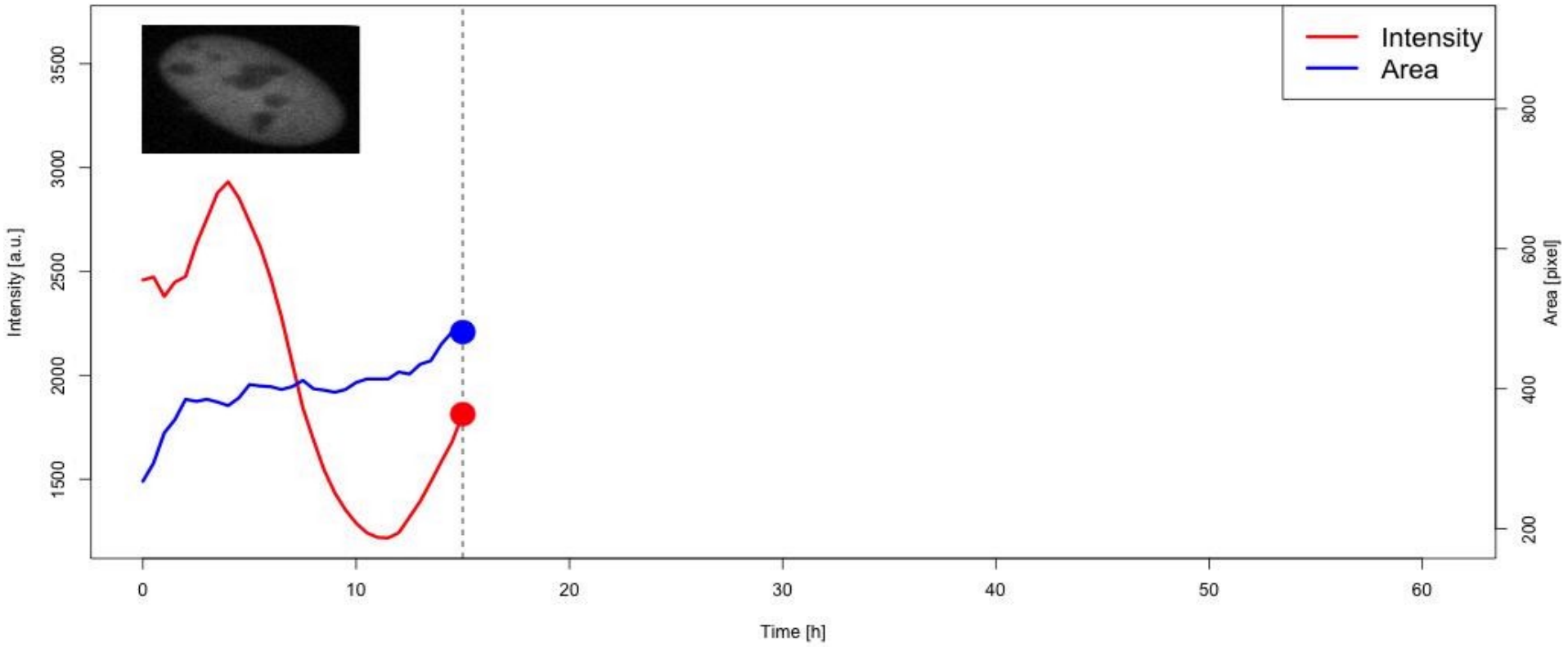
3- QC (traces)



<u>Strengths :</u>

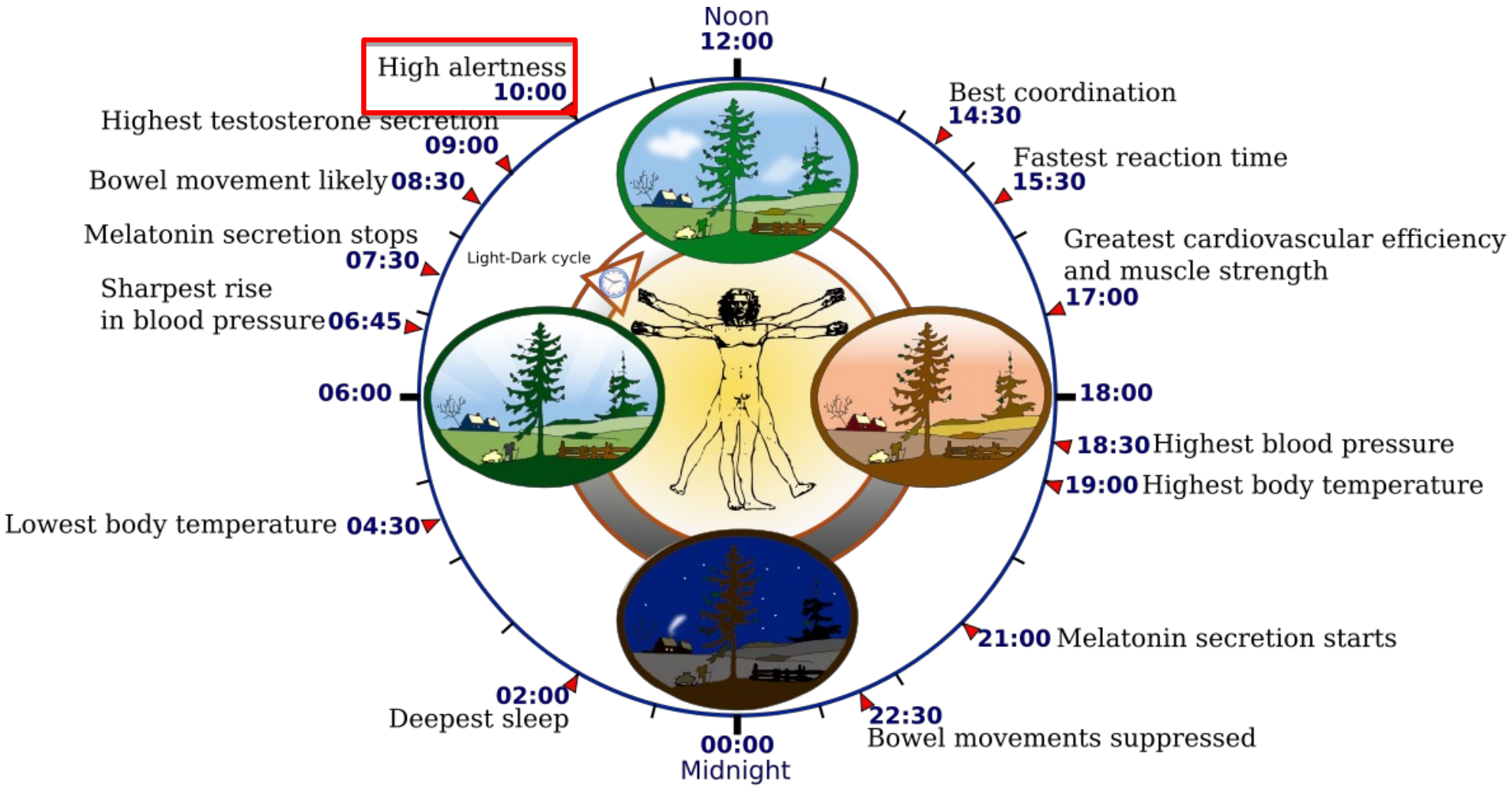
- Matlab suite of highly-customizable GUIs
  - Segment
  - Track
  - Quality control
- ~20 72h high-quality traces per field of view (20X)

Example of a trace from one cell



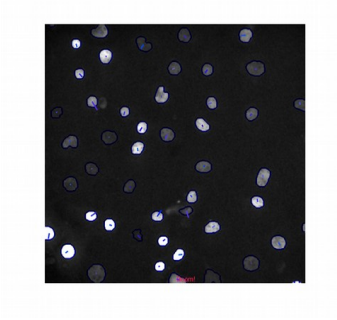
Understanding the interaction between the circadian clock and the cell cycle

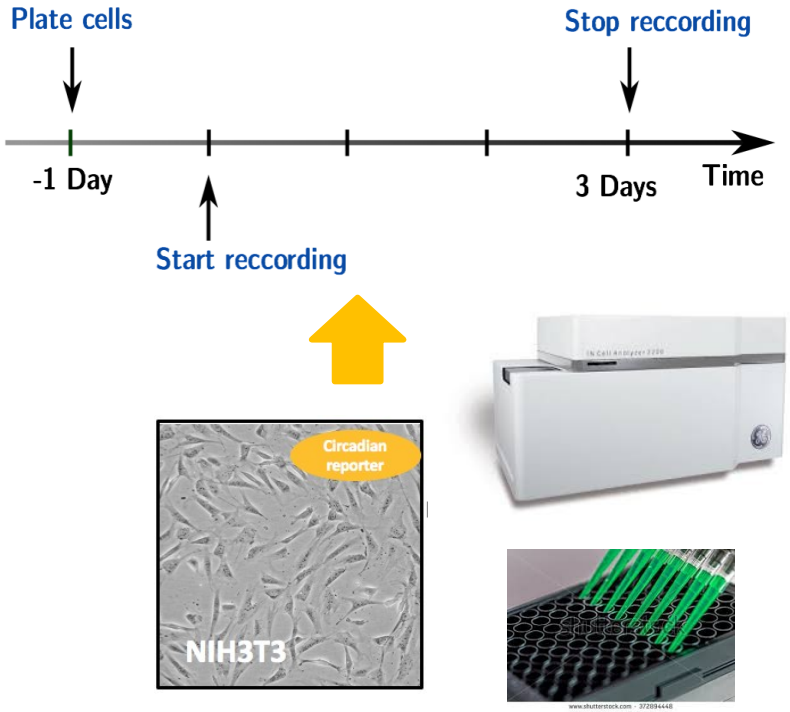
Examples of processes driven by the circadian clock

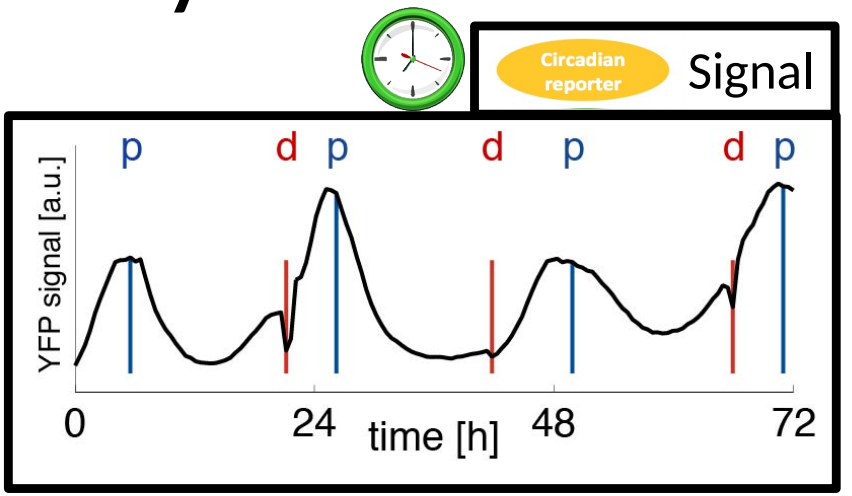


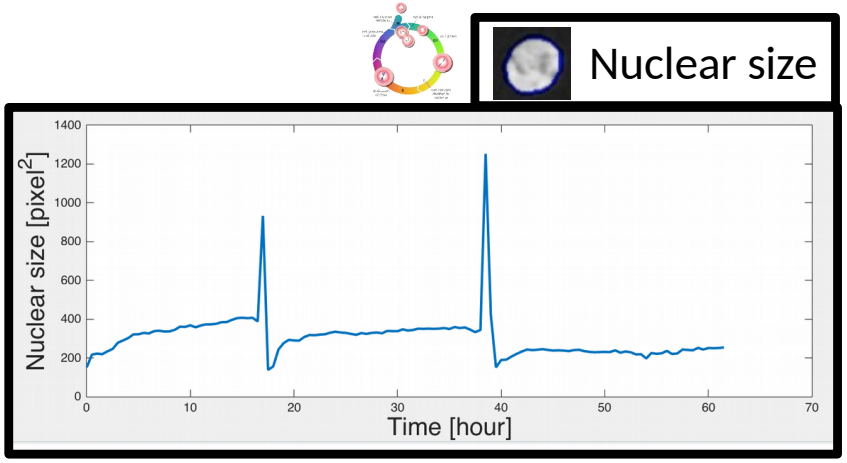
The Body Clock Guide to Better Health, Lamberg, and Smolensky, 2001. Wikipedia

How do we simultaneously track the cell and circadian cycles?

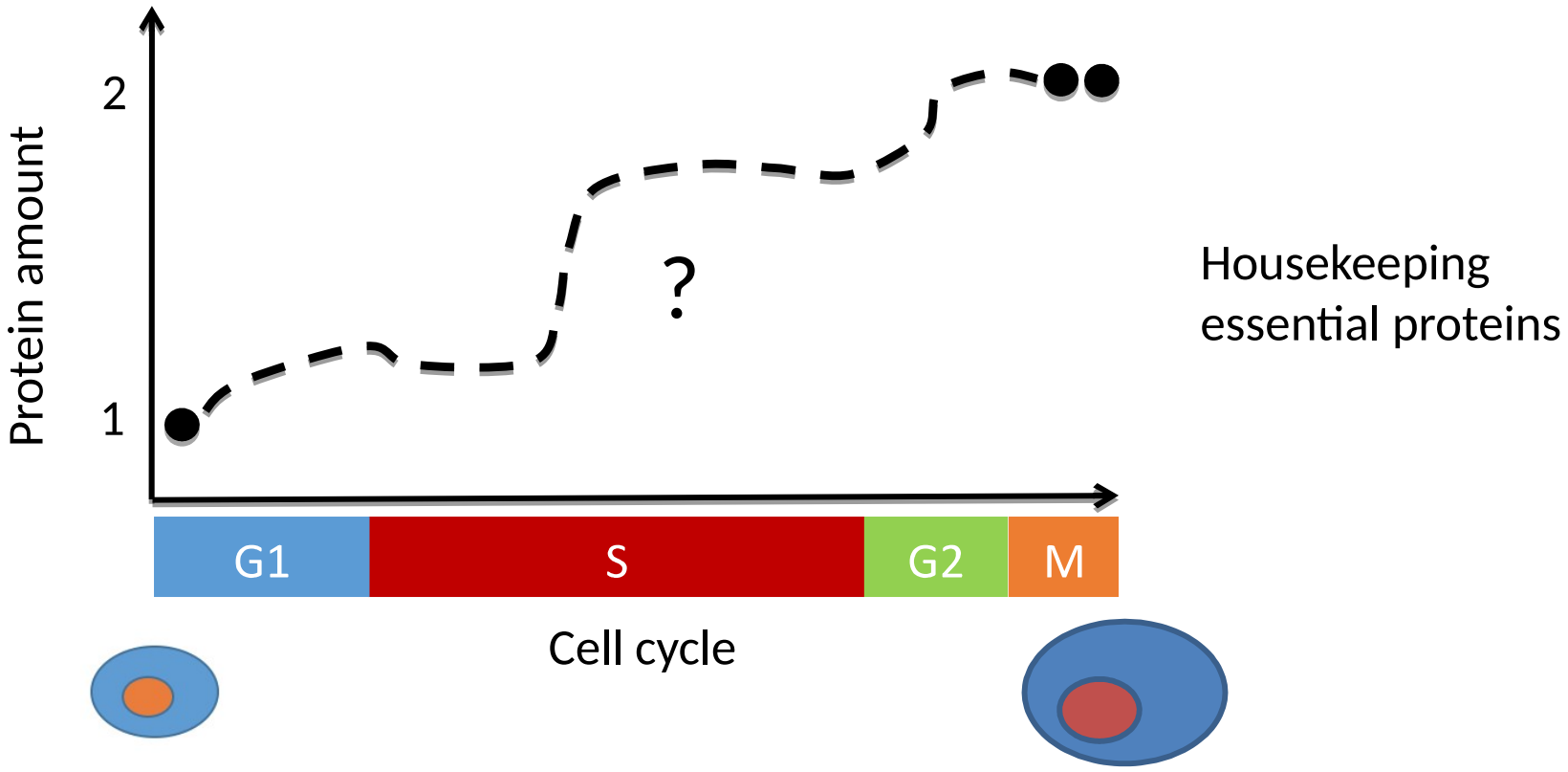








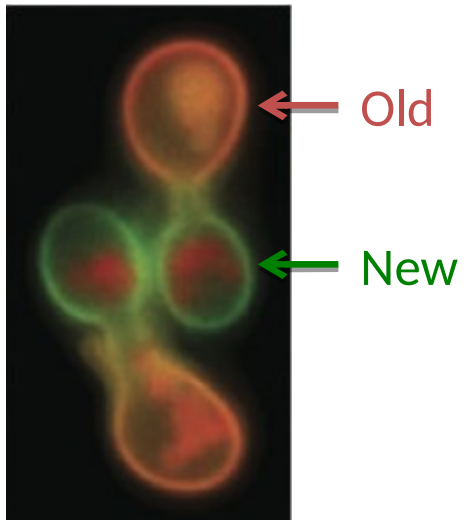
Protein dynamics around the cell cycle



How are we studying this ?

- Single cell level
- Using live cell imaging to get synthesis and degradation rates
- No need for synchronization and perturbation.
- Dual fluorescent timer :

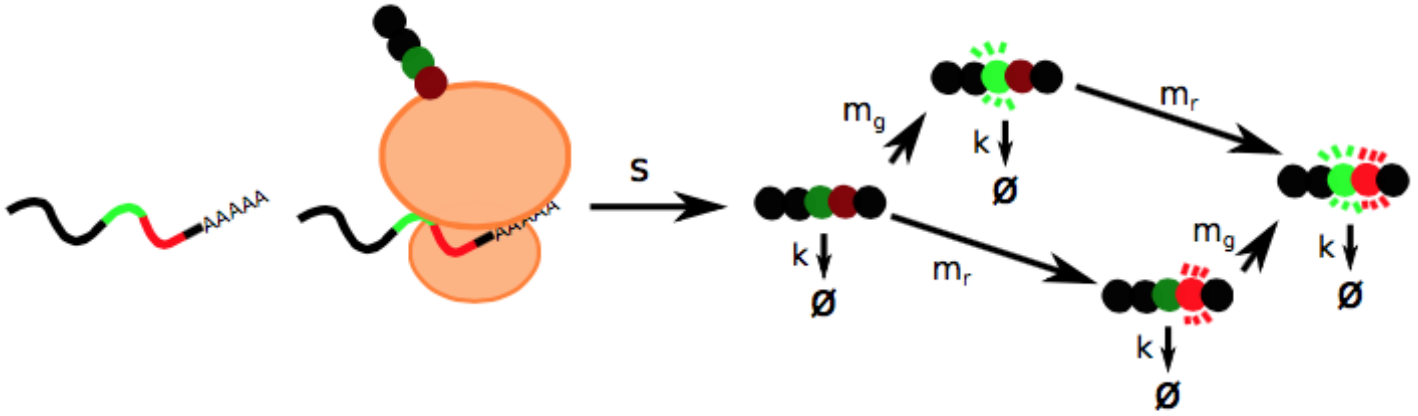
Maturation time : m_g = 5 minutes m_r = 8.5 hours



Hxt1 in yeast

Nature Biotechnology 30, 708–714 (2012)

Modeling the dual timer



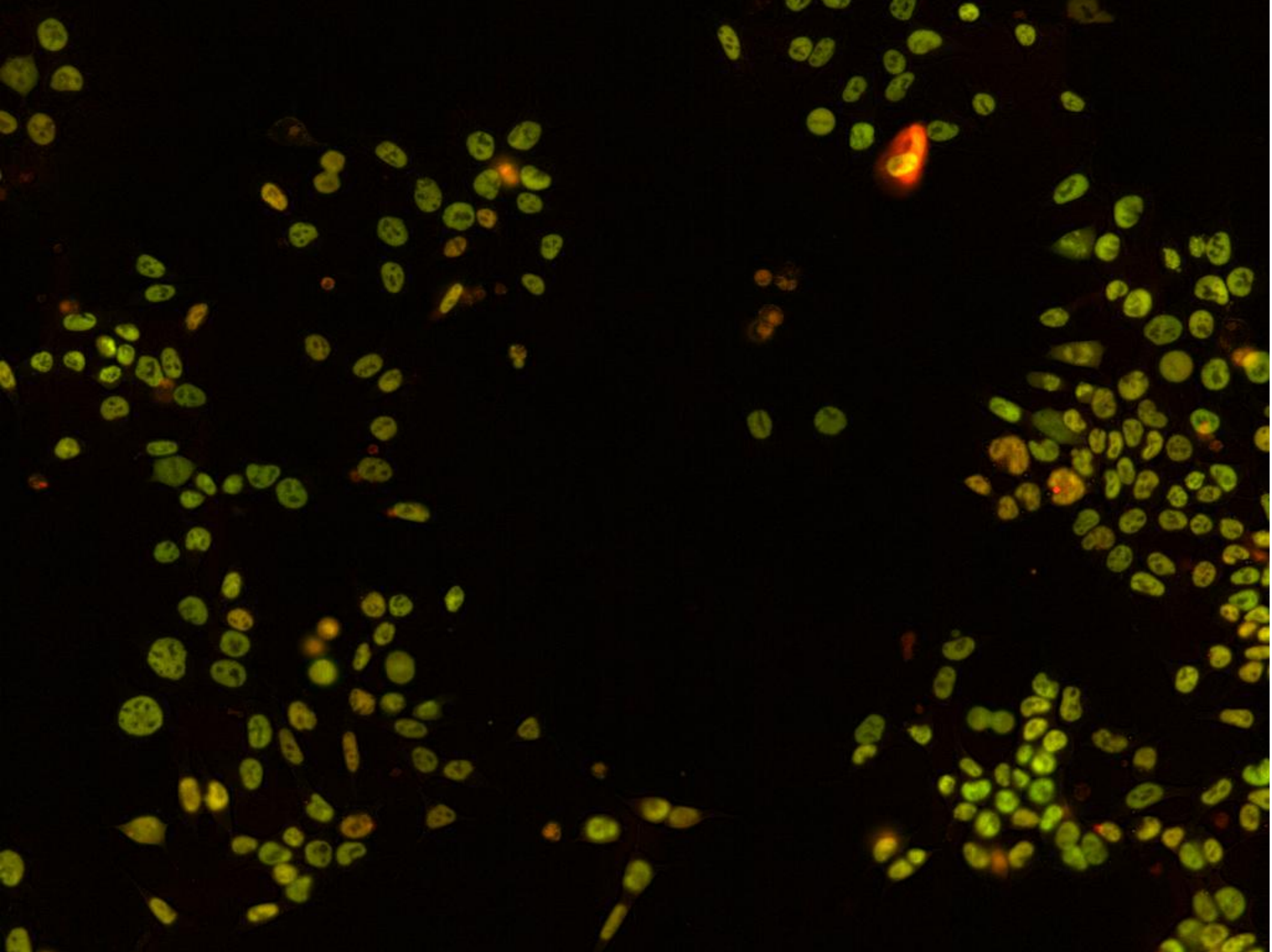
Transcription Synthesis Maturation/Degradation

$$\dot{B}_G = s - (m_g + k)B_G$$

$$\dot{G} = m_g B_G - kG$$

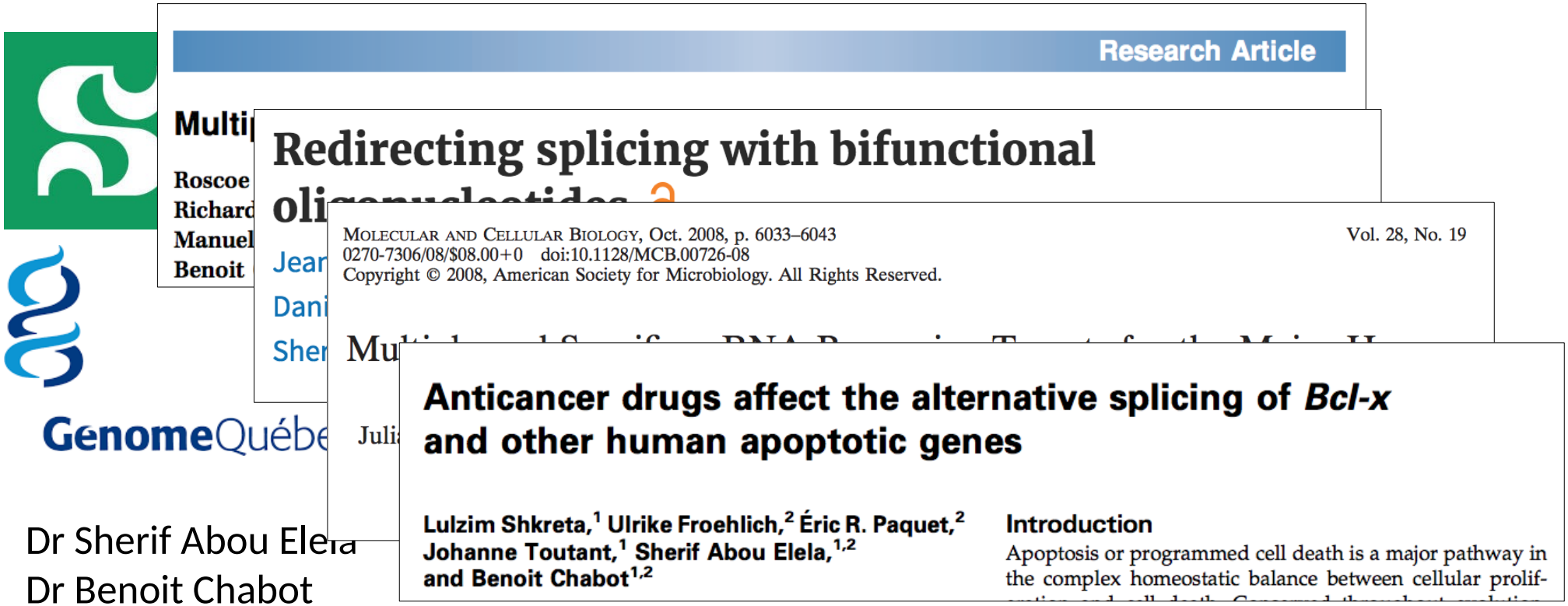
$$\dot{B}_R = s - (m_r + k)B_R$$

$$\dot{R} = m_r B_R - kR$$



Director of Bioinformatics

1999 2003 2005 2006 2007 2011 2015 2017



GenomeQuébec

Dr Sherif Abou Elela
Dr Benoit Chabot
Dr Jean-Pierre Perreault
Dr Claudine Rancourt
Dr Raymund Wellinger

RNOMICS PLATFORM

Director of bioinformatics

1999 2003 2005 2006 2007 2011 2015 2017

JEM

Article

Molecular Cell

Short Article

Cell PRESS

Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail

7788–7805 *Nucleic Acids Research, 2012, Vol. 40, No. 16* *Published online 4 June 2012*

Current Biology *18*, 1142–1146, August 5, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.06.071

Functional and Structural Basis for a Bacteriophage Homolog of Human RAD52

Mickaël Ploquin,[1] Ali Bransi,[1] Eric R. Paquet,[1]
Alicja Z. Stasiak,[2] Andrzej Stasiak,[2] Xiong Yu,[3]
Anna M. Cieslinska,[3] Edward H. Egelman,[3]
Sylvain Moineau,[4,5] and Jean-Yves Masson[1,6,*]

time, we propose a viral homolog o
acid, phylogenic, functional, and stru

Results and Discussion

Ph.D. Personalized medicine

1999 2003 2005 2006 2007 2011 2015 2017

Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype FREE

Eric R. Paquet, Michael T. Hallett

JNCI: Journal of the National Cancer Institute, Volume 107, Issue 1, 1 January 2015,

The Journal of Pathology: Clinical Research
J Path: Clin Res July 2015; **1**: 160–172
Published online 10 March 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cjp2.17

Original Article

A 12-gene signature to distinguish colon cancer patients with better clinical outcome following treatment with 5-fluorouracil or FOLFIRI

Eric R. Paquet,[1,2] Jing Cui,[3] David Davidson,[4] Natalia Pietrosemoli,[3] Houssein Hajj Hassan,[5] Serges P. Tsofack,[1] Annie Maltais,[1] Michael T. Hallett,[2] Mauro Delorenzi,[3,6] Gerald Batist,[4] Raquel Aloyz[4] and Michel Lebel[7]*

EDITORIAL

Prognostic Tests for Estrogen Receptor-Positive Breast Cancer
Need for Global Consideration and Further Evolution

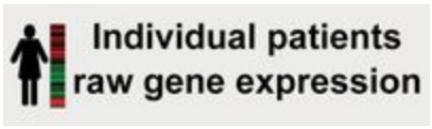
Jee Ye Kim, MD; Seung-il Kim, MD, PhD; Soonmyung Paik, MD

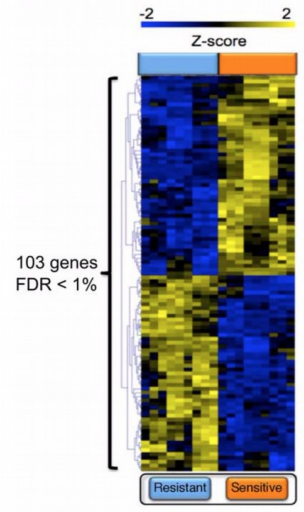
EDITORIAL

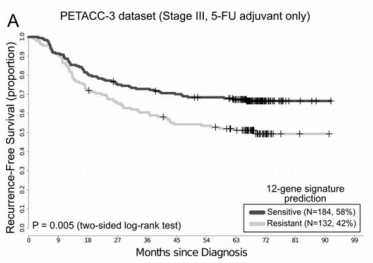
Making Breast Cancer Molecular Subtypes Robust?

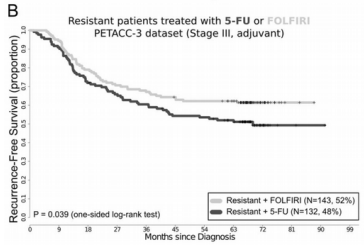
Johan Staaf, Markus Ringnér

Affiliation of authors: Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden.
Correspondence to: Markus Ringnér, PhD, Division of Oncology and Pathology, Lund University, Medicon Village, Building 404:B2, Scheelevägen 2, SE-223 81 Lund, Sweden (e-mail: markus.ringner@med.lu.se).

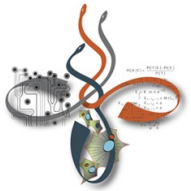








McGill



Breast cancer informatics group.
Dr Mike Hallett

Funding

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

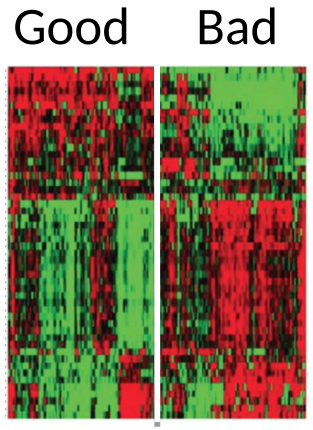
Q-CROC
Consortium de Recherche en Oncologie Clinique du Québec

Plan

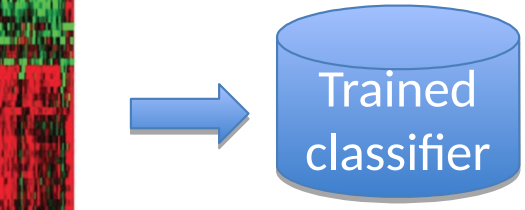
- List of pitfalls
  - Bad experimental design
  - Inadequate statistics
  - Missing background distribution
  - Not knowing what you are doing
- Applied machine learning with examples in systems biology
  - QC
    - Important plots
      - Clustering and Heatmaps
      - Boxplots
      - PCA
  - Pre-processing
  - Imputation
  - Class imbalance
  - Features selection in P >> N [mostly genomics]
    - Regularization
  - Kernel trick
  - Boosting
  - Personalized medicine and MAQC-II
  - Image analysis : features extraction

Prototype : Breast cancer personalized medicine

Training

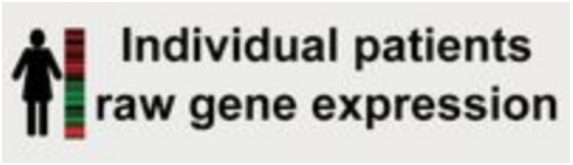


p >> n



Good outcome

Bad outcome



Good outcome (no chemo)

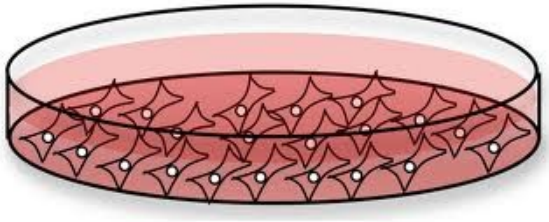
Bad outcome (chemo)

PITFALL #1
BAD EXPERIMENTAL DESIGN

Experimental design

- Different type of replicates :
  - Technical
  - Biological
- Batch effect

Different type of replicates (technical)



RNA

1/2 1/2

Note : This is **NOT** n=2

Why would you want to do this?

Comments of technical repeats

- Generally useless EXCEPT if :
  - You are developing a new protocol or a new technology and you want to show reproducibility.
  - In most cases (ie when biological replicates are not too expensive) you want to favor biological replicates.
- Technical replicates are not N = 2.
  - Negligible statistical utility.
  - Always favor biological replicates.

Biological replicates

RNA#1

RNA#2

Note : This is a n=2

Why should you favor this?

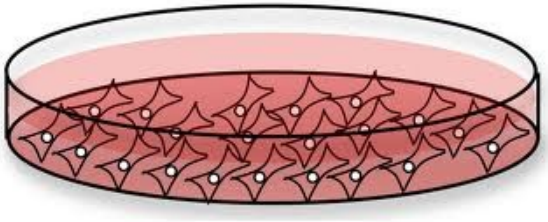
Comments on replicates

- Favor biological replicates when affordable.

BATCH EFFECT

What is a batch?

Drug 1



Full moon

Student 1

Center 1

Morning

RNA#1

Drug 2

New moon

Student 2

Center 2

Afternoon

RNA#2

Why should we take into account the batch in our experiment?

Is it frequent? Yes! It is too frequent!

Widespead batch effect in the litterature

Study description*	Known variable used as a surrogate			Principal components used as a surrogate			Association with outcome	Refs
	Surrogate‡	Confounding (%)§	Susceptible features (%)‖	Principal components rank of surrogate (correlation)¶	Principal components rank of outcome (correlation)#	Susceptible features (%)**	Significant features (%)‡‡	
Data set 1: gene expression microarray, Affymetrix (N_p = 22,283)	Date	29.7	50.5	1 (0.570)	1 (0.649)	91.6	71.9	9
Data set 2: gene expression, Affymetrix (N_p = 4167)	Date	77.6	73.7	1 (0.922)	1 (0.668)	98.5	62.2	2
Data set 3: mass spectrometry (N_p = 15,154)	Processing group	100	51.7	2 (0.344)	2 (0.344)	99.7	51.7	3
Data set 4: copy number variation, Affymetrix (N_p = 945,806)	Date	29.2	99.5	2 (0.921)	3 (0.485)	99.8	98.8	16
Data set 5: copy number variation, Affymetrix (N_p = 945,806)	Date	12.2	83.8	1 (0.553)	1 (0.137)	99.8	74.1	17
Data set 6: gene expression, Affymetrix (N_p = 22,277)	Processing group	NA	83.8	5 (0.369)	NA	97.1	NA	18
Data set 7: gene expression, Agilent (N_p = 17,594)	Date	NA	62.8	2 (0.248)	NA	96.7	NA	18
Data set 8: DNA methylation, Agilent (N_p = 27,578)	Processing group	NA	78.6	3 (0.381)	NA	99.8	NA	18
Data set 9: DNA sequencing, Solexa (N_p = 2,886)	Date	24.2	32.1	2 (0.846)	2 (0.213)	72.7	16.9	1000 Genomes Project

Cancer research

Nature genetics

The Lancet

Nature

Am. J. Hum. Genet

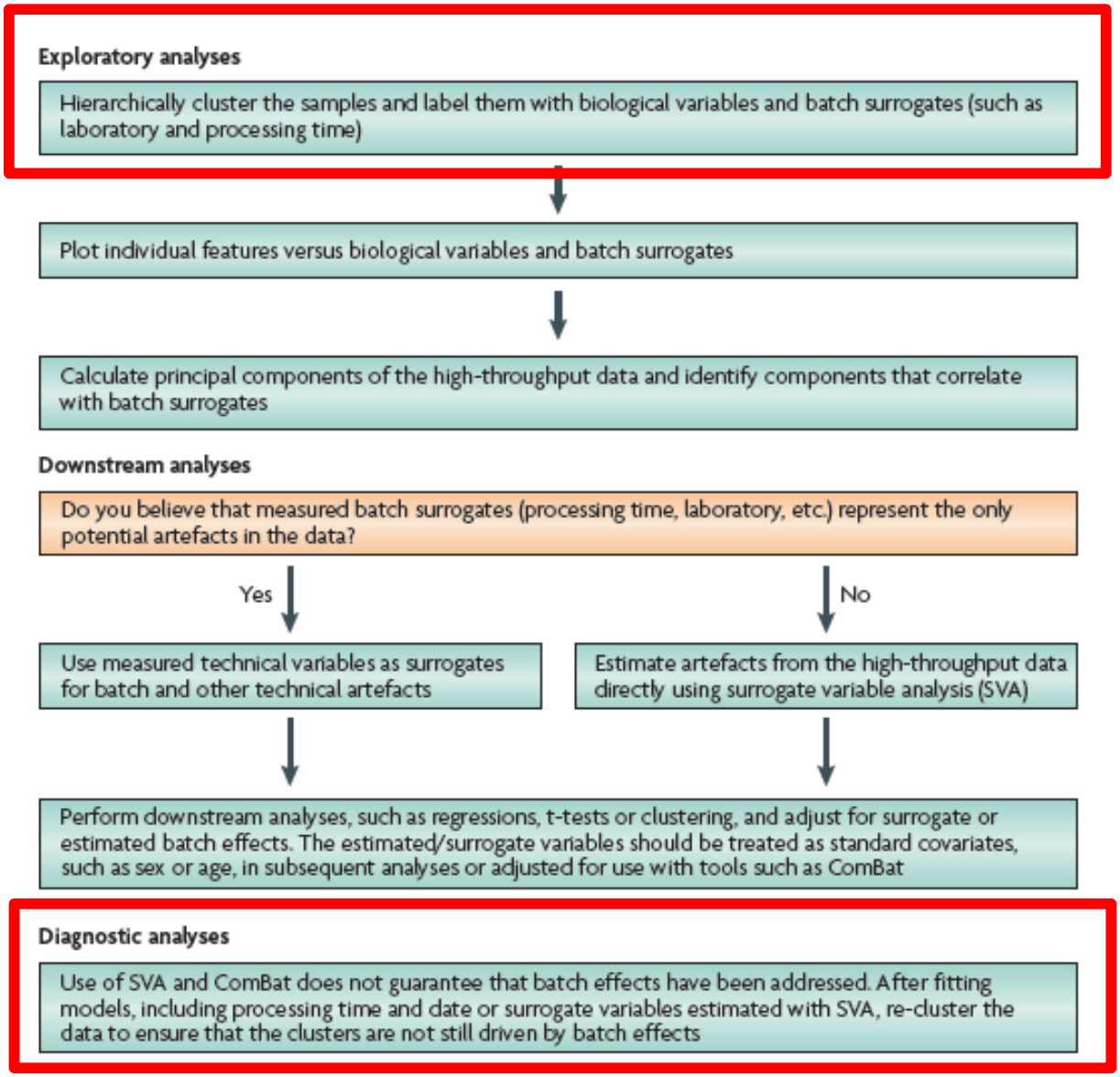
Nature

Nature

Nature

Nat Rev Genet. 2010 Oct;11(10):733-9.

How to detect and correct for batch effect



Use this sva package
In R.

Good vignette.

Nat Rev Genet. 2010 Oct;11(10):733-9.

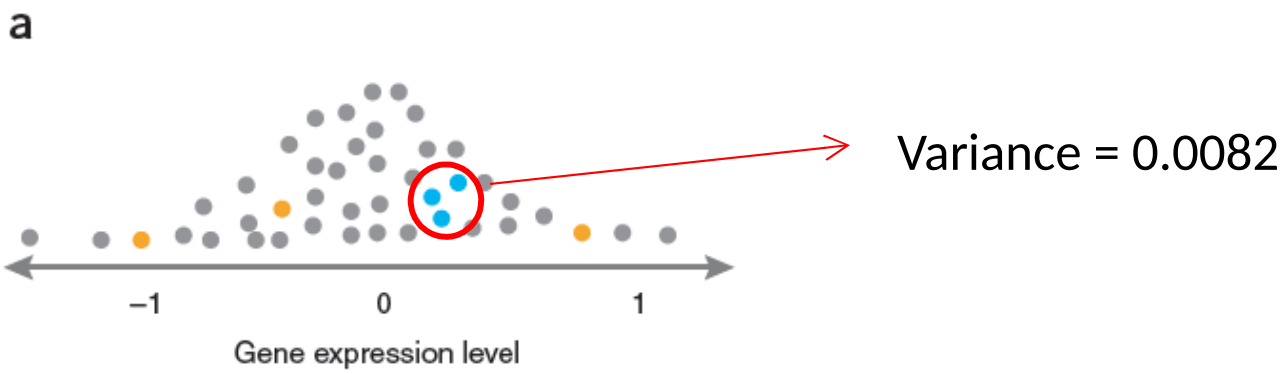
Summary batch effect

- When planning an experiment, think about all the possible variables and confoundings.
- It is important because this could introduce a lot of bias
- If the batch effect is not too strong this could be corrected using tools like combat in the R sva package.

PITFALL # 2. INADEQUATE STATISTICS

« standard » statistics and p >> n problem

T-statistic = mean / variance



What happen when the variance → 0 ?

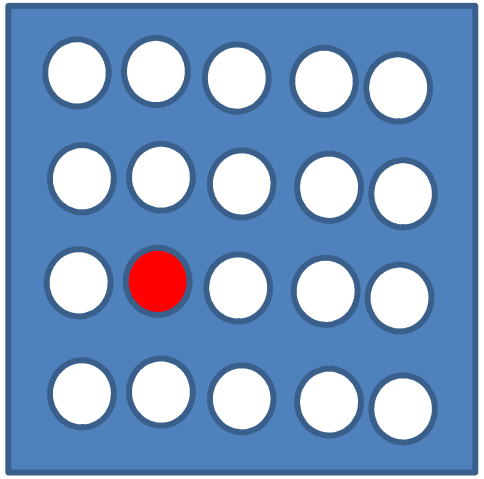
P-value → 0 !!! We need to correct for this.

Methods R packages : SAM, **limma**, Ebayes

Nature Biotechnology 2010;
28(4):337-40

Multiple hypothesis testing

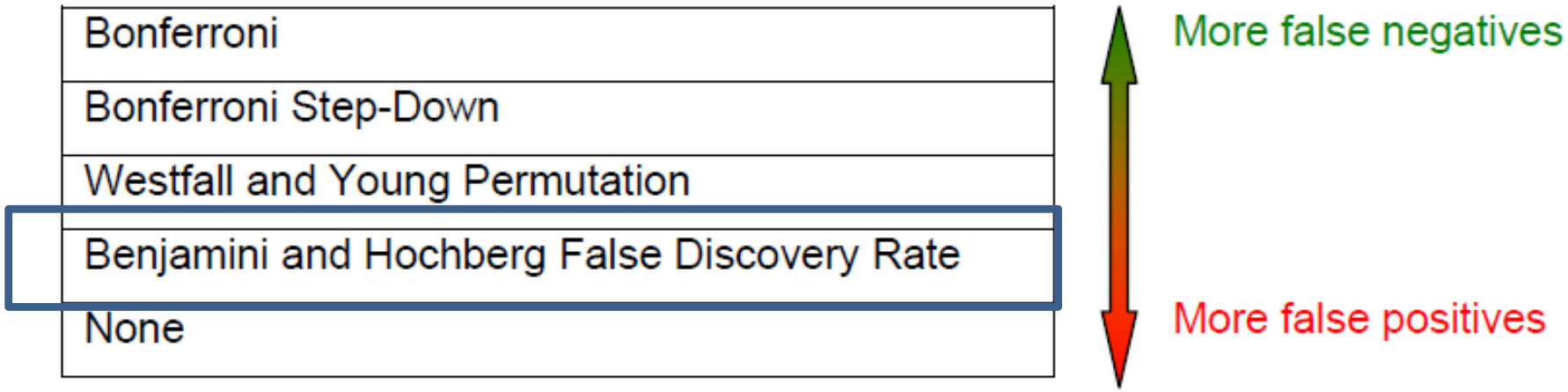
- What is the chance of picking up the red ball with one draw?
- What is the chance of picking up the red ball with 20 draws? ~ 64% 100 times = 99,4%



- Testing 20 000 times the same statistical hypothesis with a 0.05 level of significance
- False positive (balle rouge) picked = 20 000 * 0.05 = 1000

How to correct for this

- A compromise between false positive [picking up the red ball] et false negative [not picking a real gene]
- Different approaches (use p.adjust in R)



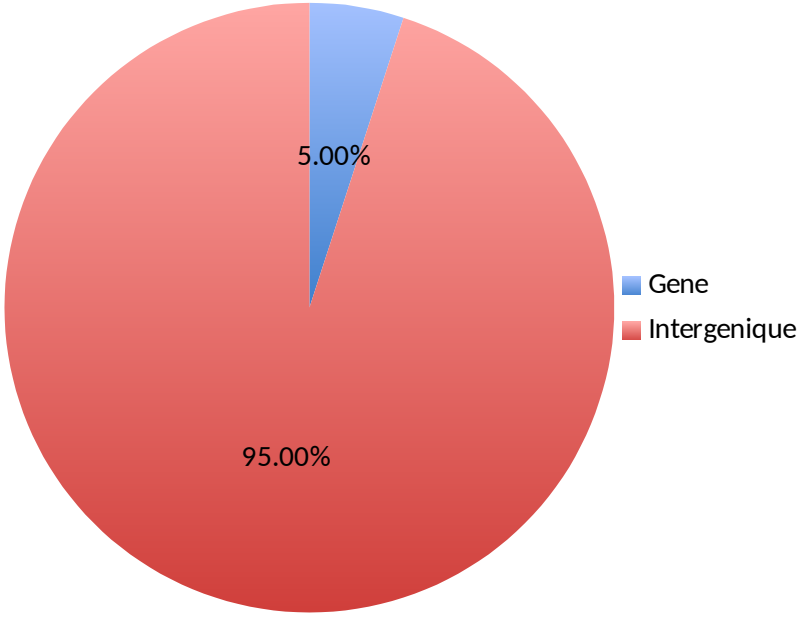
Take home message

- Some statistics are designed for genomic or systems biology (p >> n) SAM, limma, etc.
- Pay a special attention when testing more than one time a statistical hypothesis (big p). Need to correct the p-values

PITFALL #3 : MISSING THE BACKGROUND DISTRIBUTION

Example #1

- ChIP-seq of a transcription factor (TF) on the human genome

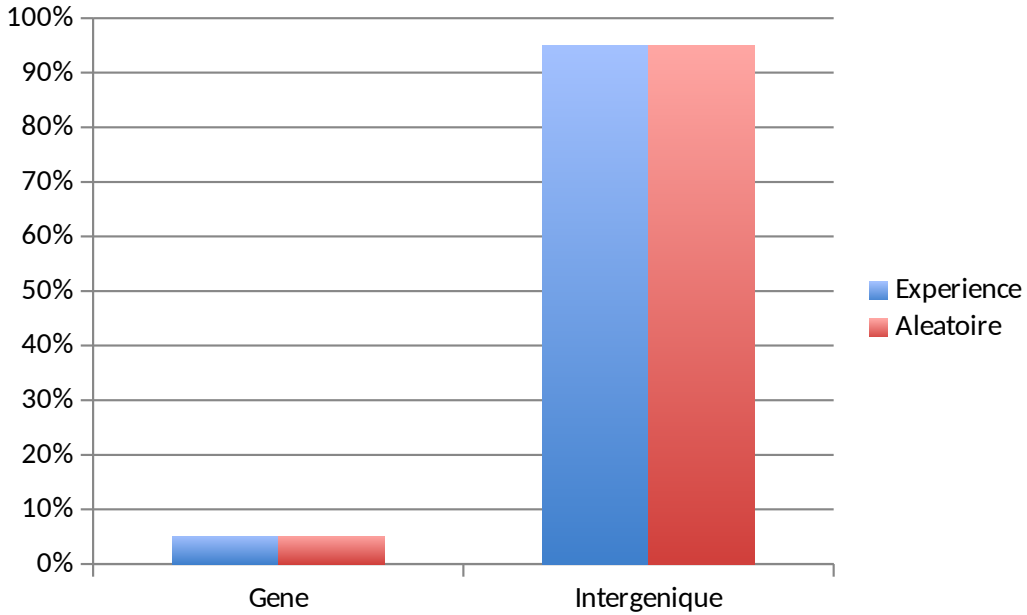


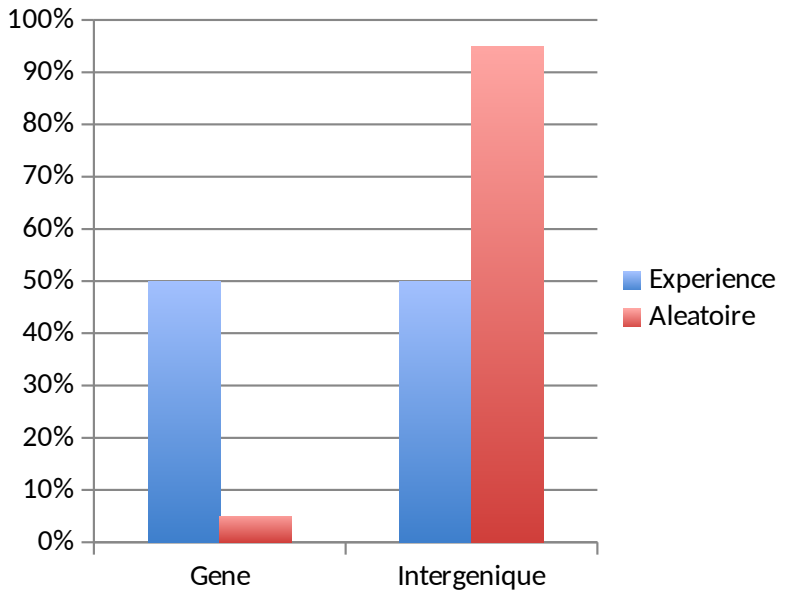
What is the background distribution?

5% of the genome code for genes the remaining is intergenic or intronic regions…

Consequently this TF follows exactly the background distribution! No enrichment

Example #1 corrected

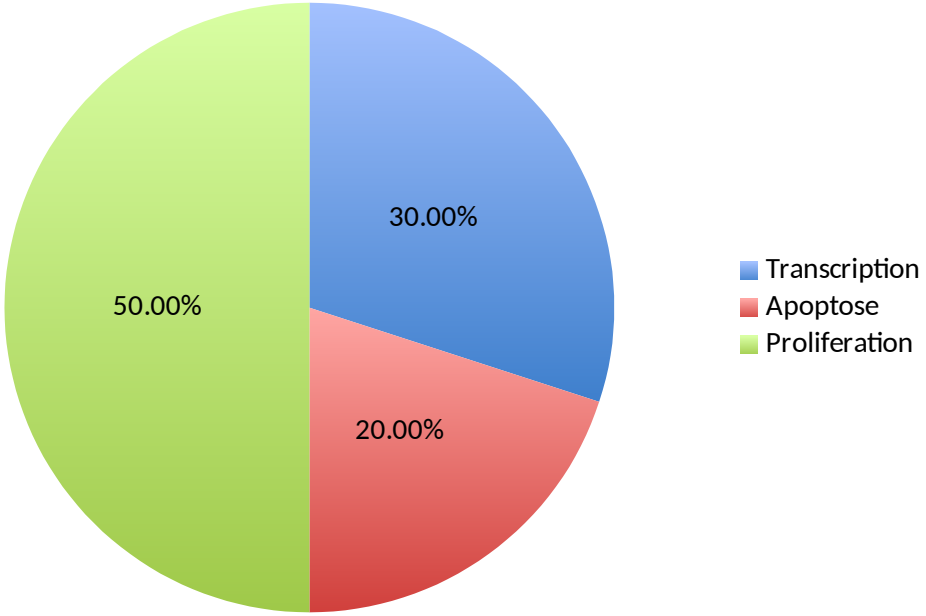




You should favor dual band barplots to piecharts. This way you could present the background distribution (test significance using a chisq.test or a fisher.test).

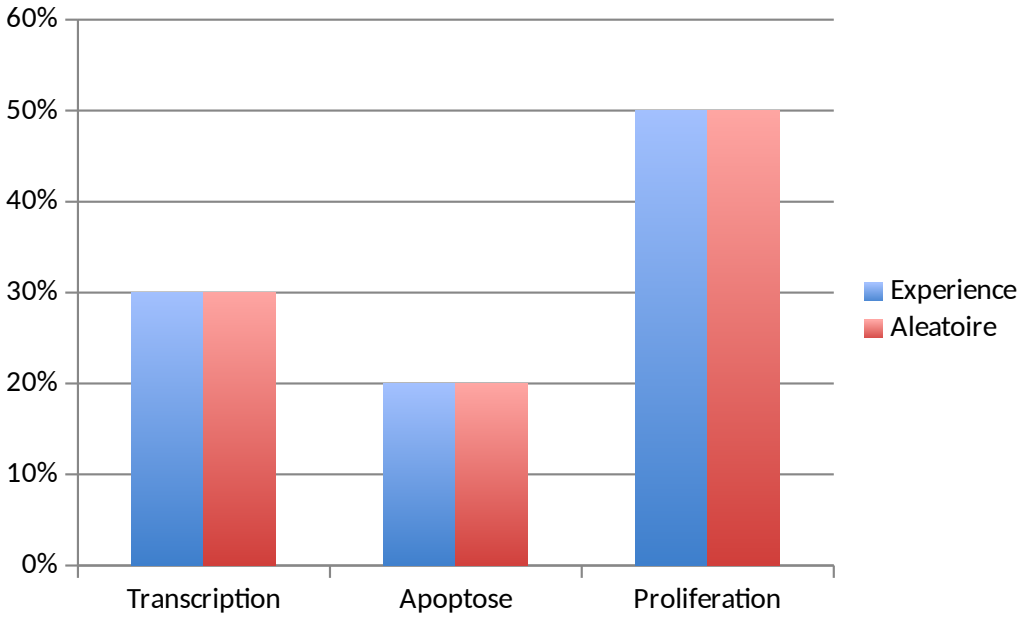
Example #2

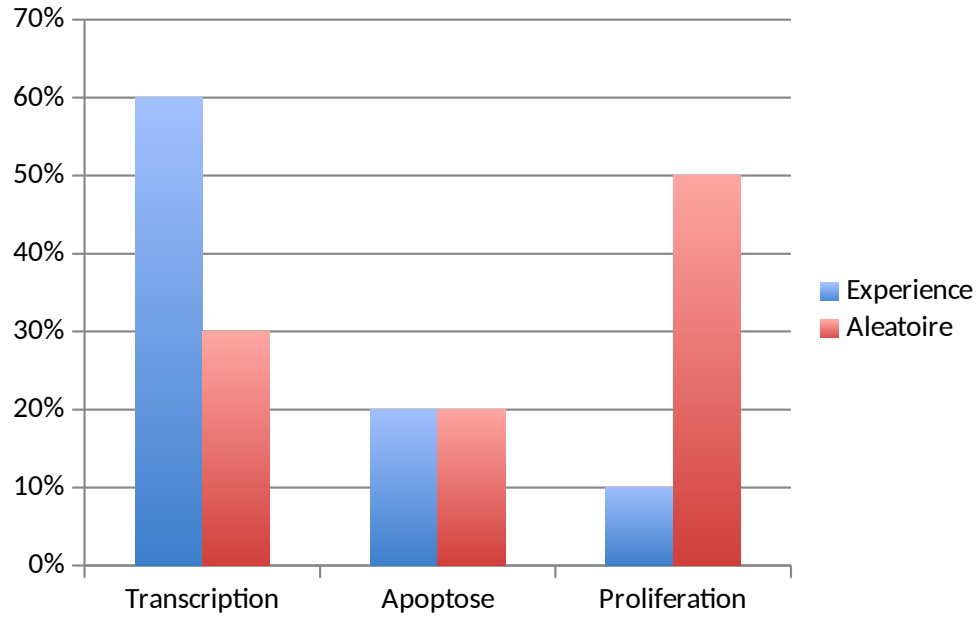
- RNA-seq experiment. You obtained 100 genes significantly modulated (human). What are the enriched biological processes in the list of 100 genes?



What is the random distribution? *ie* what is the fraction of genes in the human genome implicated in Transcription, apoptosis or Proliferation?

Example #2 corrected

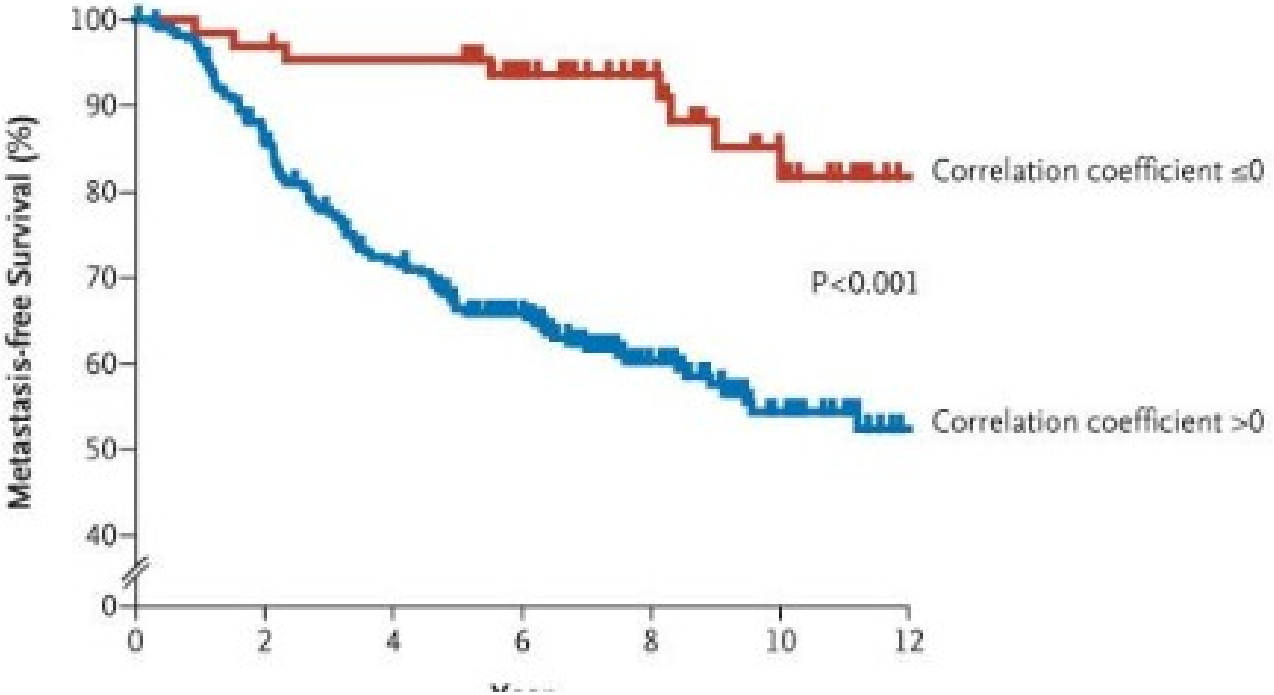




You should favor dual band barplots to piecharts. This way you could present the background distribution (test significance using a chisq.test or a fisher.test).

Example #3

- You just found a gene signature associated with outcome in breast cancer.

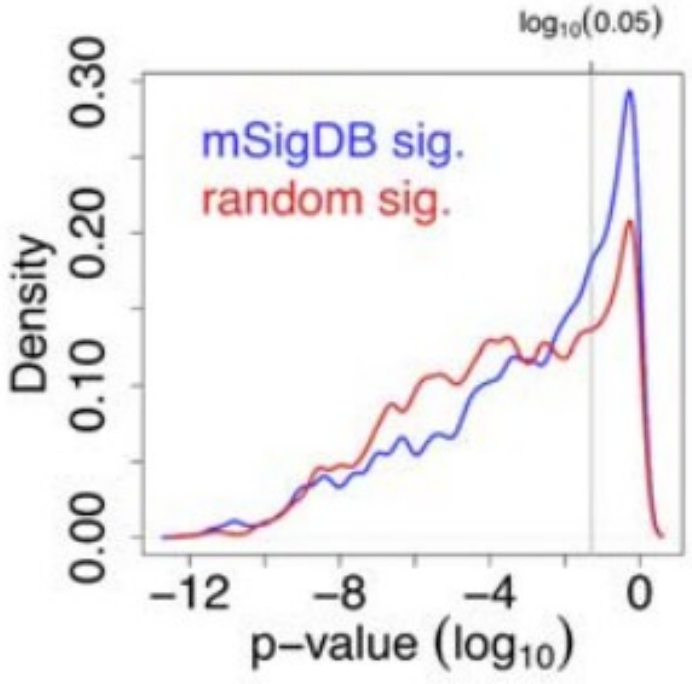


What is the likelihood of this type of signature?

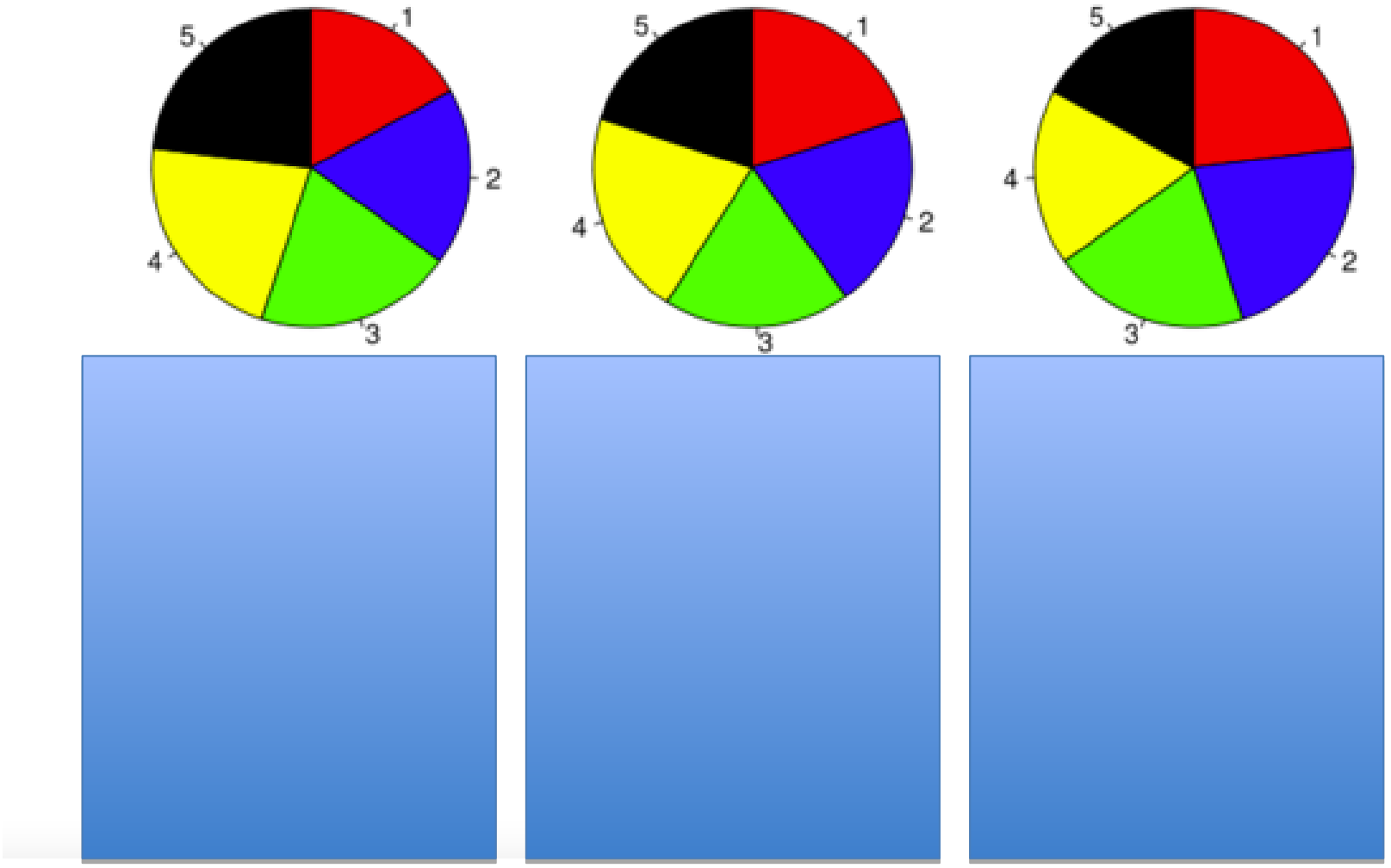
Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], **Jacques E. Dumont**[2], **Vincent Detours**[2,3]*

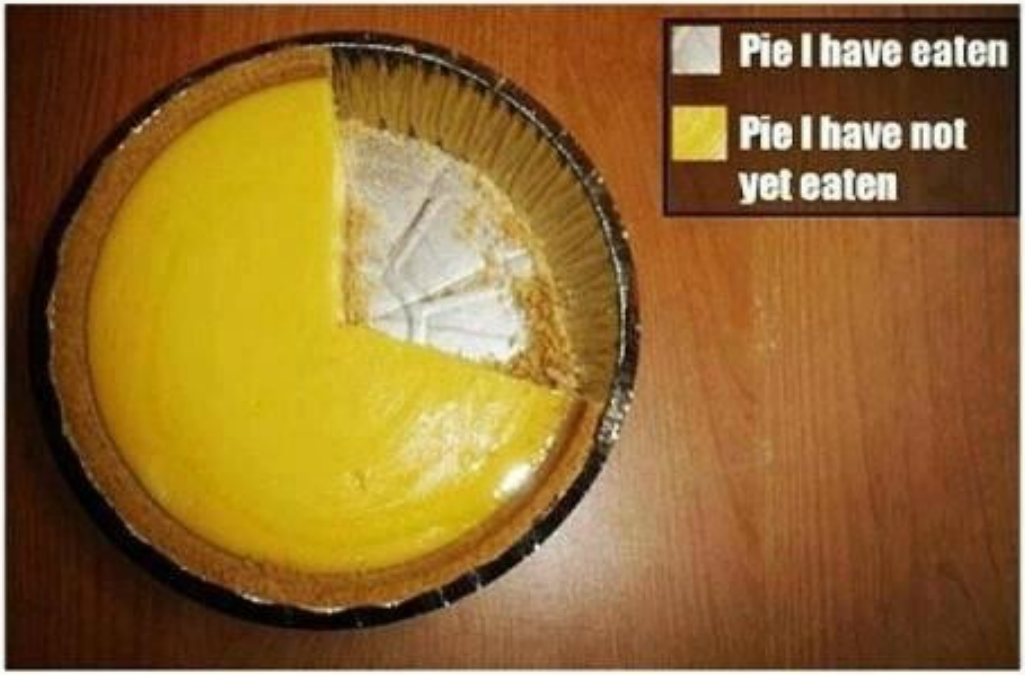
1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium



Quiz pie-chart



World's Most Accurate Pie Chart



Take home message

- Always compare to the background distribution
- Use pie-charts moderately ie you need to also show the background distribution right?
- Favor barplots (to show the background distribution).

PITFALL #4 : NOT KNOWING WHAT YOU ARE DOING

Richard Simon

Dr. Richard Simon

Associate Director, Division of Cancer Treatment and Diagnosis

Director, Biometric Research Program

Chief, Computational & Systems Biology Branch

Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting

Alain Dupuy ✉, Richard M. Simon

JNCI: Journal of the National Cancer Institute, Volume 99, Issue 2, 17 January 2007, Pages 147–157, https://doi.org/10.1093/jnci/djk018

Published: 17 January 2007 **Article history** ▾

Major Flaws Found in 40 Studies Published in 2004

- Inadequate control of multiple comparisons in gene finding
  - 9/23 studies had unclear or inadequate methods to deal with false positives
    - 10,000 genes x .05 significance level = 500 false positives
- Misleading report of prediction accuracy
  - 12/28 reports based on incomplete cross-validation
- Misleading use of cluster analysis
  - 13/28 studies invalidly claimed that expression clusters based on differentially expressed genes could help distinguish clinical outcomes

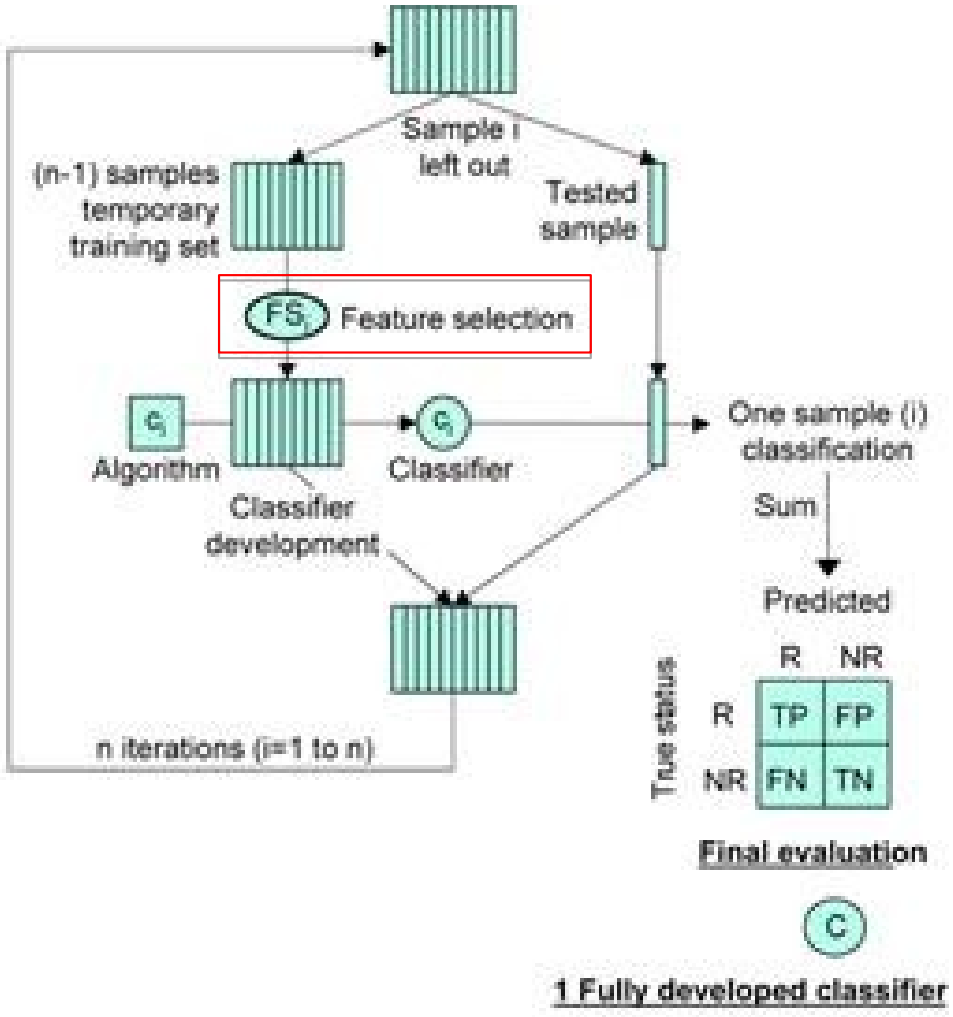
- 50% of studies contained one or more major flaws

One of the major flaw (can you spot it?)

```
require(caret)
require(gplots)

## This is important
set.seed(1234)

data <- matrix(rnorm(6000*50),nrow=50,ncol=6000)
colnames(data) <- as.character(1:6000)
cl <- c(rep(1,25),rep(2,25)) ## 1 = normal, 2 = cancer

## Select genes
pv.feat <- apply(data,2,function(x){
    t.test(x[cl==1],x[cl==2])$p.value
})
top.20 <- order(pv.feat)[1:20]

heatmap.2(data[,top.20],trace="none",
RowSideColors=as.character(cl),
col=colorpanel(50,"blue","white","red"))

## LOOCV
preds <- c()
for (looi in 1:nrow(data)){
    cur.t <- train(data[-looi,top.20],factor(cl[-looi]),method="knn")
    preds <- c(preds,predict(cur.t,data[looi,top.20,drop=F]))
}

table(preds,cl)
```

P >> n problem

Example with leave-one-out cross-validation



B. Leave-one-out cross-validation procedure

Even if it is time consuming feature selection should be done within the cross-validation

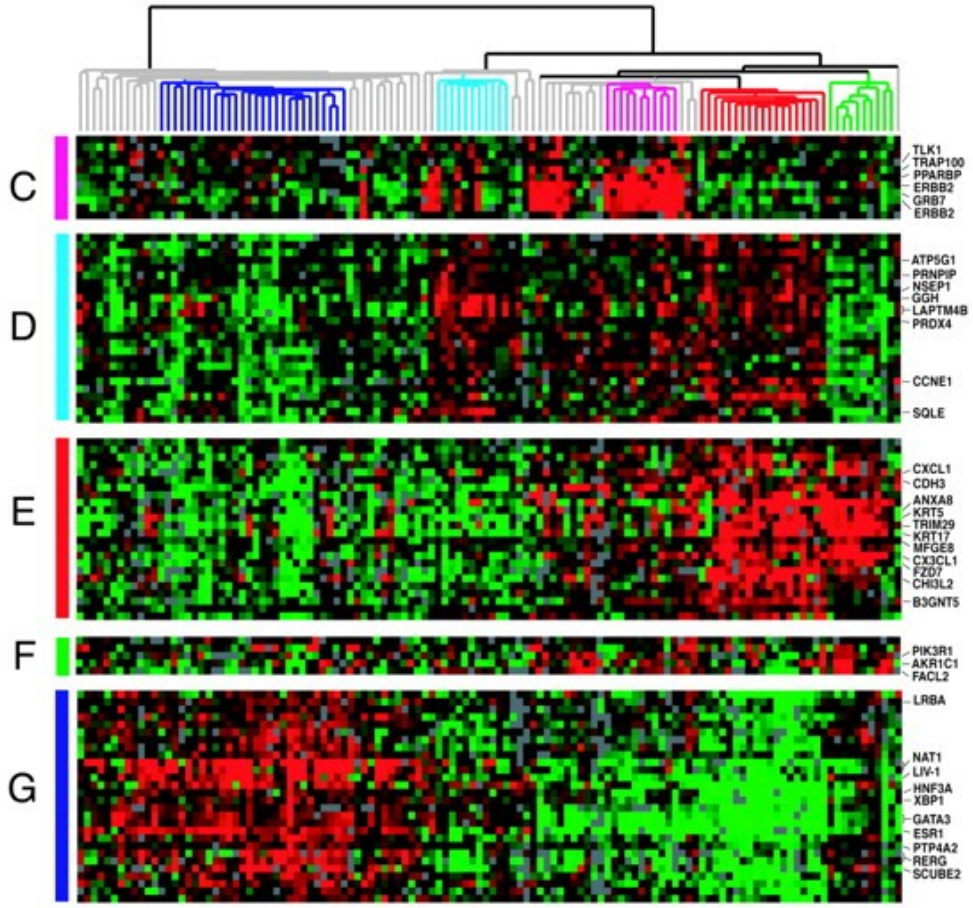
Democratization of machine learning via simple to use GUI interfaces ?

Take home message

- Know what you are doing
- The entire training process should be performed inside cross validation. DO NOT :
  - Select features
  - Normalized
  - Outside cross-validation

Clustering

Hierarchical clustering



Dependent on two things:

-Distance metrics

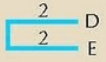
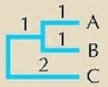
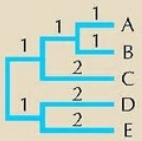
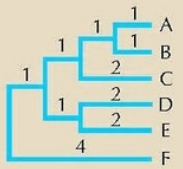
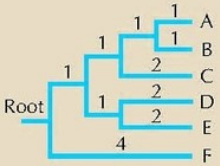
- -Euclidean
- -correlation
- -etc

-Agglomeration

- -complete
- -mean
- -ward

How to perform clustering?

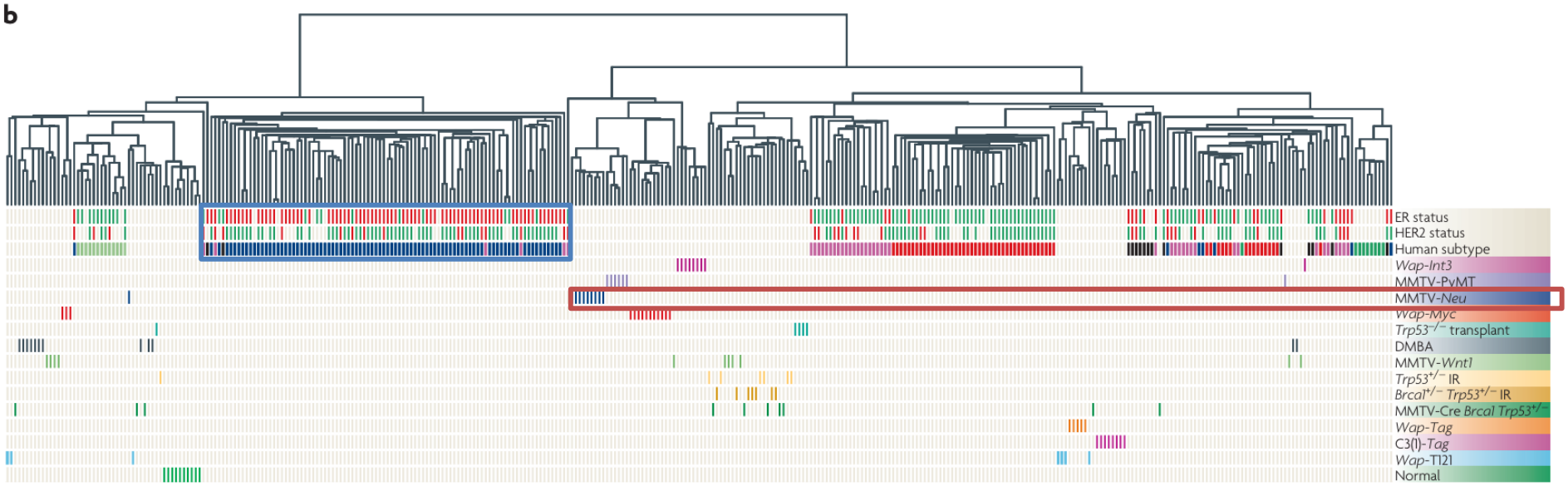
TABLE 27.9. Example of UPGMA tree construction

Step	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Distance matrix	OTUs A B C D E B 2 C 4 4 D 6 6 6 E 6 6 6 4 F 8 8 8 8 8	OTUs AB C D E C 4 D 6 6 E 6 6 4 F 8 8 8 8	OTUs AB C DE C 4 DE 6 6 DE 8 8 8	OTUs ABC DE DE 6 F 8 8	OTUs ABCDE F 8	No new matrix
Identify smallest D	A↔B = 2	AB↔C = 4 D↔E= 4	AB↔DE = 6 C↔DE = 6	ABC↔DE	ABCDE↔F	
Taxa joined	A and B	D and E	AB and C	ABC and DE	ABCDE and F	
Subtree						
Comments on tree drawing	The distance between A and B is 2 units. A subtree is drawn with the branch point halfway between the two. Thus, each branch is 1 unit in length.	Branching done as in Step 1. Because the distance from AB to C is also 4, that pair could have been selected as well.	First a subtree is drawn with AB and C:  The the AB subtree is attached to the AB branch at a point equal to the length of the A and B branches.	The tree is first done as in Step 3 with the ABC and DE subtrees replacing the branches.	The tree is now complete but unrooted.	The tree can then be rooted using midpoint rooting which tries to balance all the tips to reach the same end point. Note this is the tree that we started with to build the distance matrix.

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html.

Need two things : a distance metric + an agglomerative function. Need to mention both in publications.

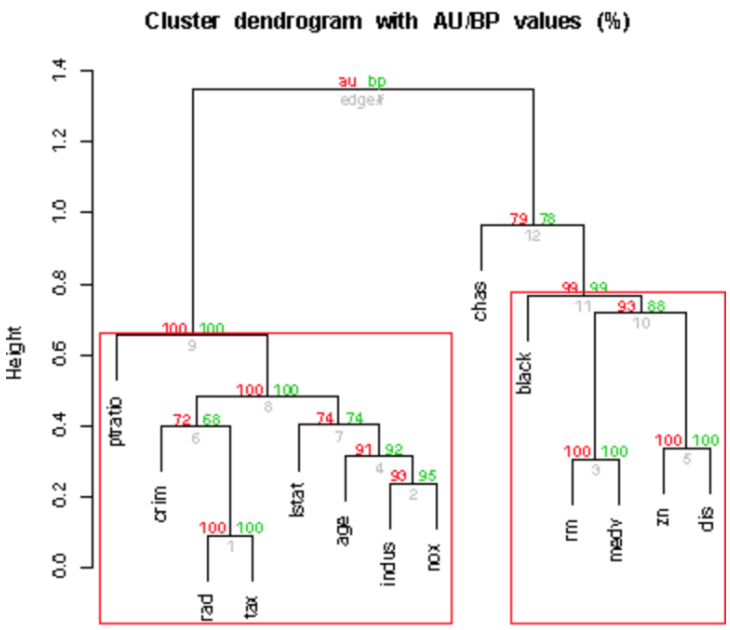
Know how to read it?



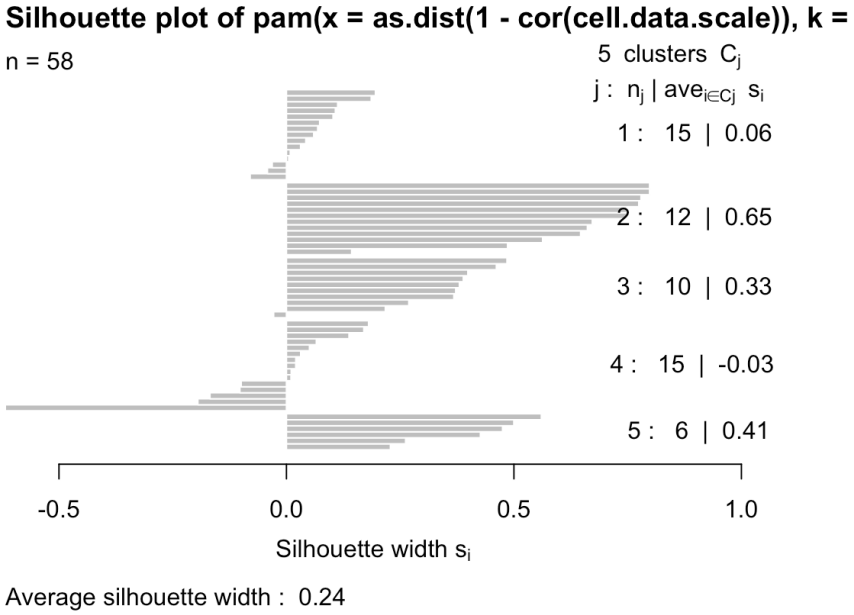
deficient models) subtypes. None of these genetically engineered mouse models were representative of ER^+ breast cancer. Furthermore, the tumours from MMTV-*Neu* GEM were more similar to human luminal tumours than to the human $ERBB2^+$ tumours. DMBA, 7,12-dimethylbenz(a) anthracene. Image reproduced from Ref. 107.

Modelling breast cancer: one size does not fit all. Tracy Vargo-Gogola and Jeffrey M. Rosen. Nat Rev Cancer. 2007 Sep;7(9):659-72

Need to test the stability of your clustering otherwise it is meaningless



Bootstrapping approach
Pvclust
How reproducible is the clustering if you repeat it multiple time on boostrapped data?

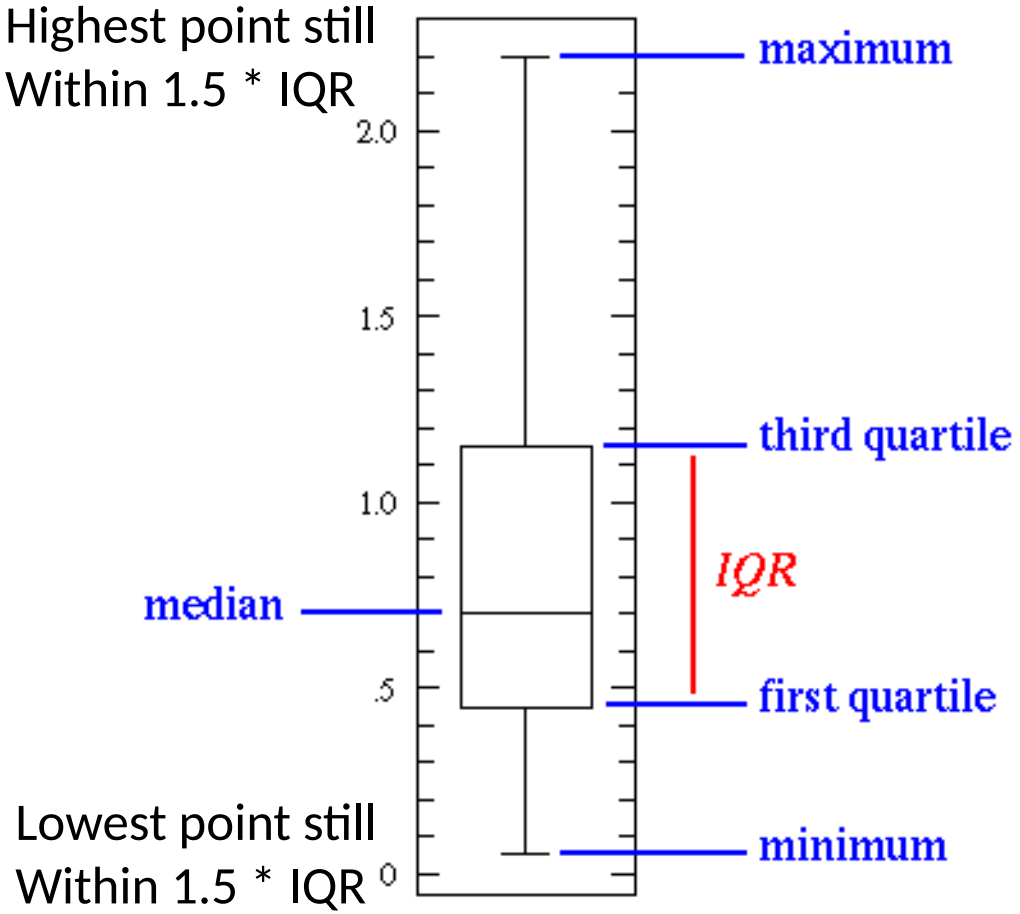


How similar is a sample compare to members of its own cluster versus members of the closest cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Boxplots

Useful to look rapidly at the distribution of your data



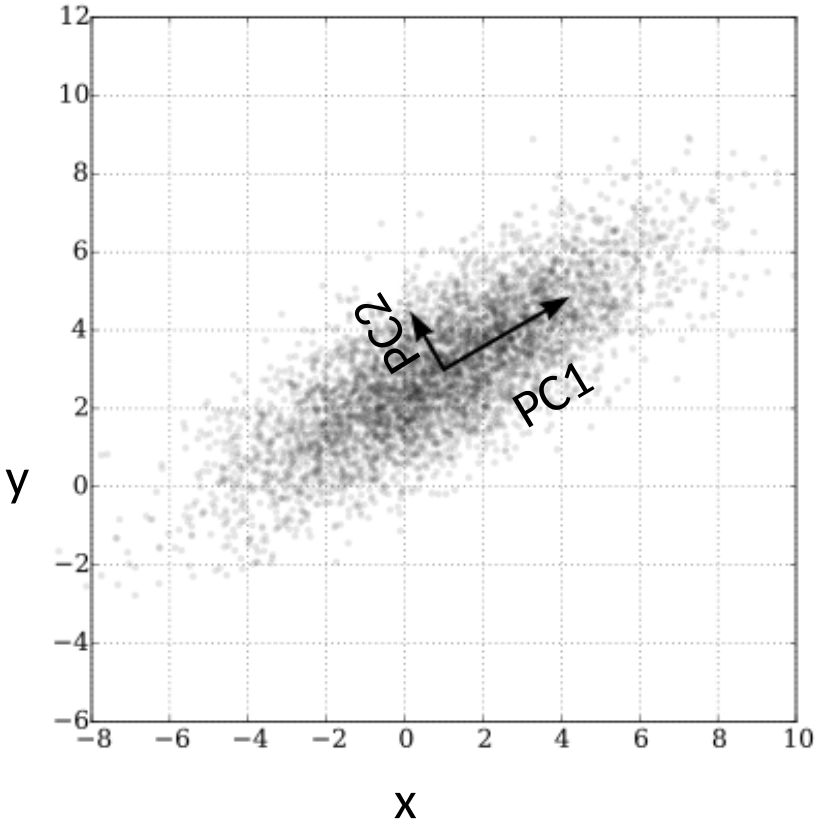
boxplot(data ~ class)

Usually nonparametric stats :

wilcox.test(...) 2 samples
kruskal.test(...) > 2 samples
dunn.test(...) posthic

Principal component analysis (PCA)

Principent component analysis



Transform the data in a way so

the first component get the largest variance

and the second othogonal to the first get the second largest variance, etc

prcomp() in R

You can use PCA to look at your data and also to reduce the dimensionality of your dataset.