Un statisticien chez les biologistes, pour quoi faire ?

Nécessité et difficultés de la conception à l'analyse

Sébastien DÉJEAN

@ sebastien.dejean@math.univ-toulouse.fr
 www.math.univ-toulouse.fr/~sdejean

INSTITUT de MATHEMATIQUES de TOULOUSE

math.univ-toulouse.fr

geno toul biostat

math.univ-toulouse.fr/biostat

Une feuille de route possible ◄ ◄ ◄

1) énoncer clairement une question précise
2) prévoir les méthodes d'analyse des données
3) mettre en place un plan d'expérience

4) acquérir les données

5) analyser des données
6) interpréter des résultats
7) répondre à la question posée

Biologiste et Statisticien en symbiose

Quelques définitions du mot Symbiose :

• association **durable** et **réciproque** entre deux ou plusieurs organismes vivants

• association **intime** et **durable** entre deux organismes hétérospécifiques (**espèces différentes**)

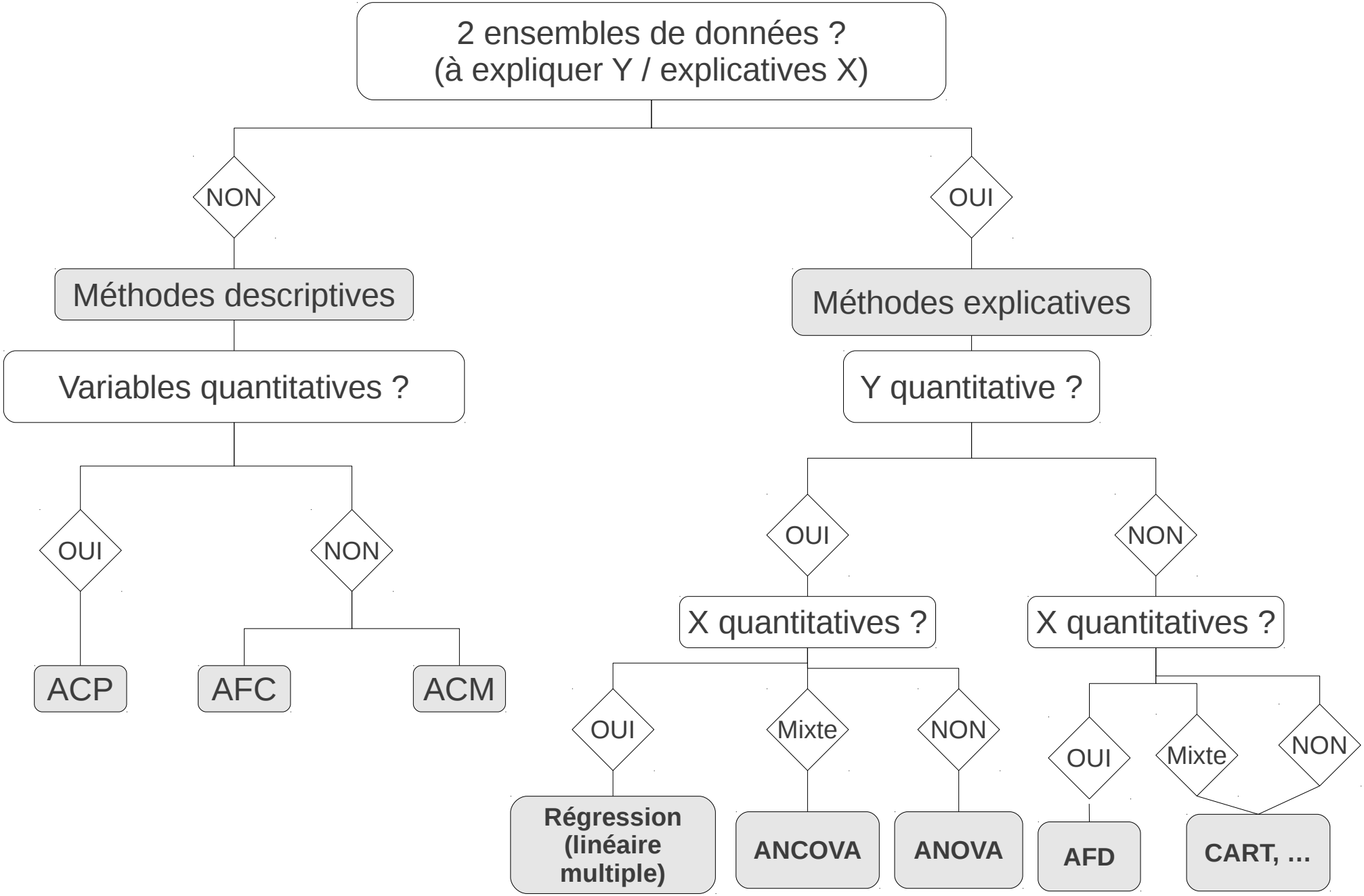
• relation écologique **obligatoire** qu'entretiennent des organismes d'espèce différente **vivant en contact direct** les uns avec les autres

• association **obligatoire** de deux ou de plusieurs organismes différents, les symbiotes, avec **bénéfice réciproque** et **qui leur permet de vivre**

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

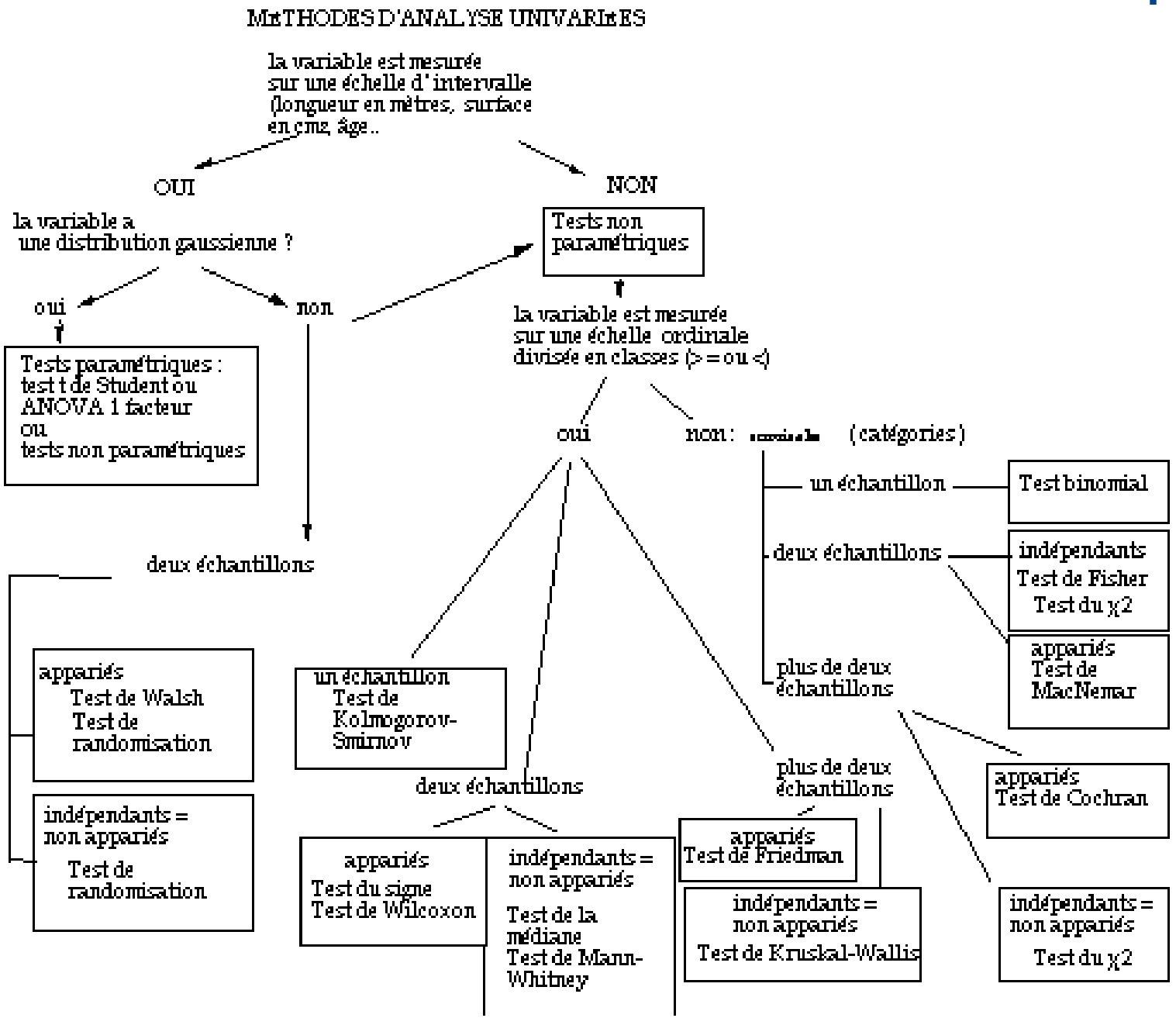
Quelques repères

Analyse multidimensionnelle



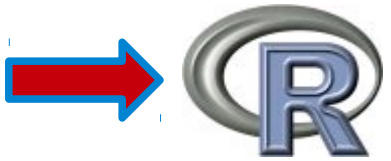
Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Tests statistiques



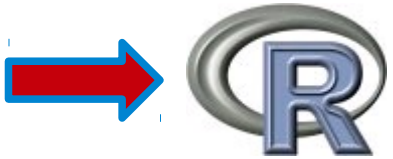
Logiciels ◀ ◀ ◀


XLSTAT
your data analysis solution




MathWorks


SYSTAT



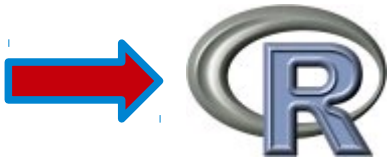

statgraphics

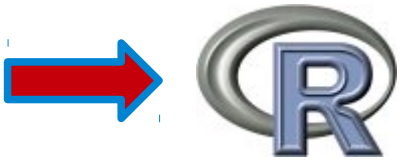

UMETRICS
AN MKS COMPANY
SIMCA
BY UMETRICS


STATA
Data Analysis and Statistical Software

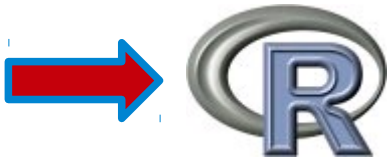

TANAGRA


sas
THE POWER TO KNOW.






Minitab



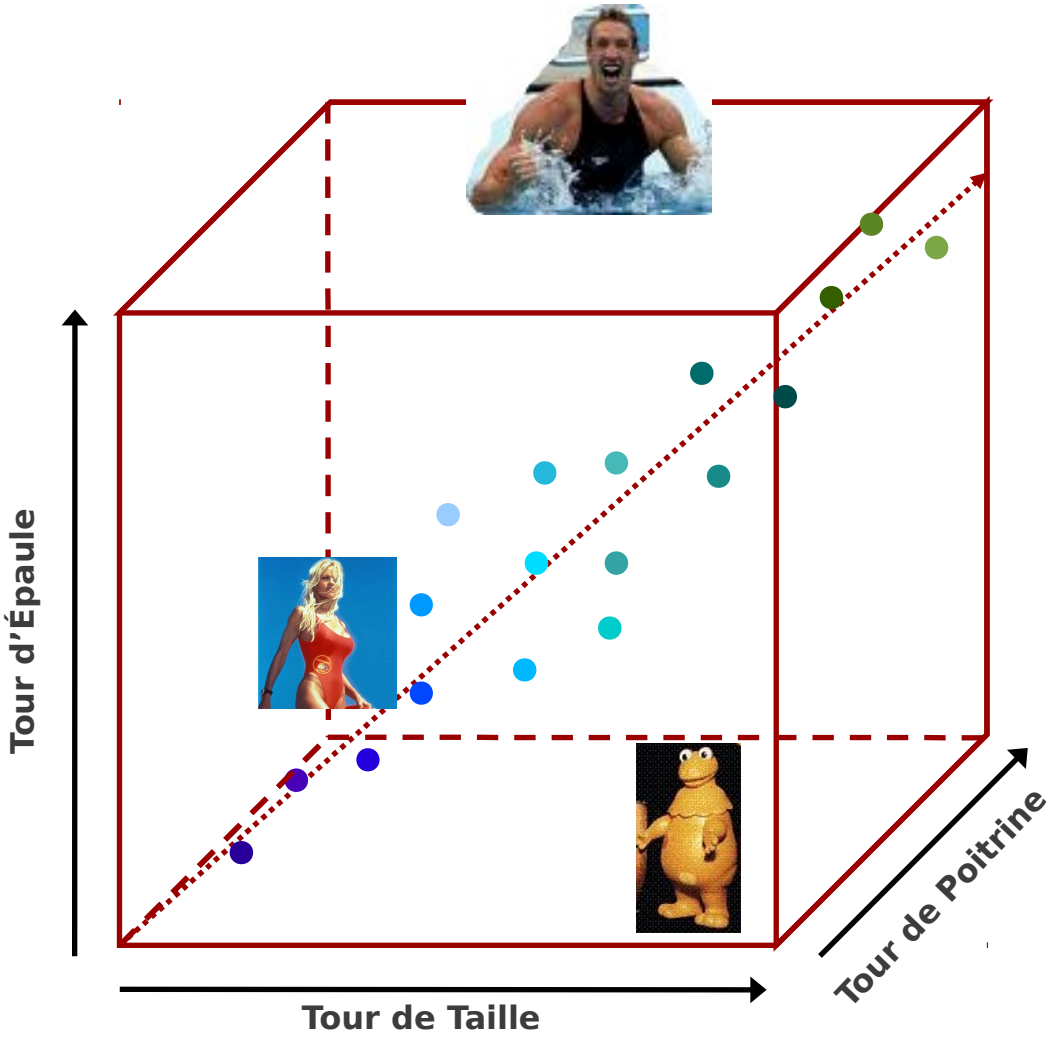

jmp
Discover JMP® 9
Download your free 30-day trial
9

Exemple 1

Analyse en Composantes Principales

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

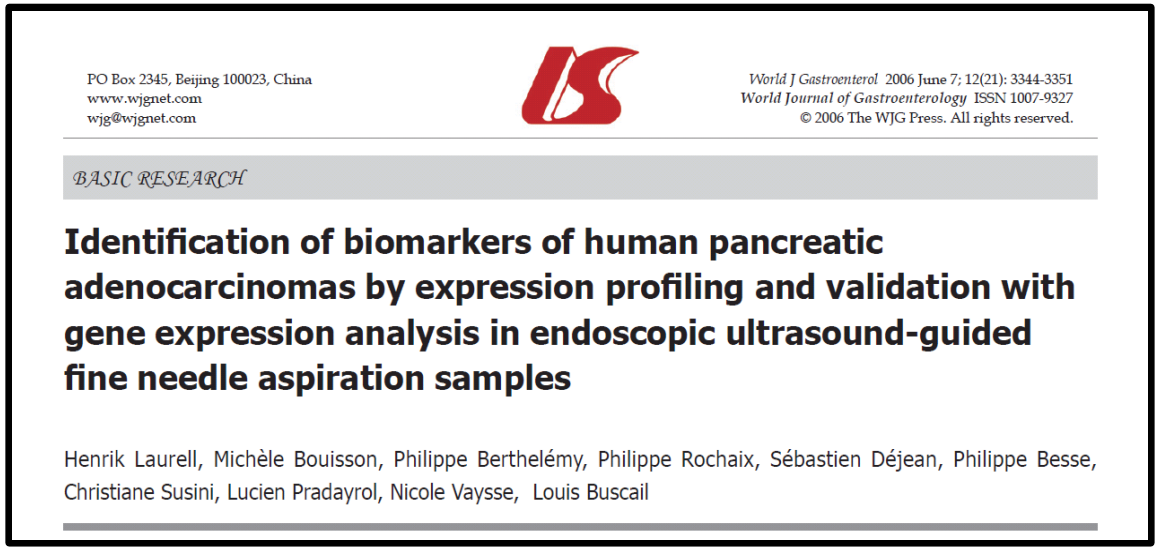
L'ACP en 1 dessin



1ère CP : « costauditude »

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Des données ◀ ◀ ◀



PO Box 2345, Beijing 100023, China
www.wjgnet.com
wjg@wjgnet.com

World J Gastroenterol 2006 June 7; 12(21): 3344-3351
World Journal of Gastroenterology ISSN 1007-9327

BASIC RESEARCH

Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation with gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples

Henrik Laurell, Michèle Bouisson, Philippe Berthelémy, Philippe Rochaix, Sébastien Déjean, Philippe Besse, Christiane Susini, Lucien Pradayrol, Nicole Vaysse, Louis Buscail

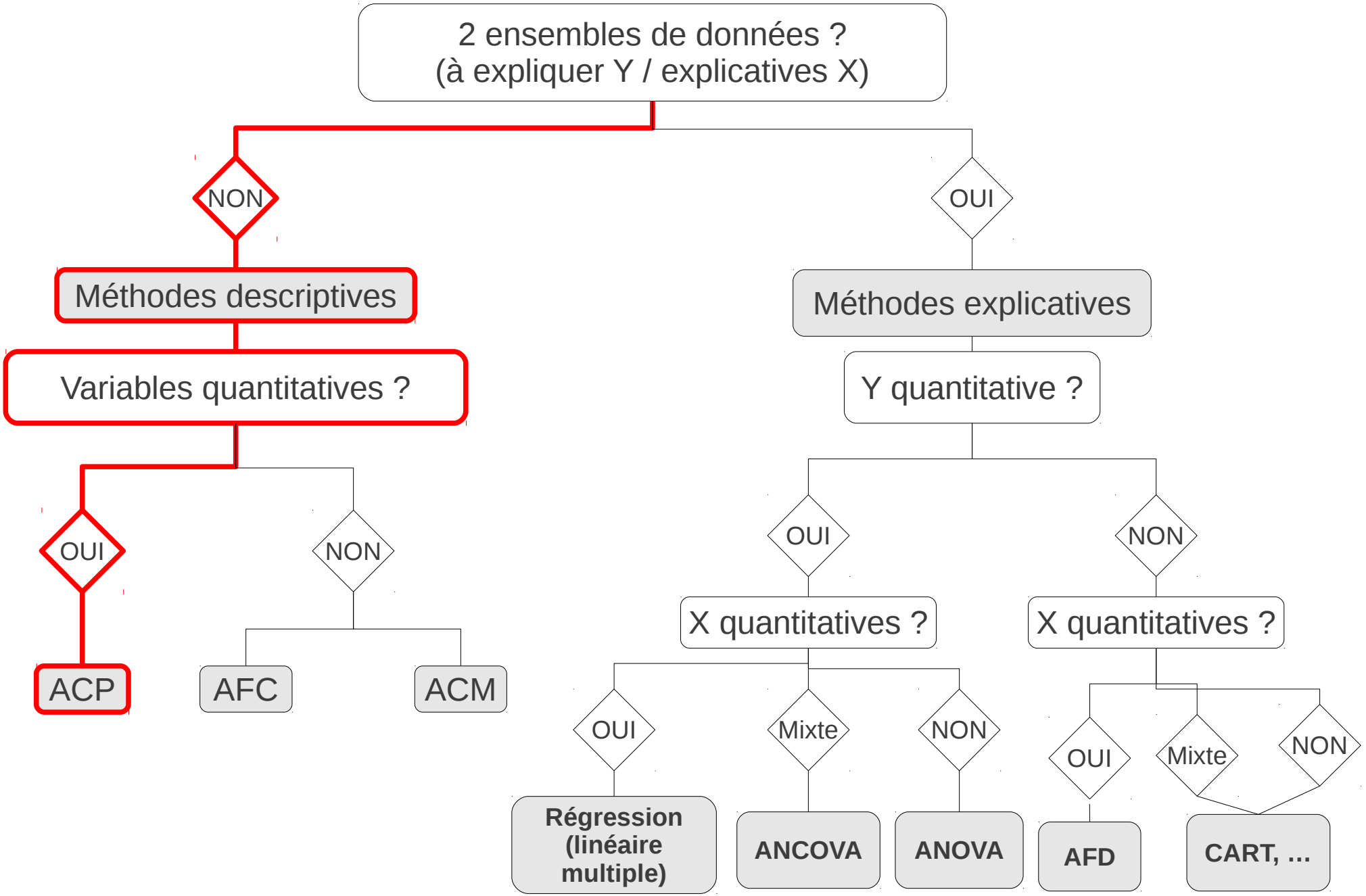
Expression de **868** gènes mesurée sur **22** échantillons :

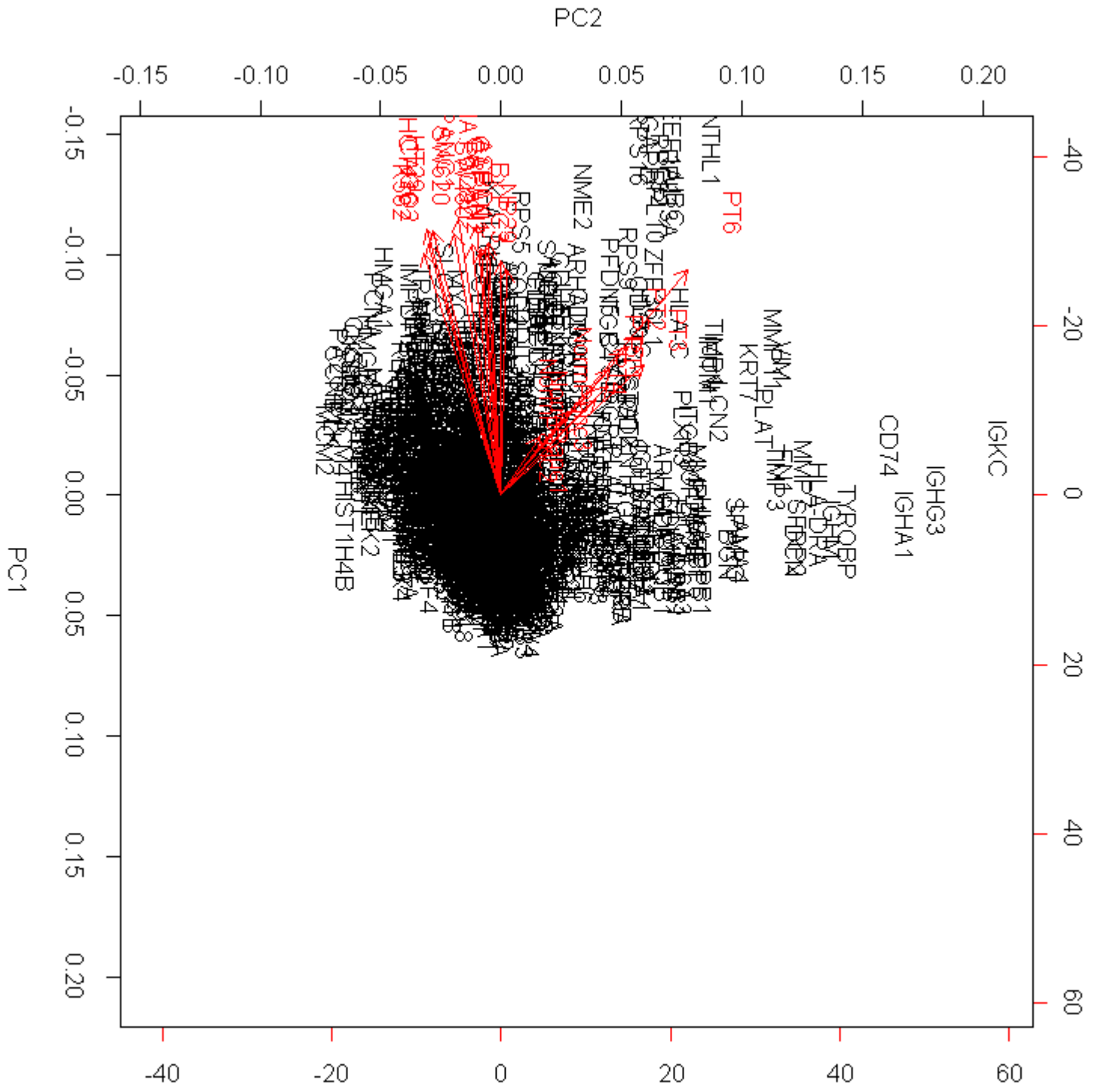
- Lignées pancréatiques (7 échantillons) : ASPC1, Bx-PC3, Capan 1, Capan 2, Mia-PaCa2, NP 29, Panc1 ;
- Lignées coliques (5 échantillons) : CaCo2, HCT116, HT29, SW480, SW620 ;
- Lignée leucémique (1 échantillon) : K562 ;
- Pièces tumorales (6 échantillons) : PT1, PT2, PT3, PT4, PT5, PT6 ;
- Pancréas normal (3 échantillons) : PancNorm1, PancNorm2, PancNorm3 ;

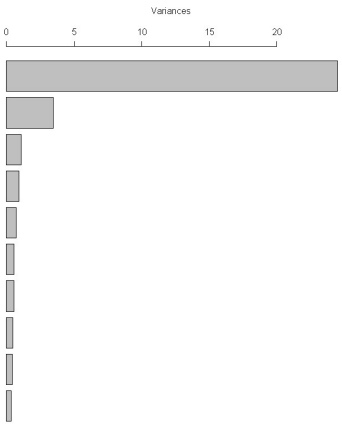
Extrait des données

	ASPC1	Bx-PC3	CAPAN1	CAPAN2	NP29	PANC1	MIA-PaCa2	PT1	PT2	PT3	PT4	PT5	PT6	CACO2	....
MAPRE1	1,838	1,736	1,523	2,062	1,353	2,488	2,319	-0,133	0,086	0,555	-0,036	0,238	1,279	2,551	
VIL2	1,458	1,687	1,429	0,788	0,605	0,736	2,243	0,02	0,745	0,25	-0,267	0,19	1,606	0,999	
NME2	3,82	4,452	4,966	4,719	4,031	4,912	5,252	2,958	3,167	3,11	2,743	2,327	3,641	4,141	
NME1	1,819	2,069	3,088	2,648	2,346	3,609	2,85	0,489	1,423	0,53	0,616	0,877	1,353	2,485	
MARK3	0,962	0,363	0,933	1,082	0,446	1,108	0,786	0,004	-0,045	-0,289	0,134	0,193	0,585	1,101	
JUN	2,157	1,417	0,887	-0,204	1,402	1,898	3,404	2,877	2,151	3,219	0,591	2,398	3,606	-0,054	
MYC	2,852	2,965	3,32	2,69	2,997	2,009	3,856	0,376	0,941	1,981	1,225	1,582	1,274	3,028	
FOSL1	2,342	1,996	2,233	1,345	1,963	3,229	3,36	-0,065	0,171	0,812	0,596	-0,774	-0,216	-1,167	
JUNB	-0,486	-0,046	-0,179	-0,649	-0,035	-0,757	-0,642	0,399	0,499	0,56	0,368	-0,3	1,208	-1,231	
AXL	0,741	1,194	-0,433	-0,513	0,326	1,353	1,358	0,018	1,122	0,358	0,501	0,362	1,281	-1,012	
ERBB3	2,733	2,499	2,727	2,35	2,503	1,29	3,142	1,555	0,928	2,503	0,619	0,443	1,877	2,449	
FLT1	2,023	2,674	3,294	3,043	2,686	3,287	3,762	2,178	1,402	2,282	0,679	0,94	2,677	2,436	
...															

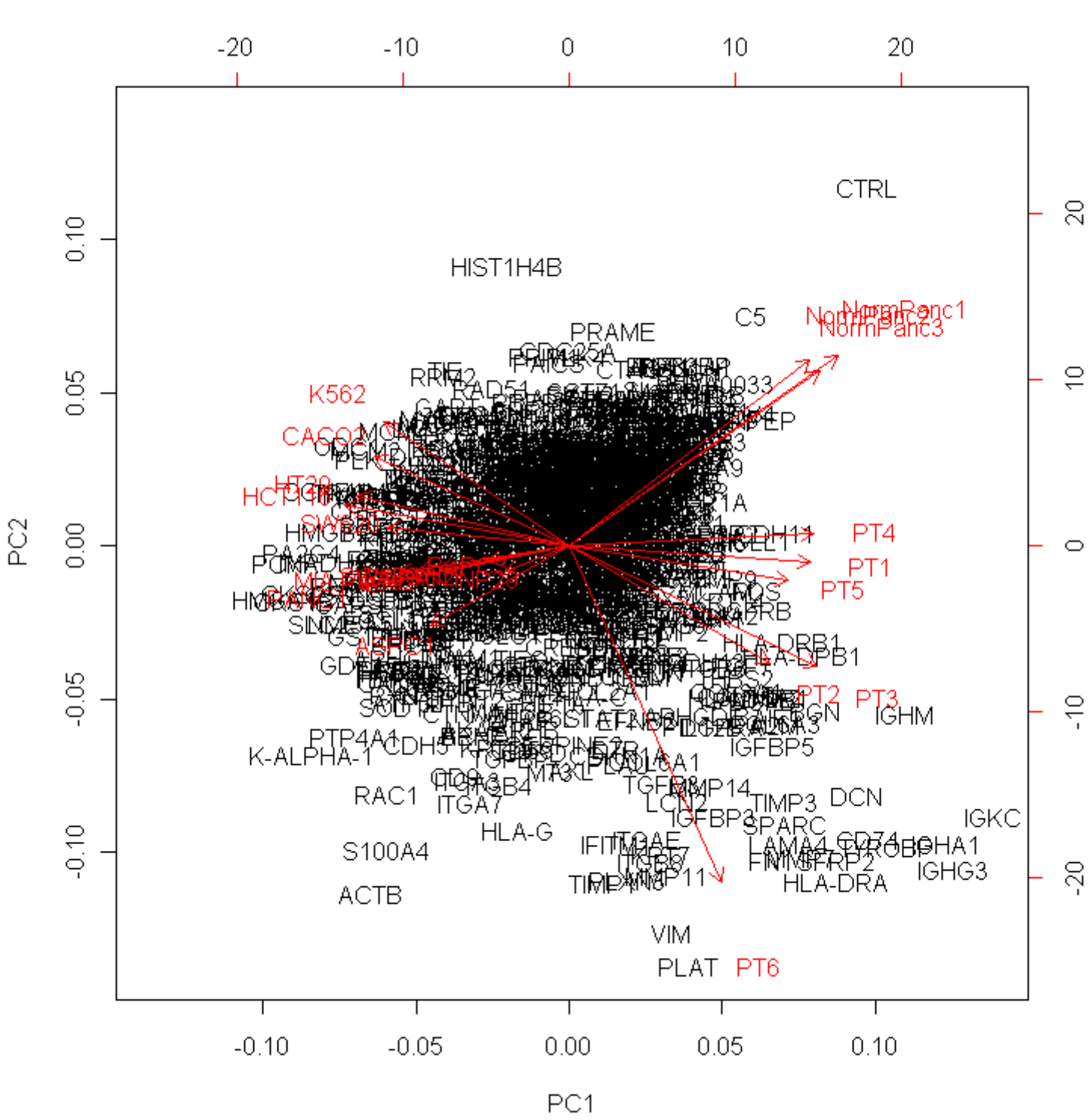
Analyse multidimensionnelle

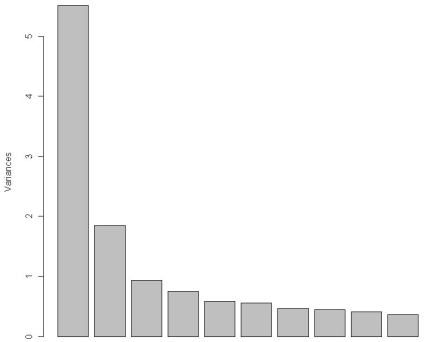


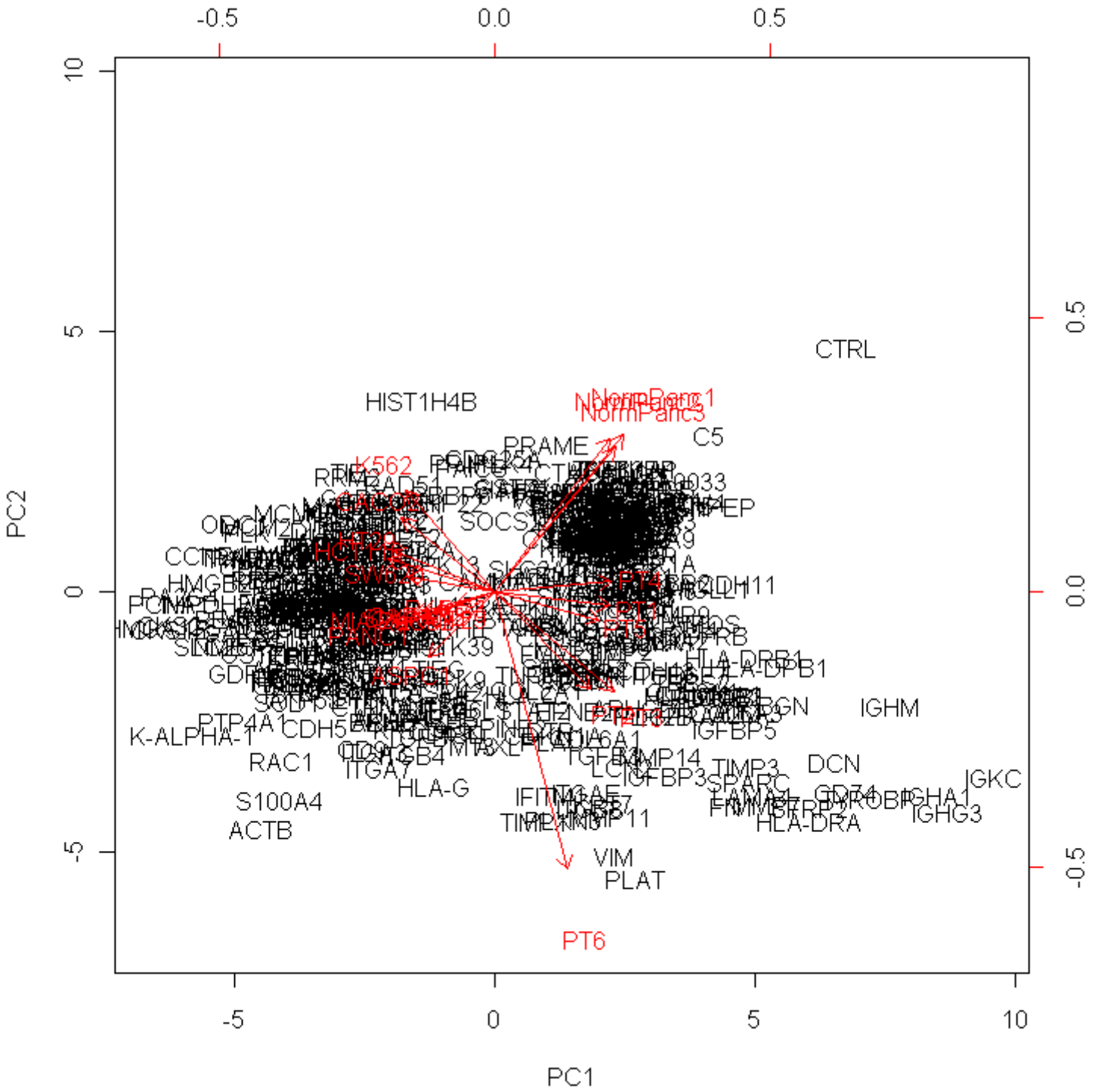

PC2
PC1
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
-40
-20
0
20
40
60
PT6
IGKC
IGHG3
IGHA1
CD74
NME2
NTHL1
KRT7
PLAT
TYROBP

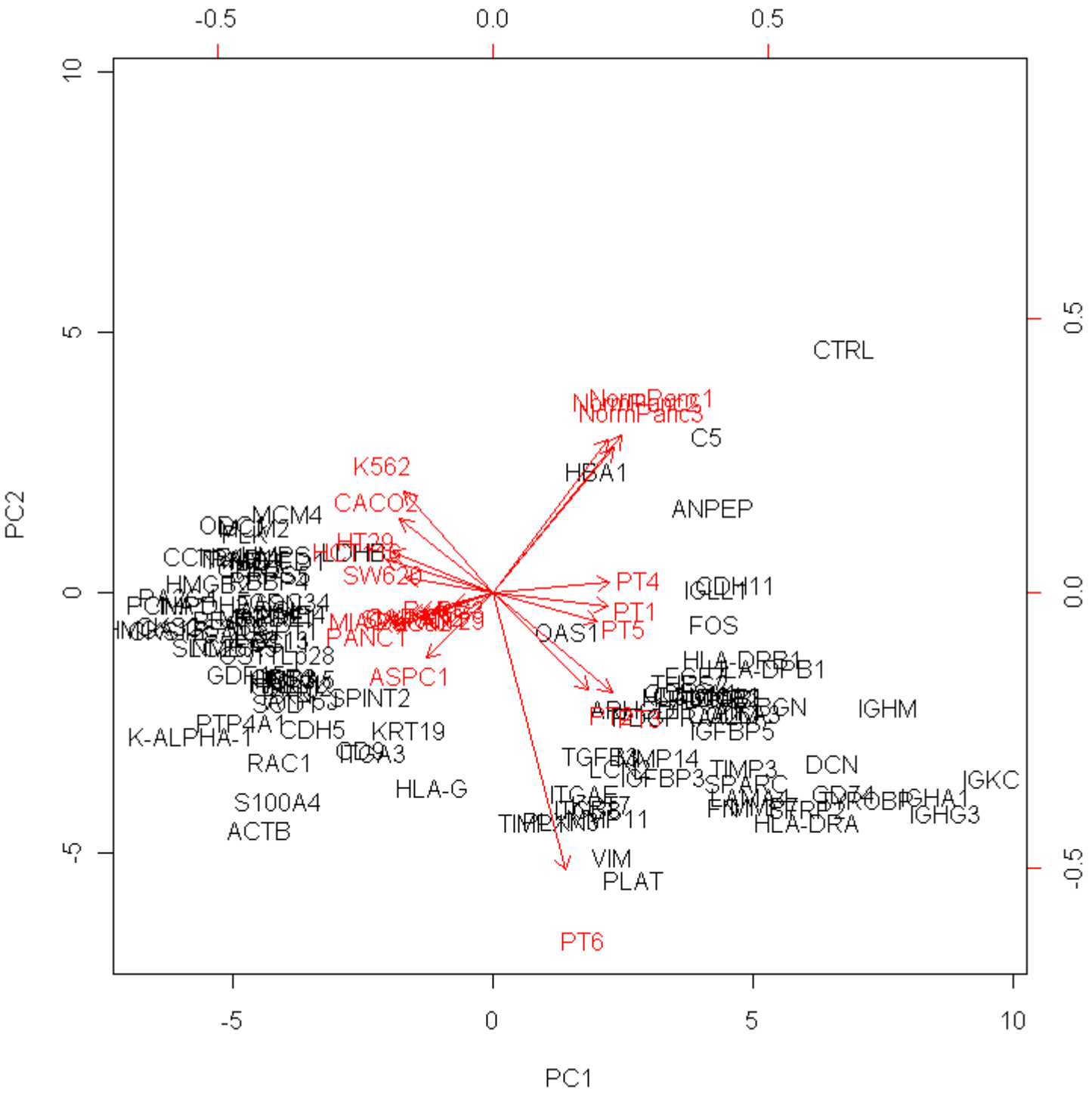

Variances
0
5
10
15
20

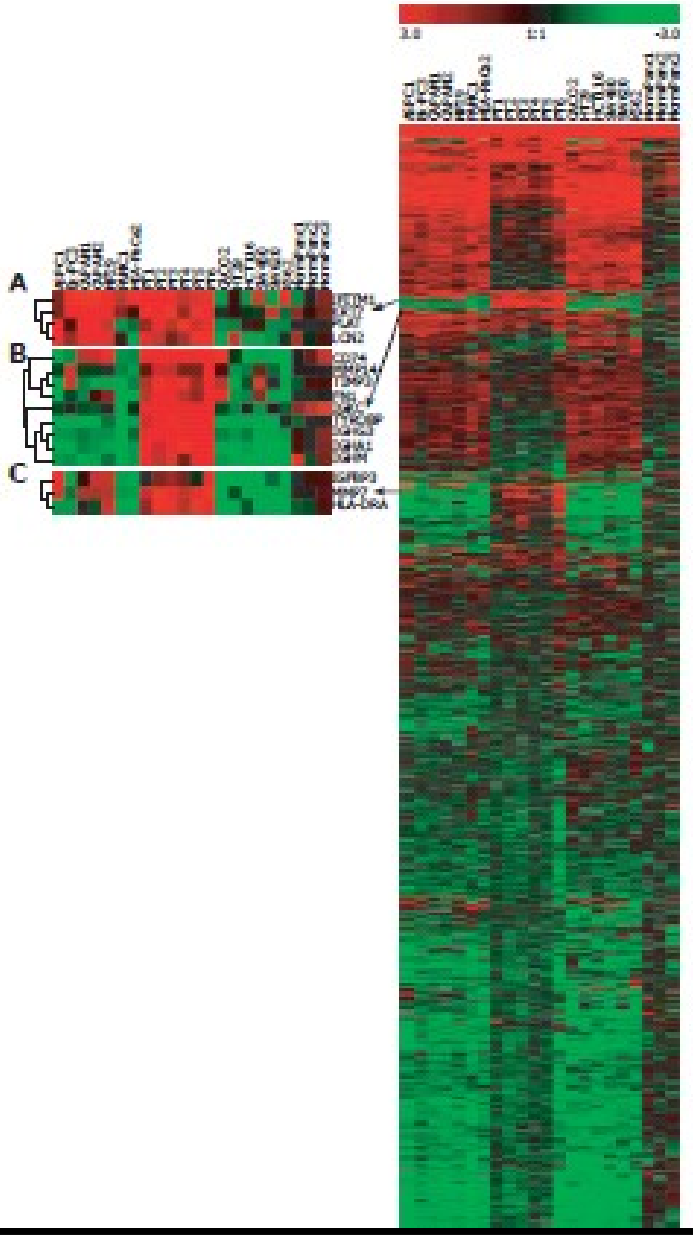
ACP ?
ACP
!











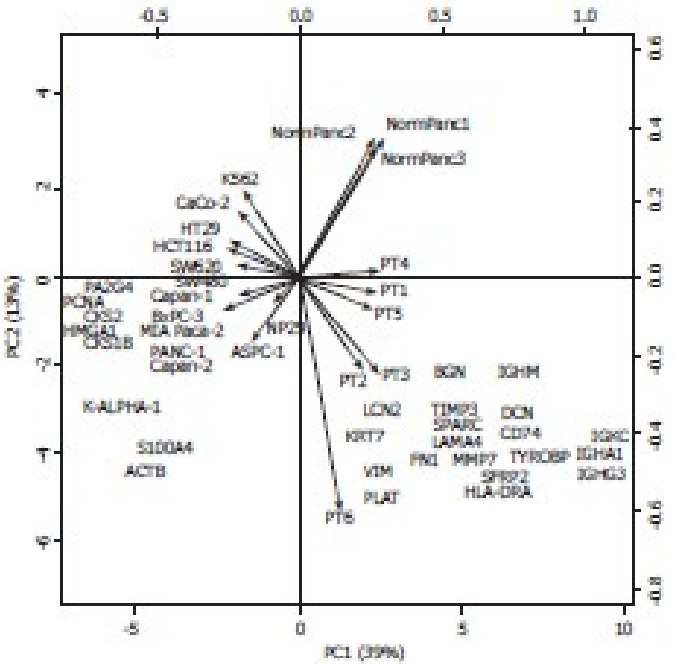
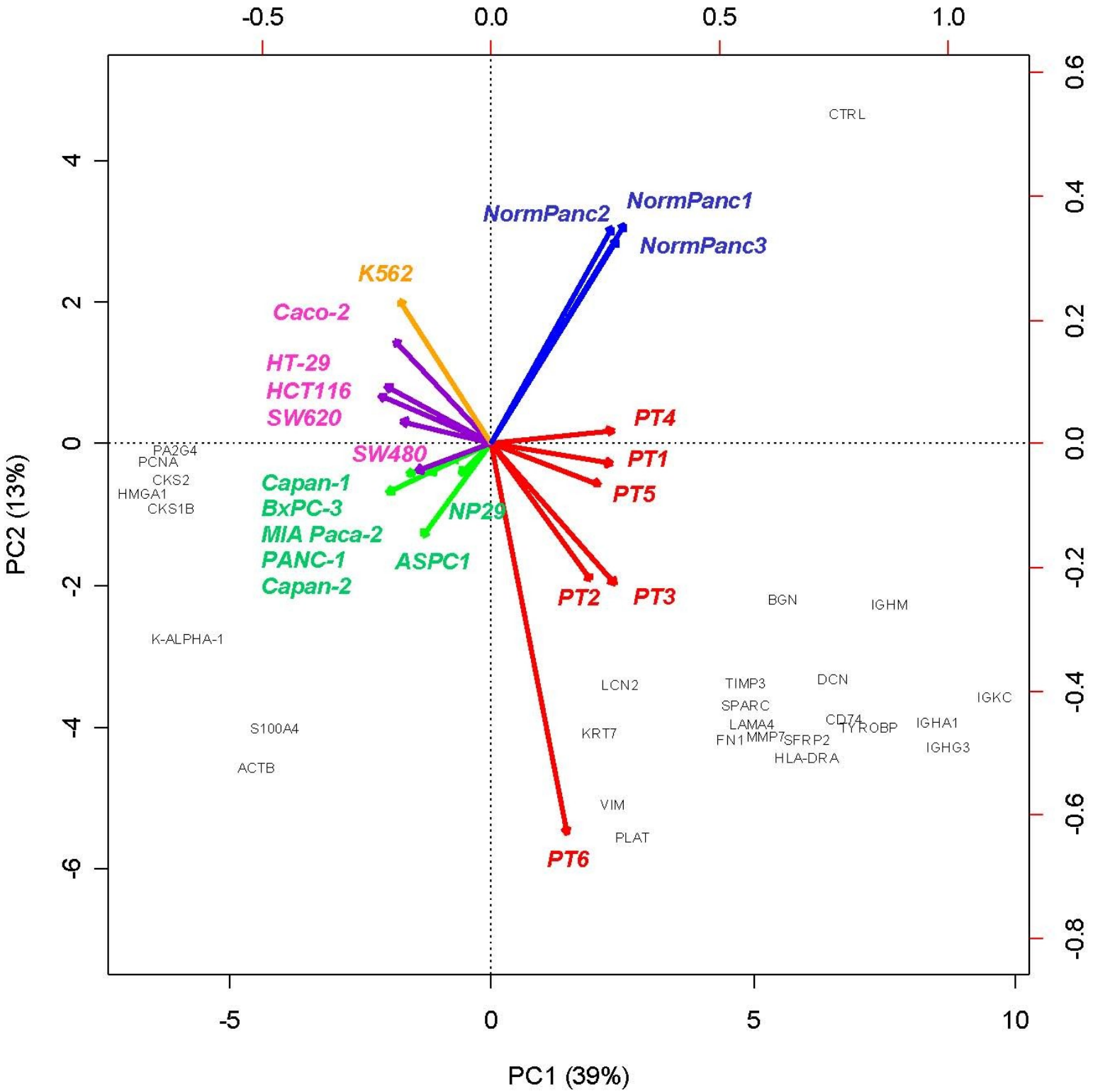


Figure 3 Principal component analysis (PCA) of gene expression data. Biplot resulting from a PCA of the line and column centered data containing 871 genes (individuals) in lines and 22 samples in columns (variables). The 28 genes contributing the most to the total variability are shown. The two principal components (PC1 and PC2) contribute to more than 50% (39% and 13%) of the total variability and resolve four biological sample categories: PC1 on the horizontal x-axis distinguish between cell lines (left) and tissue samples (right) whereas PC2 on the vertical y-axis distinguish between malignant pancreas samples (bottom) and other sample categories (top).

experimental protocols and data acquisition procedures (data not shown). We used the SAM software to detect the genes that were significantly over-expressed in malignant pancreas, referring to the mean expression in pancreatic tumors (PT) and pancreatic cancer cell lines (PCL) as compared with the other groups normal pancreas

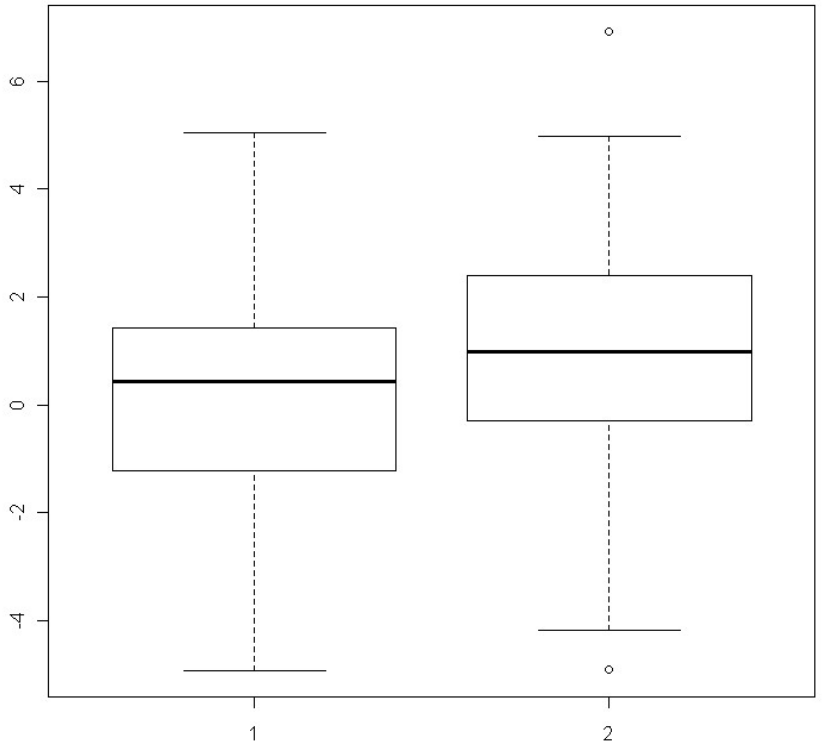


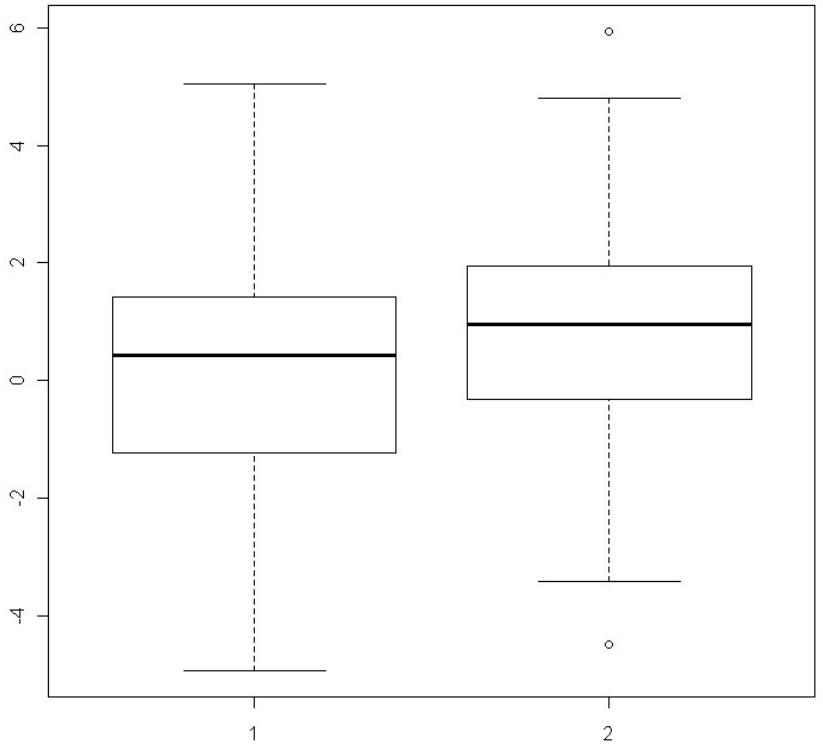
Exemple 2

Tests statistiques

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Test de comparaison de moyennes





`Two Sample t-test : p-value =` **`0.08284`**

`Two Sample t-test : p-value =` **`0.03556`**

P-value > 5%

P-value < 5%

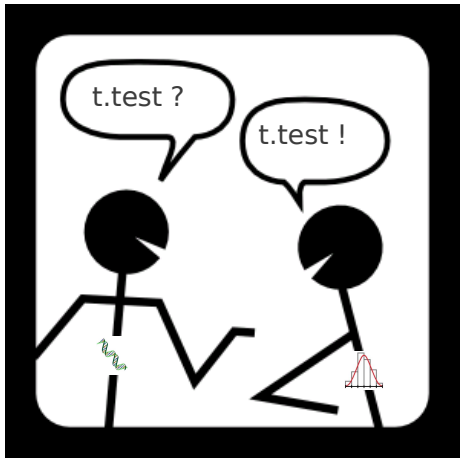
*

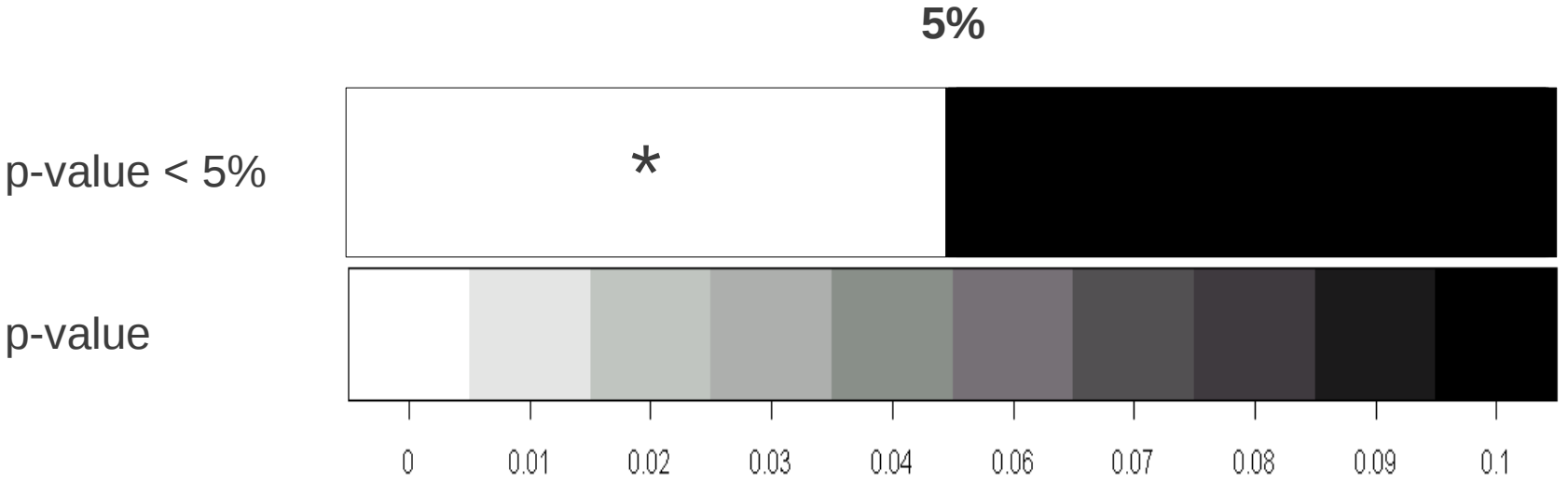
P-value < 10%

*

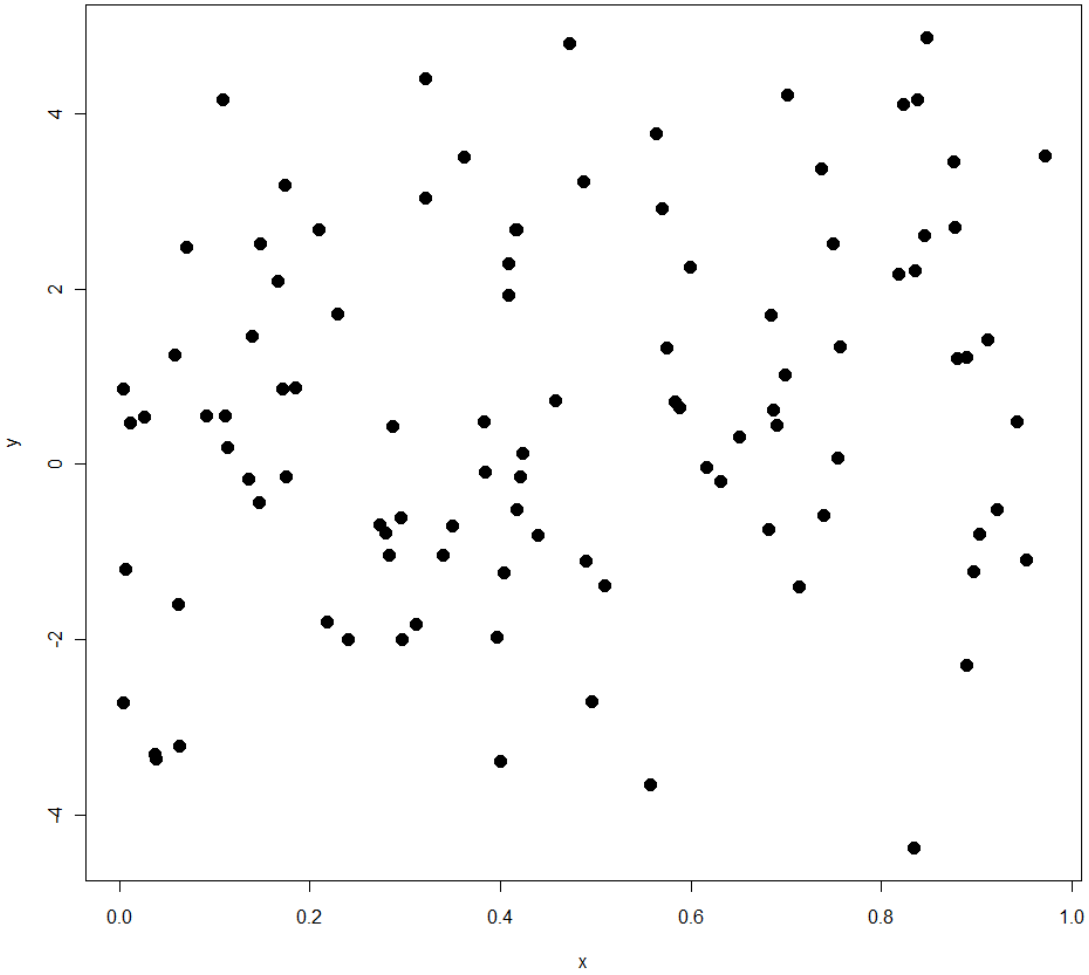
Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Test de comparaison de moyennes

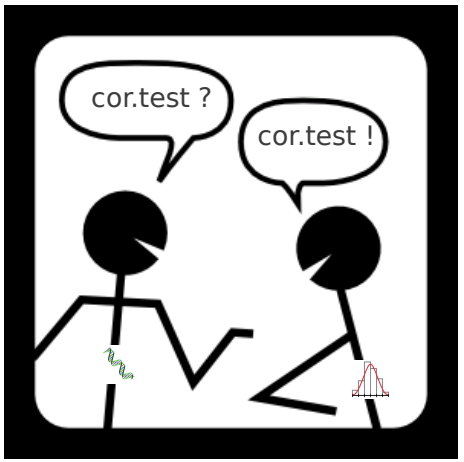




Test sur le coefficient de corrélation ◀ ◀ ◀

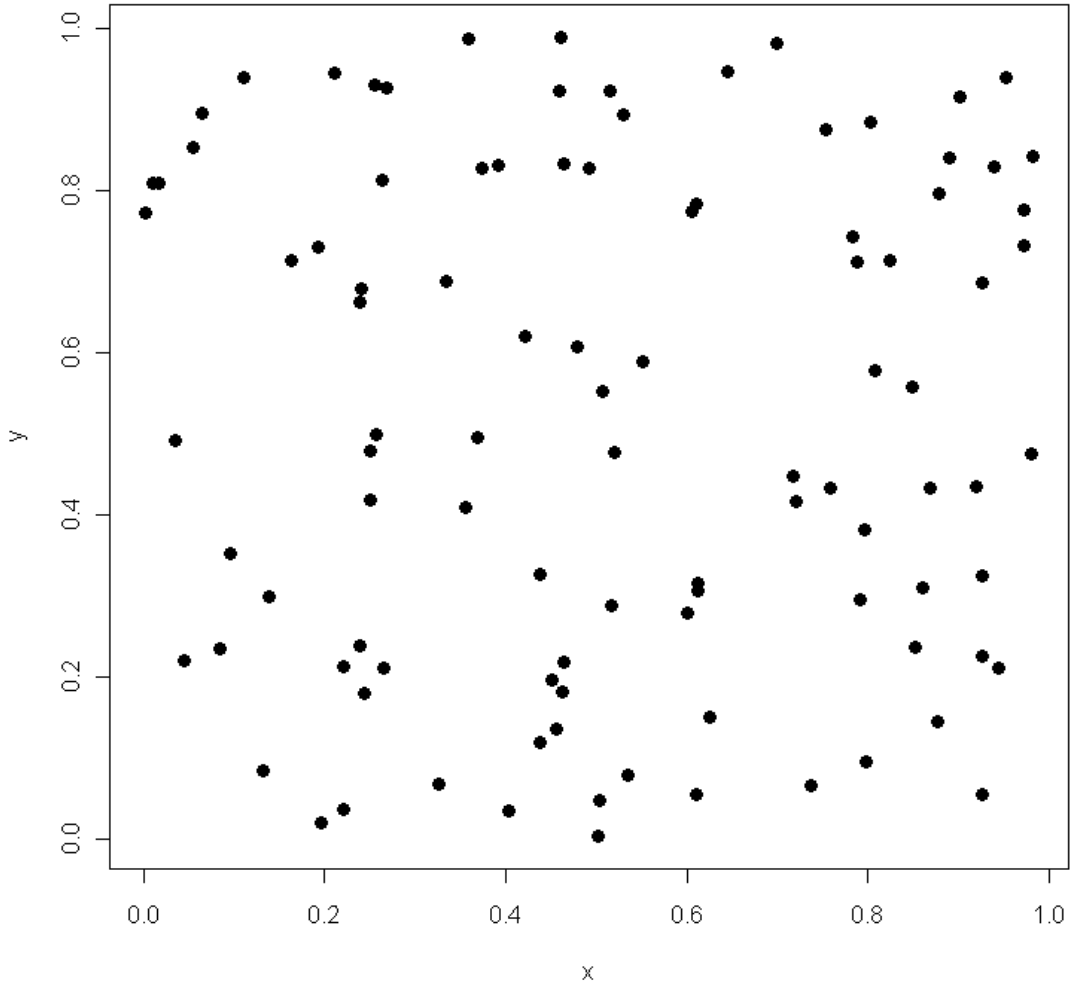



```
> cor.test(x,z)

 Pearson's product-moment correlation

data:  x and z
t = 2.0884, df = 98, p-value = 0.03935
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 0.01042556 0.38714248
sample estimates:
      cor
0.206421
```



Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Test sur le coefficient de corrélation




```
> cor.test(x,y)

 Pearson's product-moment correlation

data:  x and y
t = 0.0303, df = 98, p-value = 0.976
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 -0.1934772  0.1993555
sample estimates:
        cor
0.003057046
```

D'autres points sur les tests ◀ ◀ ◀

- Risque de 2ème espèce
- Puissance
- Calcul d'effectifs
- Gestion de la multiplicité

		Décision	
		H1 (rejet de H0)	H0 (accept. H0)
Réalité	H0	α	Bonne décision
	H1	Bonne décision	β

- 1 comparaison : α = P[Rejeter H0 // H0 vraie] = 5%

  ➔ Probabilité de ne pas commettre d'erreur : 1-0.05 = 0.95

- 3 comparaisons :

$$\begin{cases} \alpha_1 = P[\text{Rejeter } H0_1 \text{ // } H0_1 \text{ vraie}] = 5\% \\ \alpha_2 = P[\text{Rejeter } H0_2 \text{ // } H0_2 \text{ vraie}] = 5\% \\ \alpha_3 = P[\text{Rejeter } H0_3 \text{ // } H0_3 \text{ vraie}] = 5\% \end{cases}$$

➔ Probabilité de ne pas commettre d'erreur = produit des probabilités de ne pas commettre d'erreur à chacune des 3 comparaisons = (1-0.05) * (1-0.05) * (1-0.05) = 0.86

Le risque (global) de commettre au moins une erreur est :

1-0.86 = **0.14**

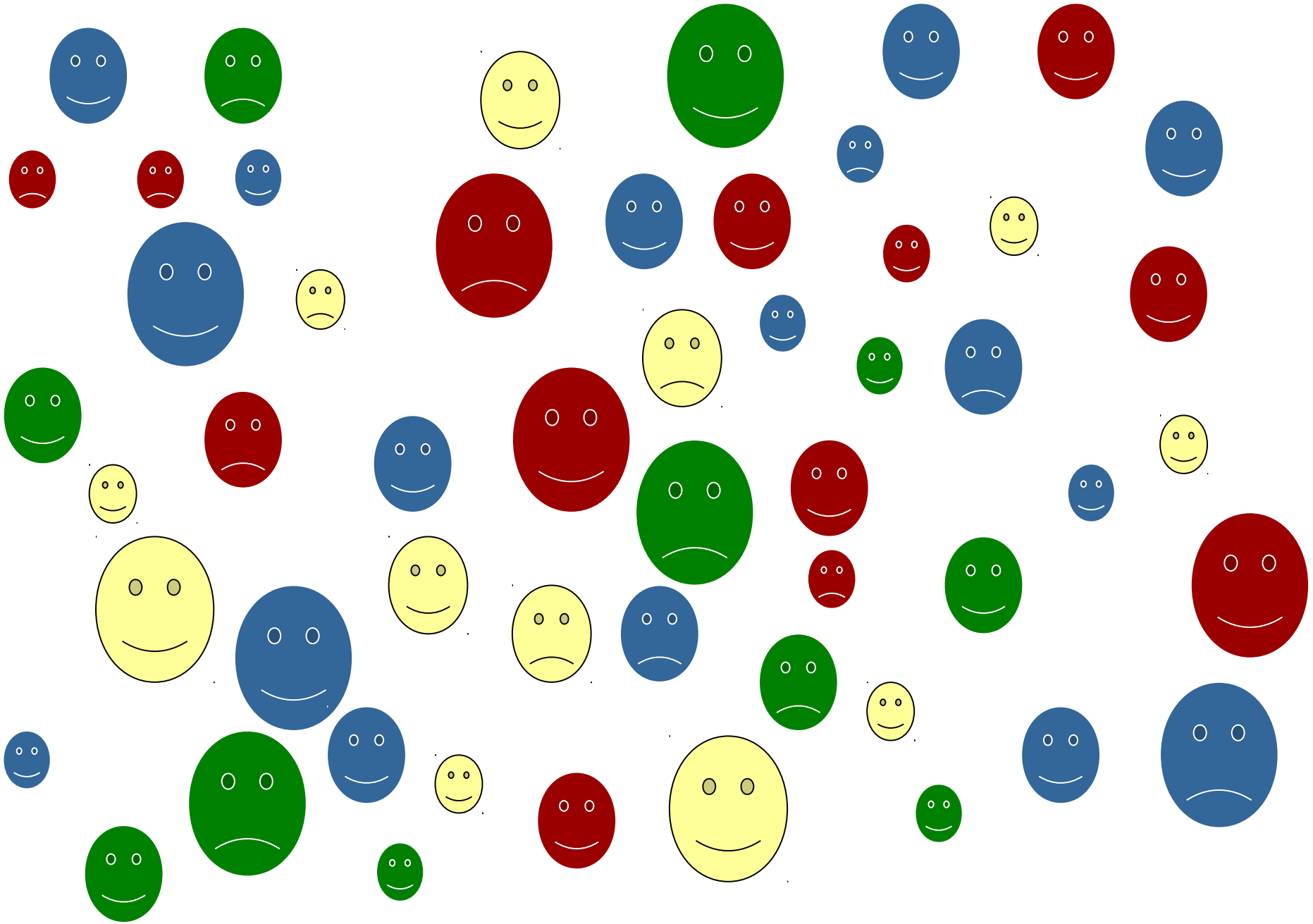
Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Exemple 3

Classification

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

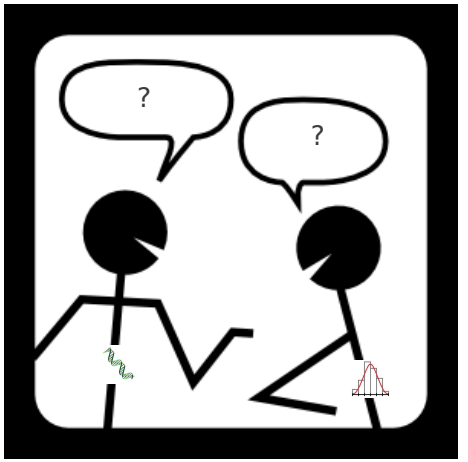
Regouper des objets qui se ressemblent

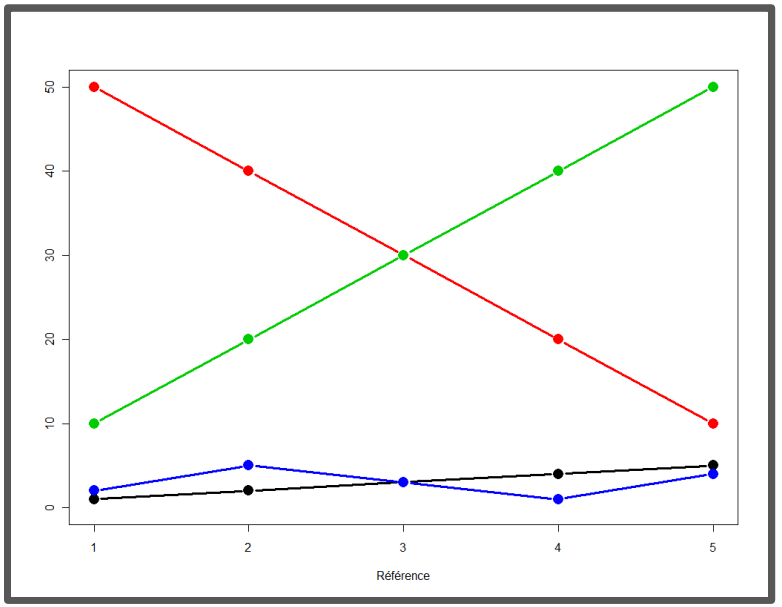


Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Quelle distance choisir ?

- Distance euclidienne
- 1 – corrélation
- 1 – corrélation^2





Référence	1	2	3	4	5
Individu 1	50	40	30	20	10
Individu 2	10	20	30	40	50
Individu 3	2	5	3	1	4

Exemple 4

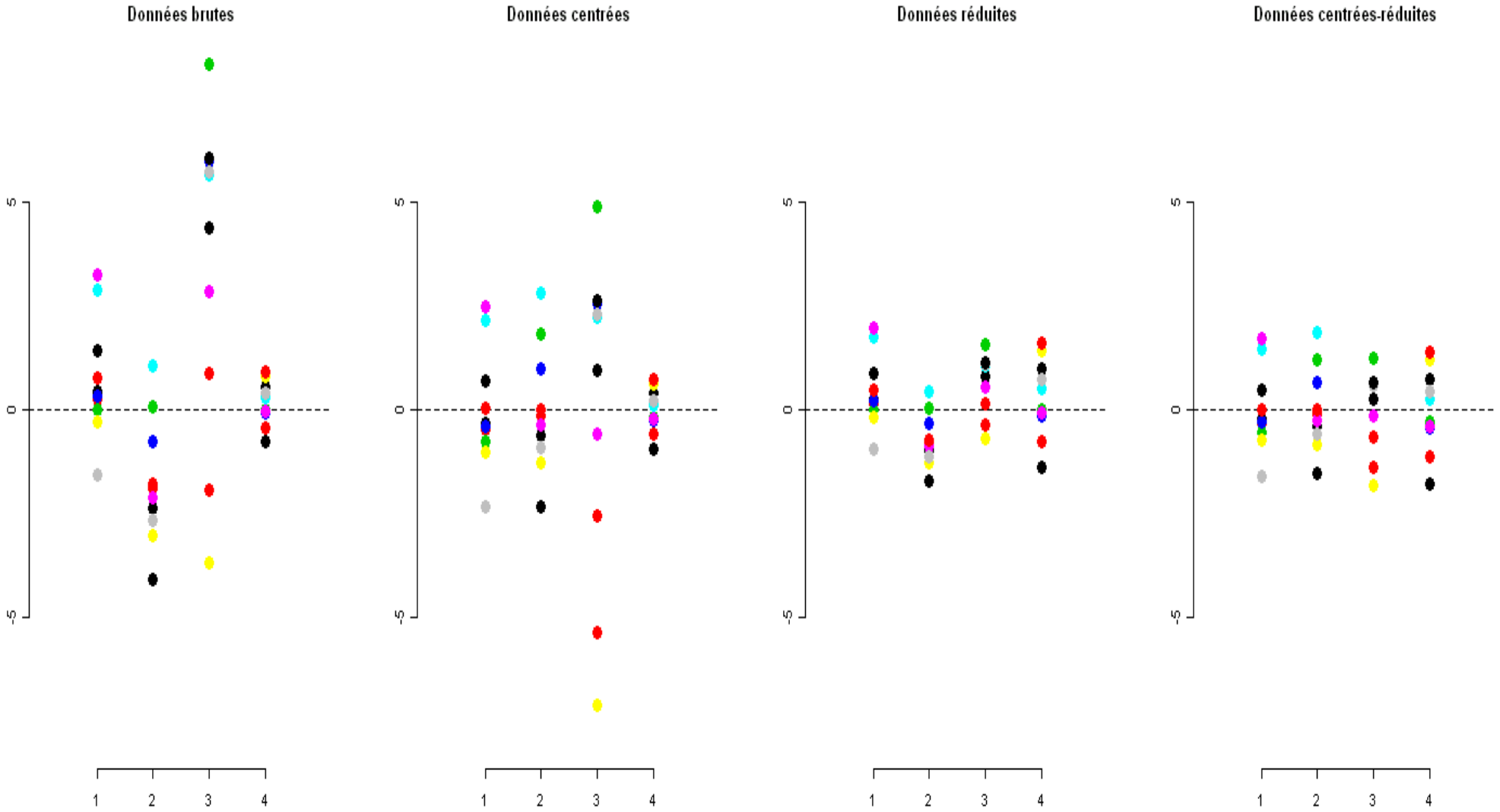
Transformation de données

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Centrage - Réduction

- **Centrer** : retrancher la moyenne
- **Réduire** : diviser par l'écart-type

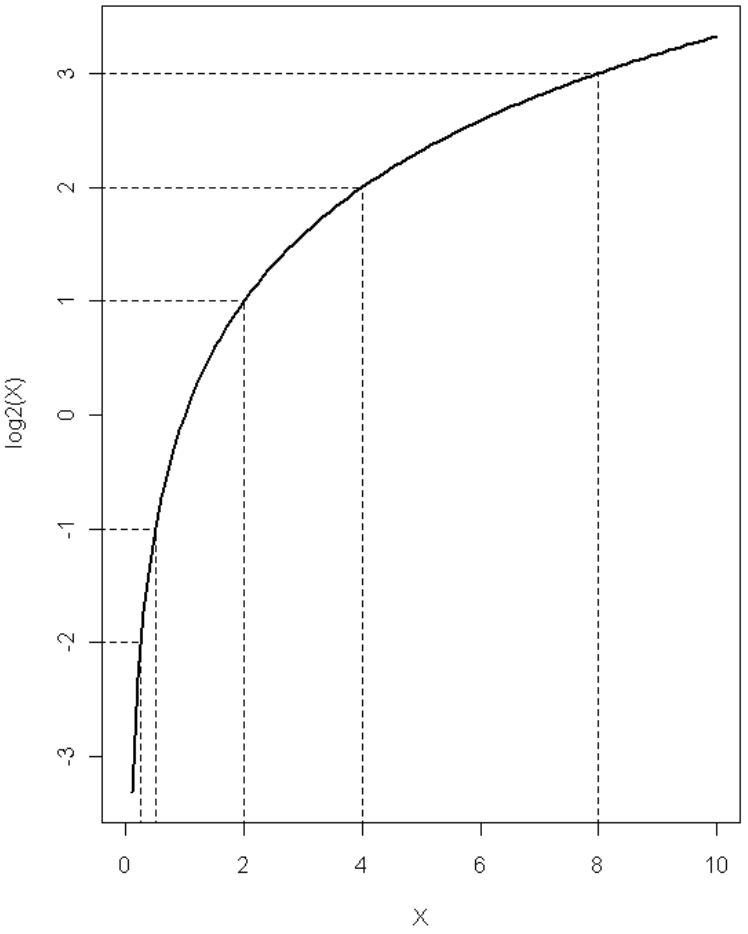
$$Z_i = \frac{X_i - \overline{X}}{\sigma_X}$$



Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Transformation log

X	0,125 $= 2^{-3}$	0,25 $= 2^{-2}$	0,5 $= 2^{-1}$	1 $= 2^{0}$	2 $= 2^{1}$	4 $= 2^{2}$	8 $= 2^{3}$
$\log_2(X)$	-3	-2	-1	0	1	2	3



- Utile pour la conversion de ratio
- Exemple : le ``double'' et la ``moitié'' sont rendues symétriques par rapport à 0
- Pour les p-values, on aura plus intérêt à utiliser $\log_{10}$.
- La fonction réciproque de « log » est la fonction puissance :

$$Y = \log_2(X) \leftrightarrow X = 2^Y$$

$$Y = \log_{10}(X) \leftrightarrow X = 10^Y$$

$$Y = \ln(X) \leftrightarrow X = e^Y = \exp(Y)$$

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Exemple 5

Planification expérimentale

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

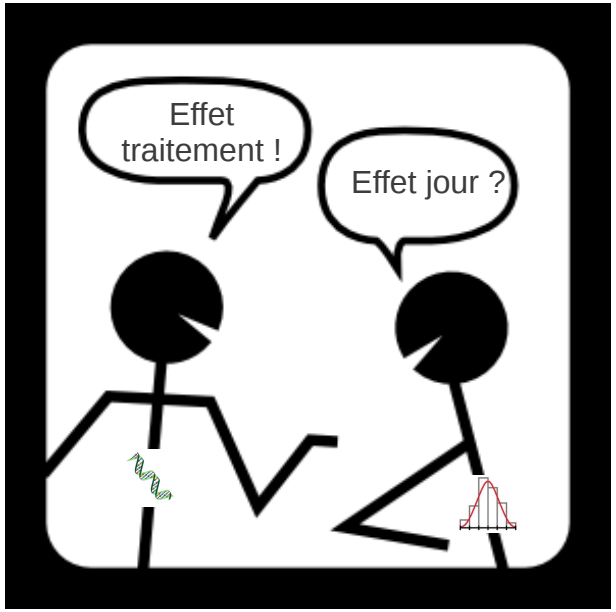
Confusion des effets

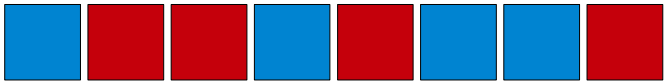
2 conditions à l'étude : Contrôle / Traitement

Jour 1  8 échantillons Contrôle

Jour 2  8 échantillons Traitement

Test statistique : les moyennes des 2 séries sont significativement différentes ! *

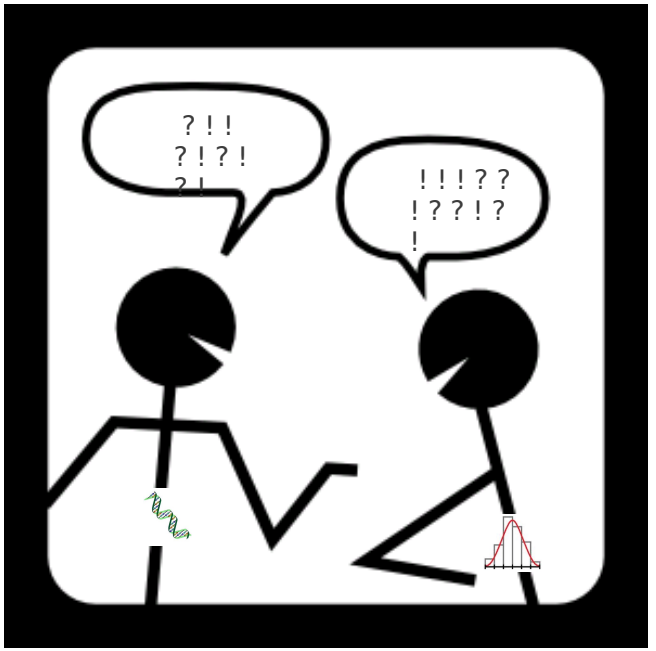


Jour 1 

Jour 2 

Randomisation

Sébastien Déjean - math.univ-toulouse.fr/~sdejean



Interaction

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

1) énoncer clairement une question précise
2) prévoir les méthodes d'analyse des données
3) mettre en place un plan d'expérience

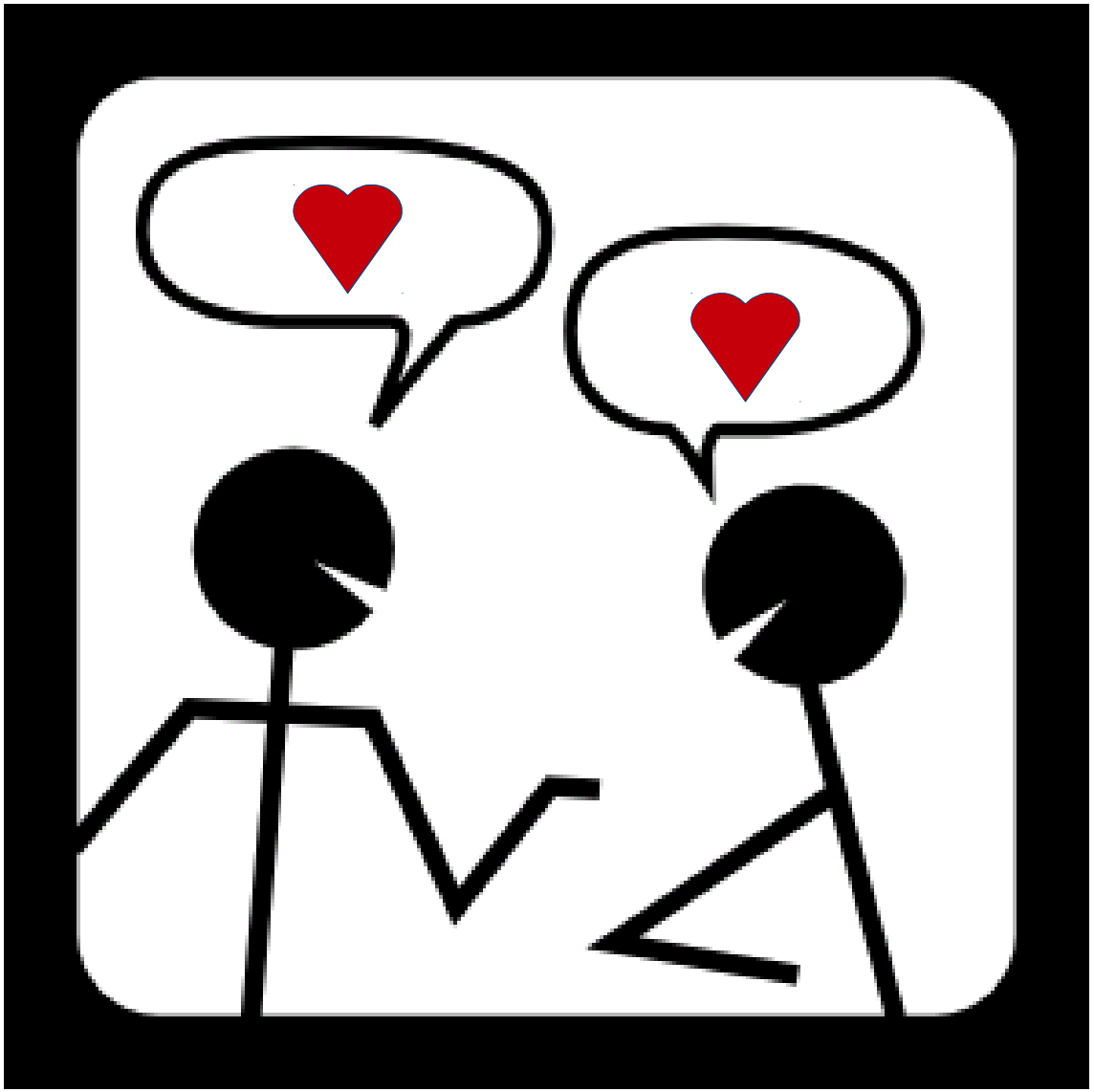
4) acquérir les données

5) analyser des données
6) interpréter des résultats
7) répondre à la question posée

Avant d'être exécutés, on accorde une dernière volonté à un statisticien et à un biologiste. Le statisticien demande l'autorisation de donner une dernière conférence sur sa Grande Théorie des Statistiques. Le biologiste demande à être exécuté en premier (...avant la conférence...)

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

R.A. Fisher



The End

Sébastien Déjean - math.univ-toulouse.fr/~sdejean