



Génomique

Guillaume LAMIRAULT

guillaume.lamirault@univ-nantes.fr

UE M1S&S-Thorax, 10 octobre 2022, Nantes

Vue d'ensemble du génome humain (2004)

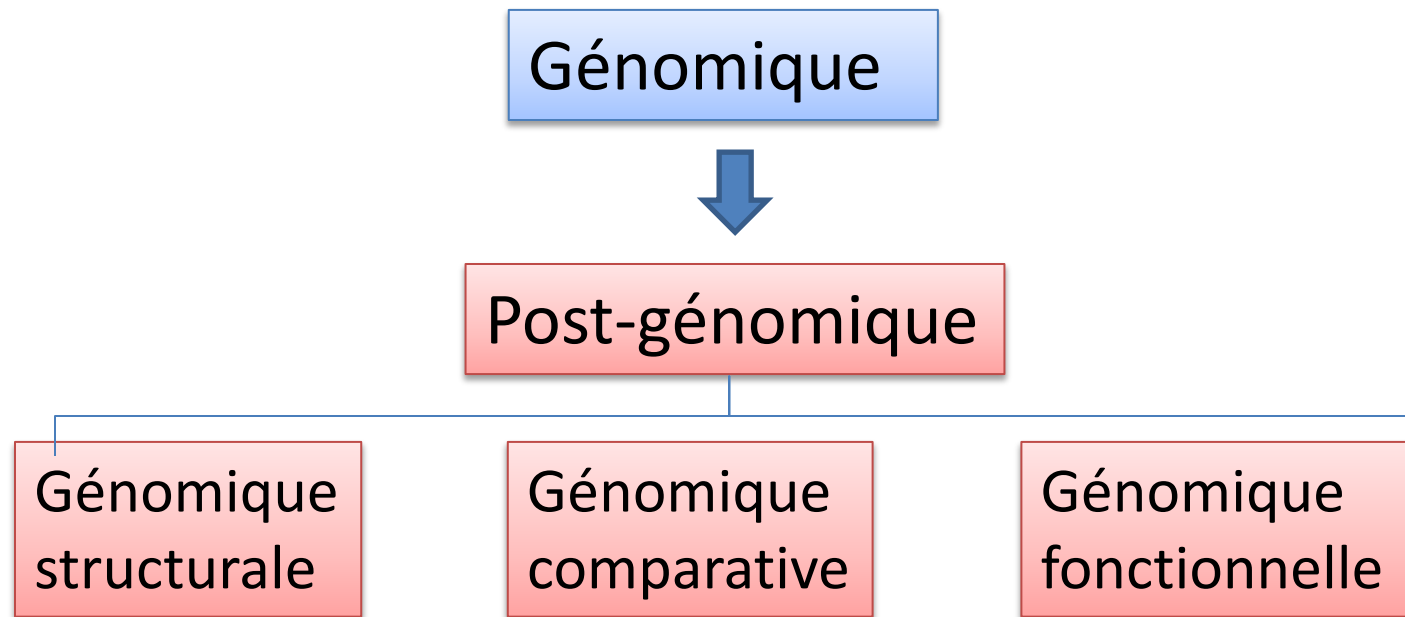
Taille du génome	3.08 Gbases
Taux d'erreur dans la séquence	1/100000 bases
Nombre de gènes identifiés	19438
Nombre de gènes (prédits et identifiés)	22287
Nombre moyen d'exon par gène	10,4
Régions traduites en protéines	34 Mbases (1,1%)
Régions transcrites non traduites	21 Mbases (0.7%)

Génomique aujourd'hui

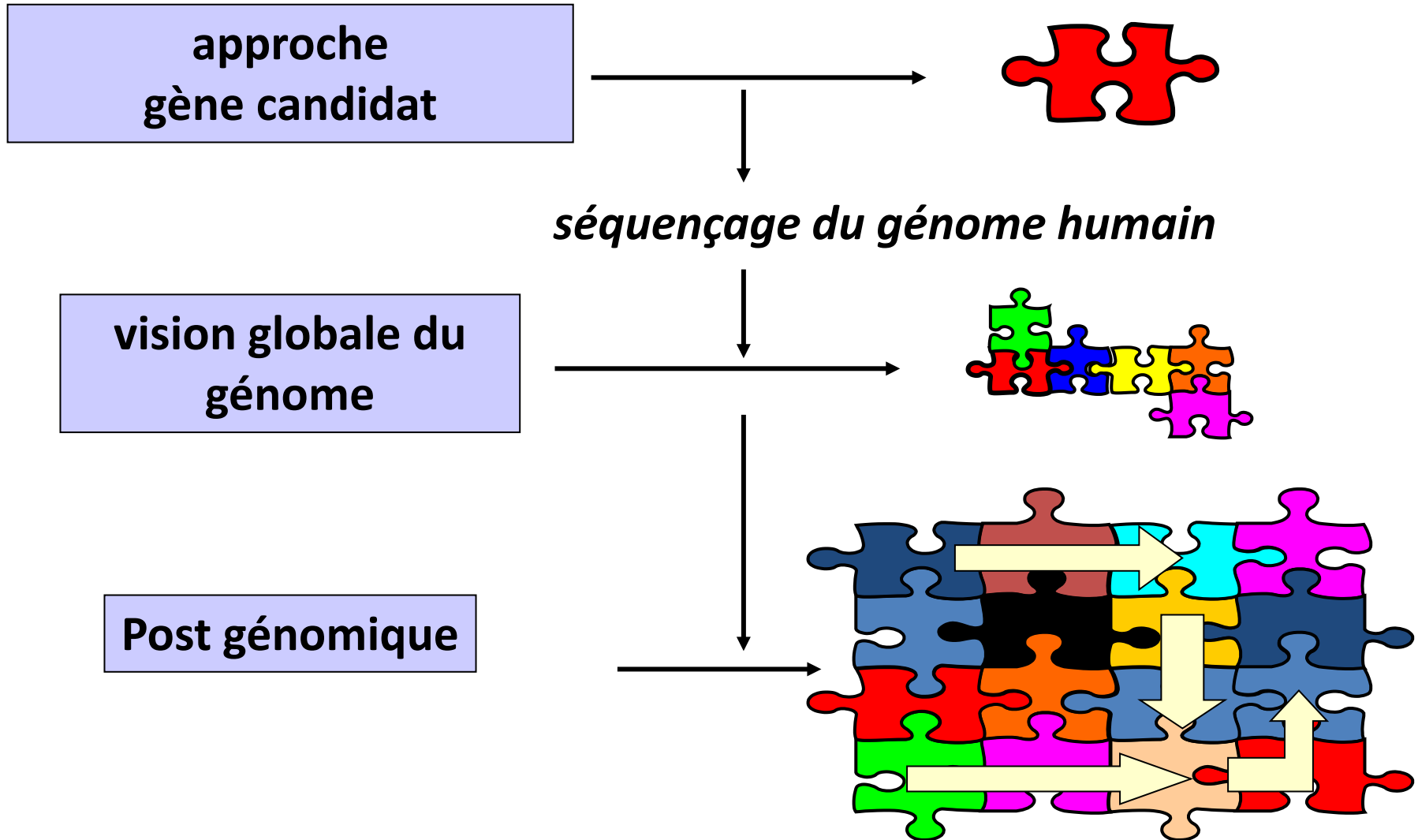
Définition initiale: Discipline scientifique dont l'objectif est de cartographier, séquencer et analyser les génomes

- Le séquençage à haut-débit des génomes est devenu une activité de routine.
- Développement de nouvelles disciplines scientifiques (transcriptomique, protéomique, bioinformatique, pharmacogénomique,...)
- Pour 40-60% des gènes identifiés, la fonction reste inconnue

La post-génomique



- Analyse globale du génome permettant d'assigner des fonctions aux gènes
- Définir l'organisation et le contrôle des processus génétiques qui s'associent pour faire le fonctionnement physiologique ou pathologique d'une cellule, d'un organe ou d'un organisme



Tous les gènes sont des candidats potentiels !

Génomique structurale

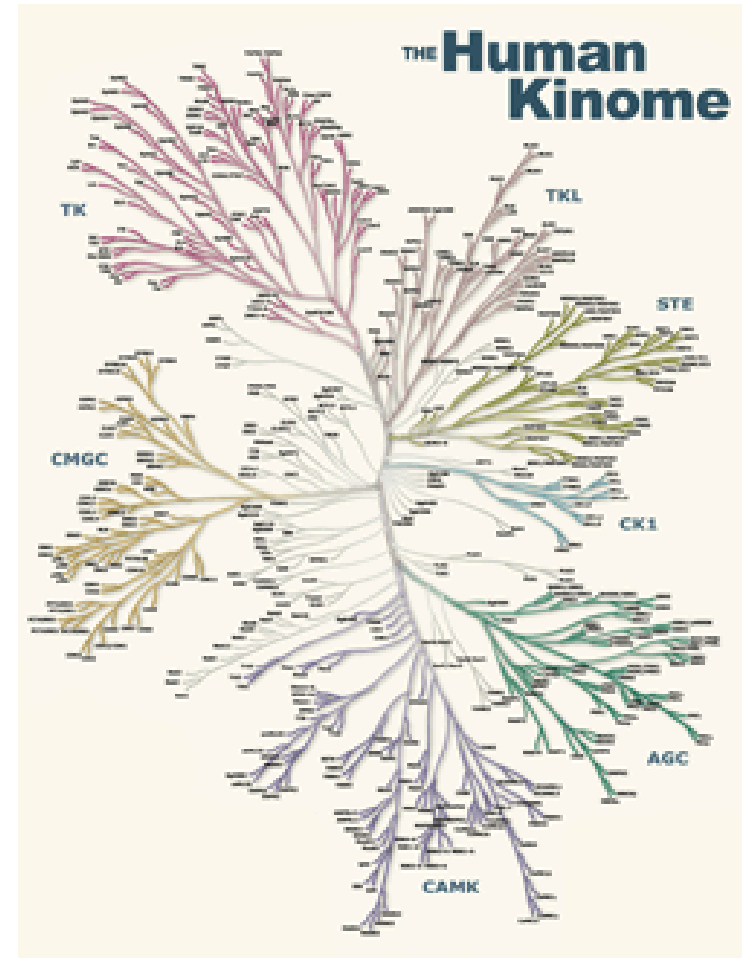
- Objectif: détermination de la structure tridimensionnelle des protéines à partir des données du génome
- Etudier la fonction des milliers de gènes « inconnus » à partir de la relation structure fonction des protéines

« Human kinome »

- = Ensemble des protéines kinase exprimées chez l'Homme

1- Définition d'une séquence consensus de site catalytique à partir des kinase connues

2- Identification de nouveaux gènes codant potentiellement des protéines kinases à partir de leur séquences

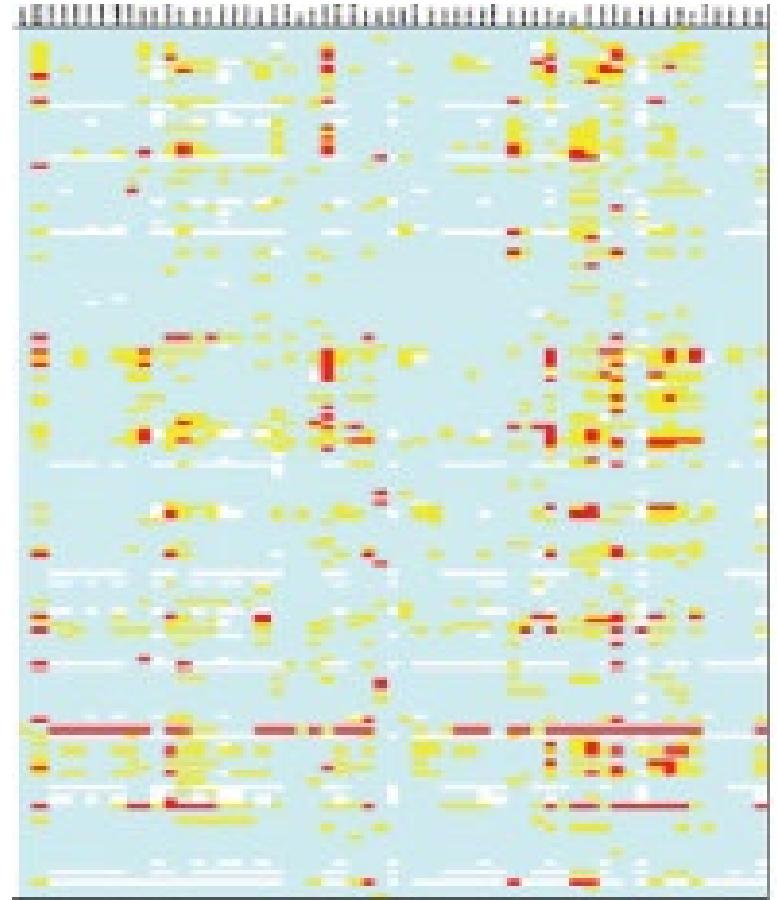


« human kinome »

57 Ser/Thr Kinases

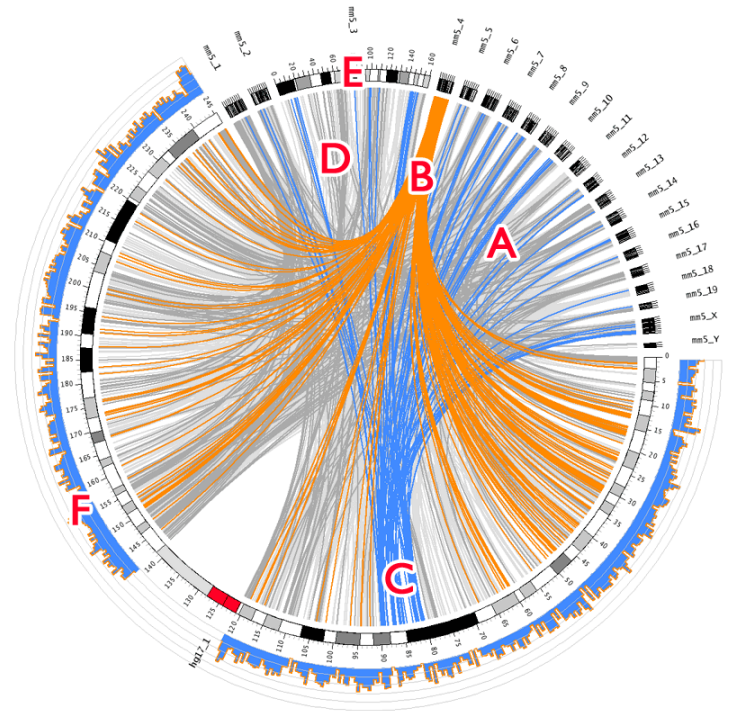
- Production des protéines à partir des séquences
- Test d'inhibiteurs connus des kinases
- Identification de cibles moléculaires

173 inhibiteurs



Génomique comparative

- Analyse et comparaison de l'ADN de différents organismes
- Conservation de l'ADN entre espèces en particulier au niveau des exons
Similarité Homme - singes 95%
Similarité Homme - souris 85%
- ~99% des gènes humains ont un équivalent (orthologue) chez la souris



Génomique comparative

- Identification des séquences conservées entre différents organismes
 - ☐ Séquences codantes
- Identification de la fonction d'un gène à partir de la fonction des orthologues

	226	233	236	238	239	242	247
humain	L I L T L S L I L V V I L V L L T V L A L L						
souris	L I L T L S L I L V L I S L L L T V L A L L						
rat	L I L T L S L I L V L I S L L L T V L A L L						
mouton	L I L T L S L V L V L I L L L L A V L A L L						
bétail	L I L T L S L V L V L I L L L L A V L A L L						
cochon	L I L T L S L I L V L I L L L L A V L A L L						

Génomique Fonctionnelle

Génomique

=

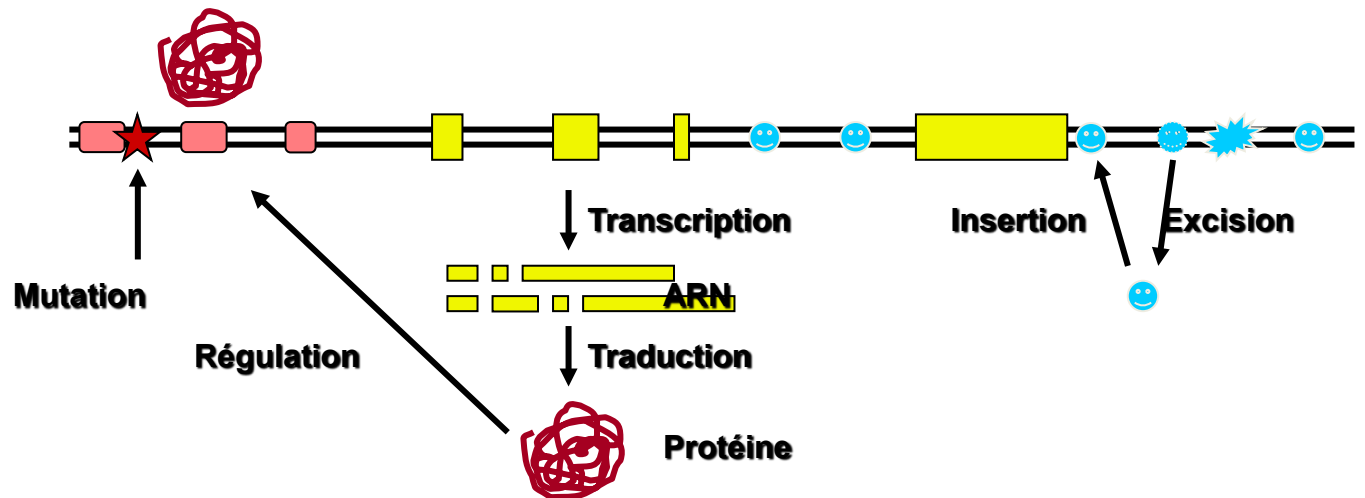
Description
physique



La Génomique
Fonctionnelle

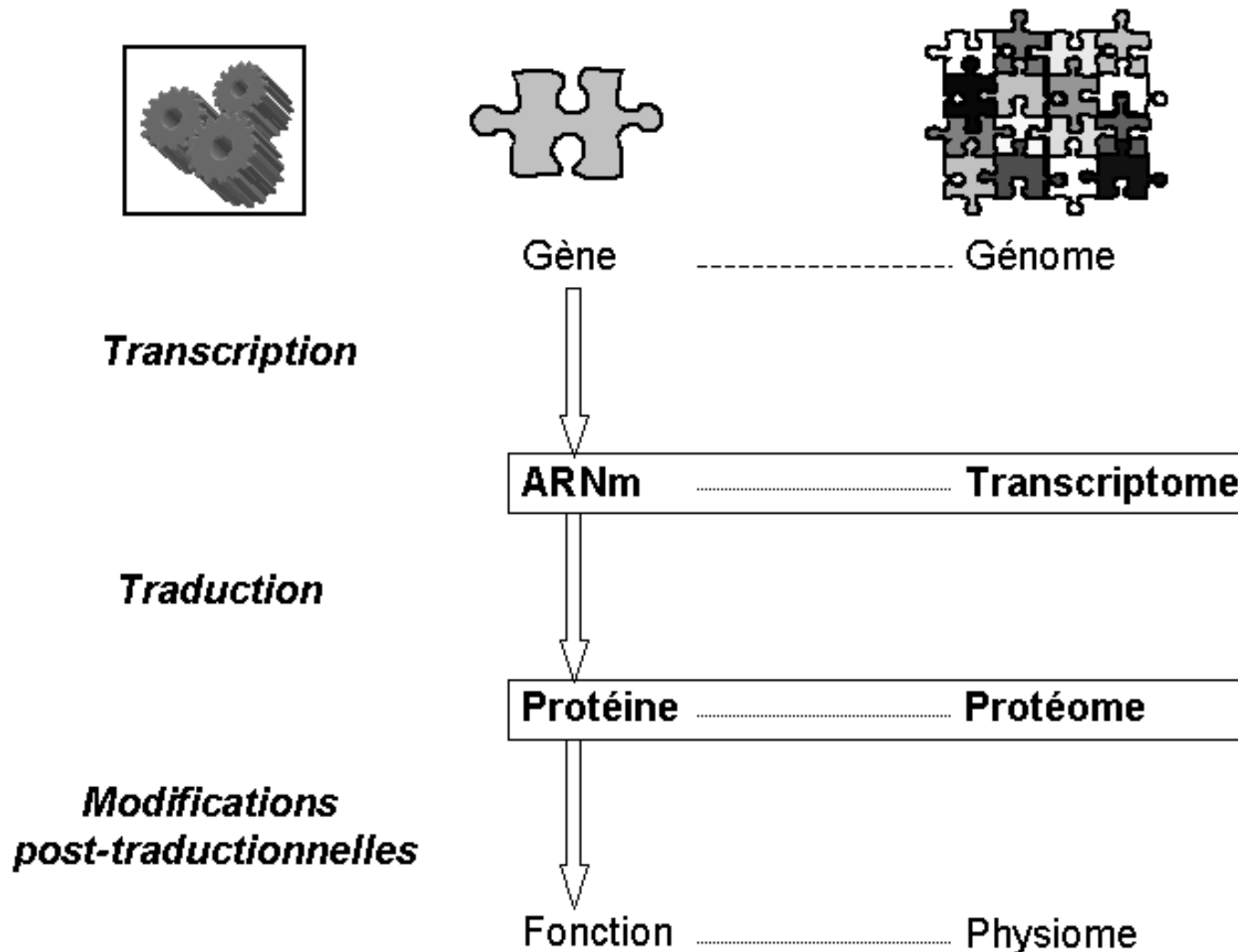
=

fonctionnement
des génomes



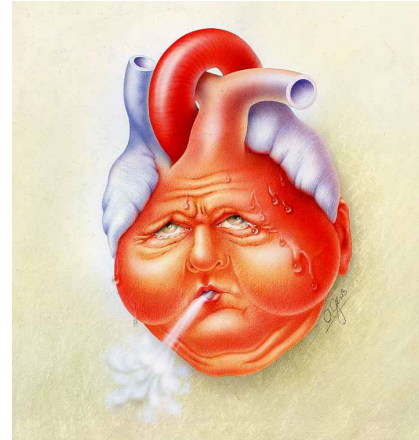
Ex: évaluer le fonctionnement global des gènes dans une maladie

Etude d'expression des gènes

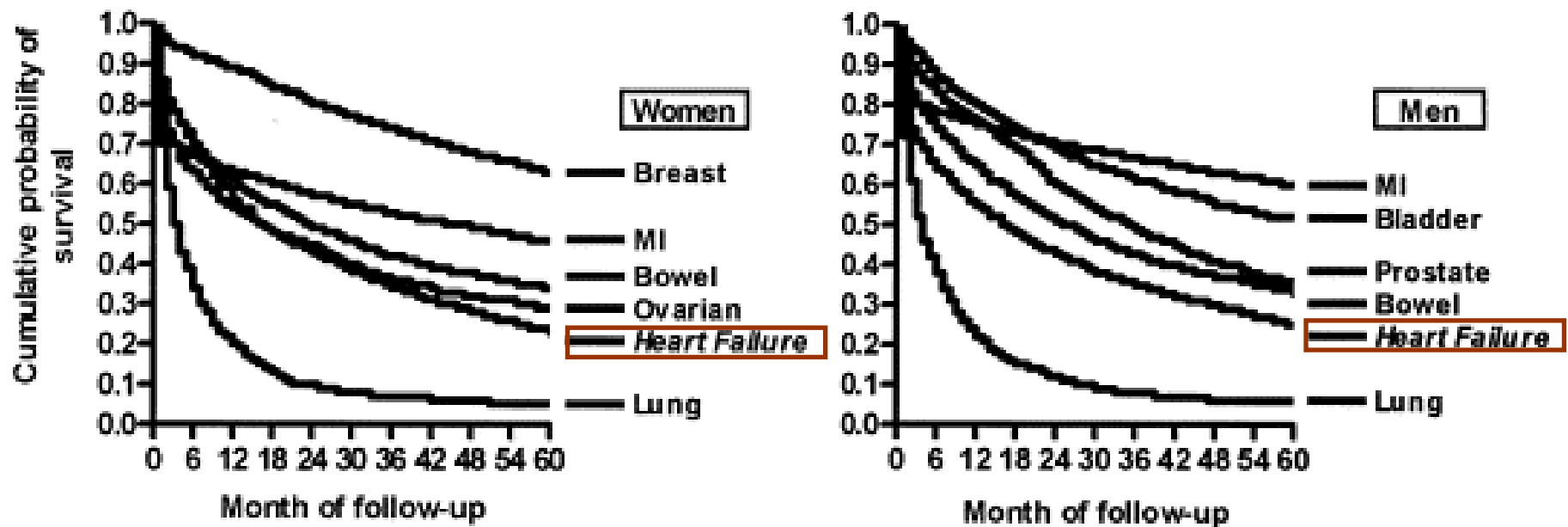


Pathologies cardiaques

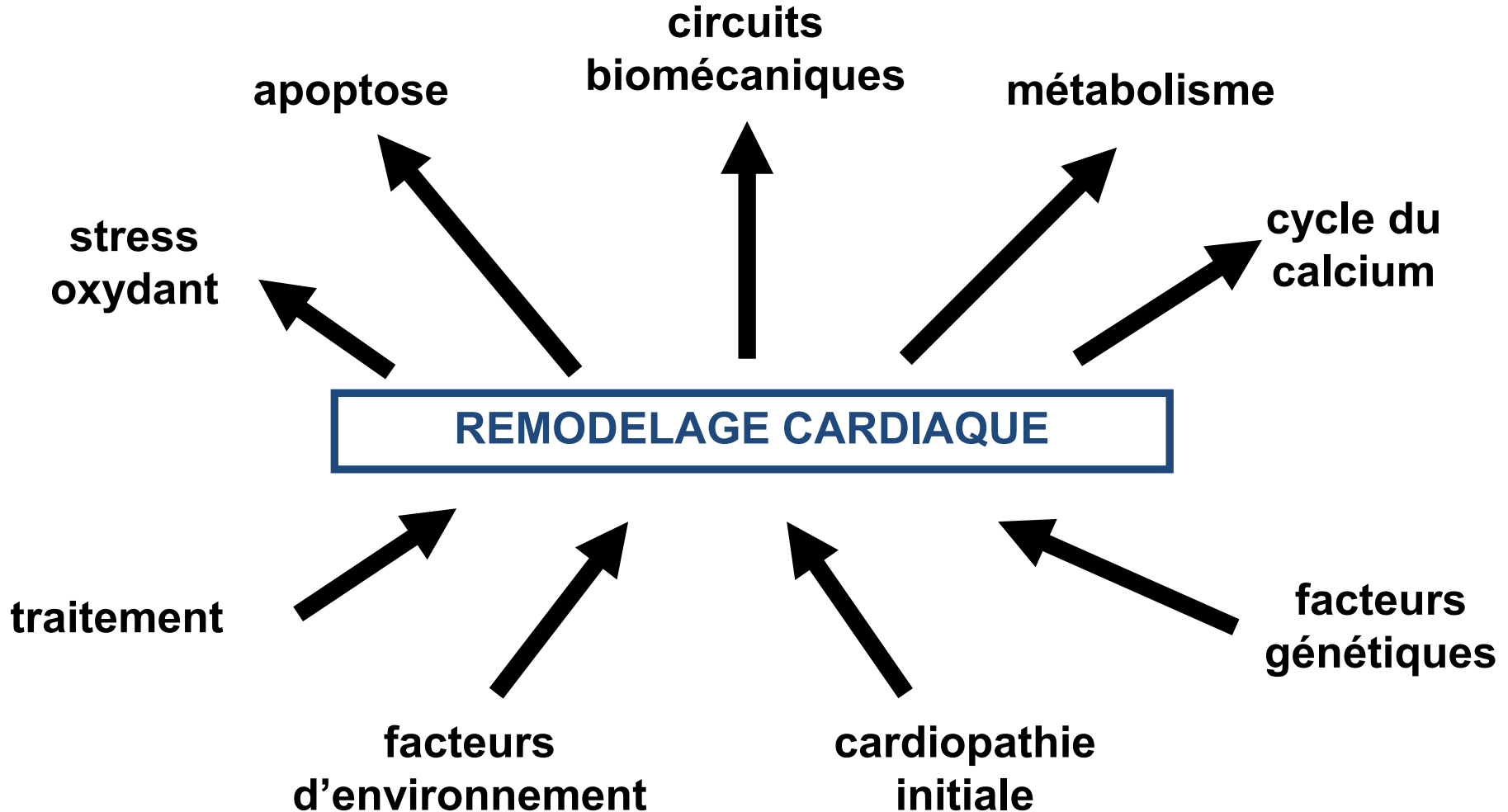
- **30% des causes de décès**
- **Forte prévalence**
- **Renforcement du poids de ces maladies avec le vieillissement de la population**
- **Facteurs de risque environnementaux et génétiques**



Insuffisance cardiaque: plus grave que le cancer?

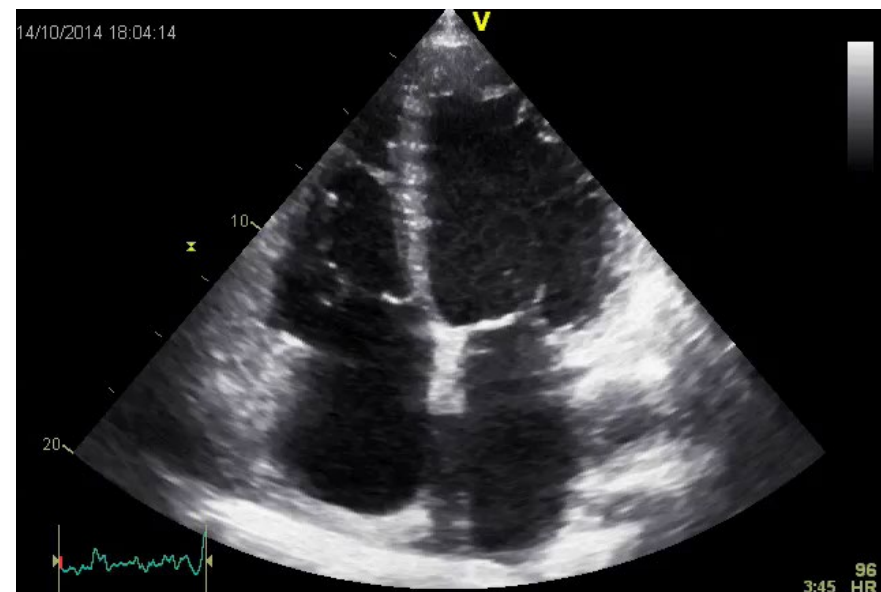
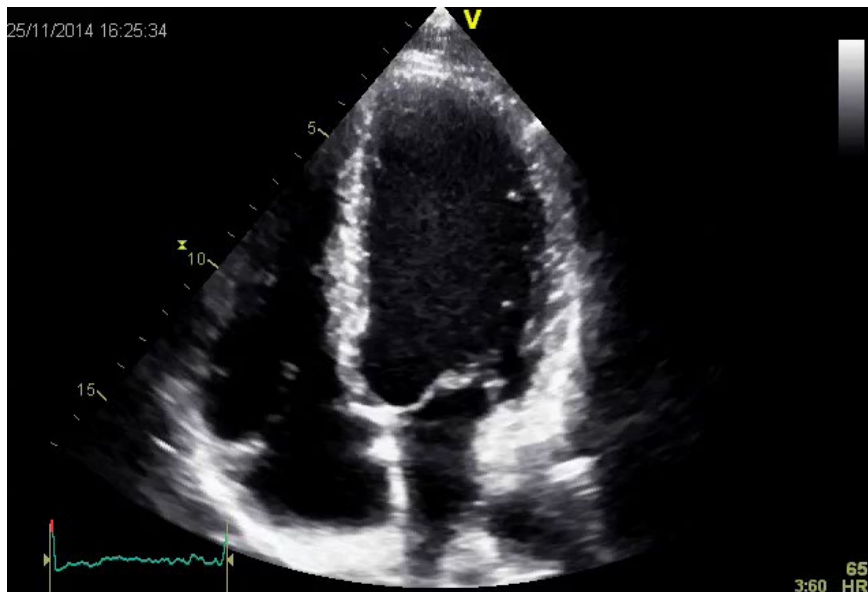


Insuffisance cardiaque: processus de remodelage complexe

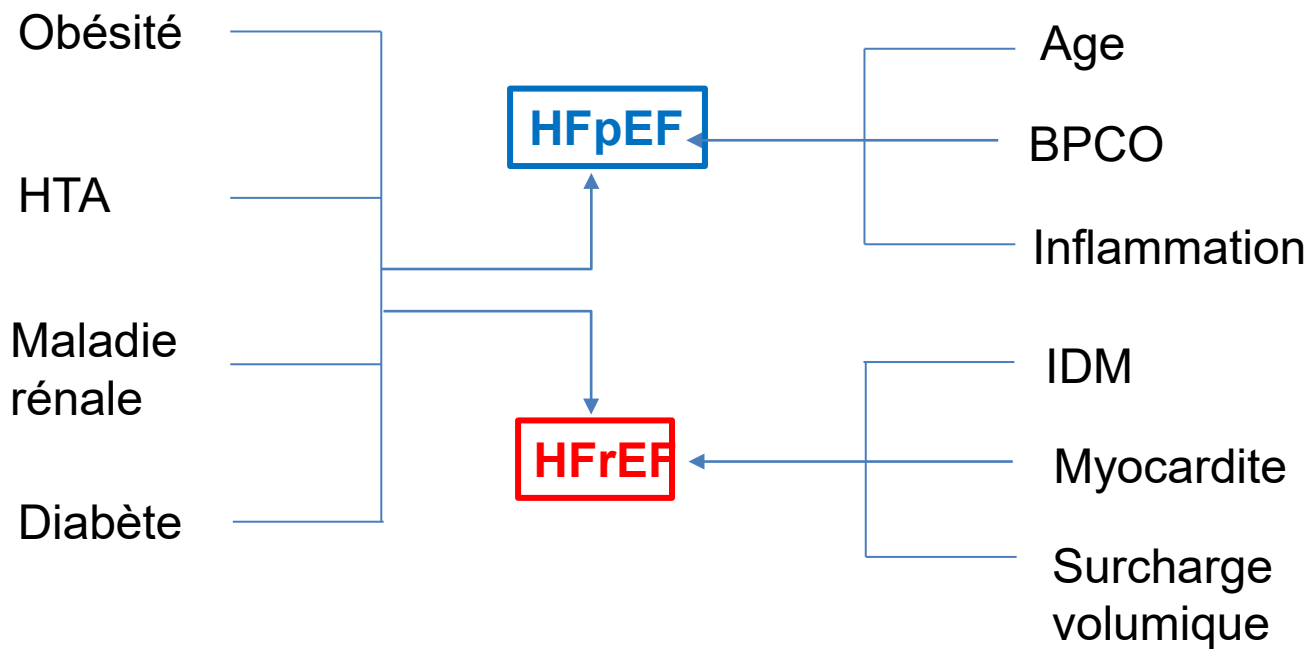


2 présentations d'IC

- Selon la fraction d'éjection ventriculaire gauche (FEVG) :
 - IC à FEVG altérée (ou insuffisance cardiaque systolique)
 - IC à FEVG préservée.



Facteurs de risque et comorbidités



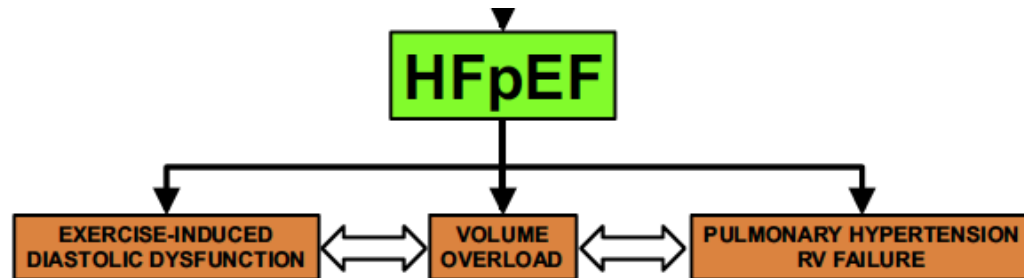
Les co-morbidités diffèrent chez les pts IC HFpEF et HFrEF

Pts + HFpEF : plus âgés, femmes, prévalence↑ comorbidités non cardiaques

Incidence hospitalisations plus fréquentes en lien avec les comorbidités si HFpEF

La mortalité liée aux comorbidités est identique

HFPEF Phenotypes (« phenomapping »)



SAMPLE PATIENTS



- 72-year-old woman
- Long-standing HTN
- NYHA II
- Exercise intolerance
- Minimal fluid retention
- No HF hospitalizations
- LVEF 70%, 2+ LAE
- Grade I DD
- PASP 30 mmHg at rest
- Exercise $E/e' = 14$

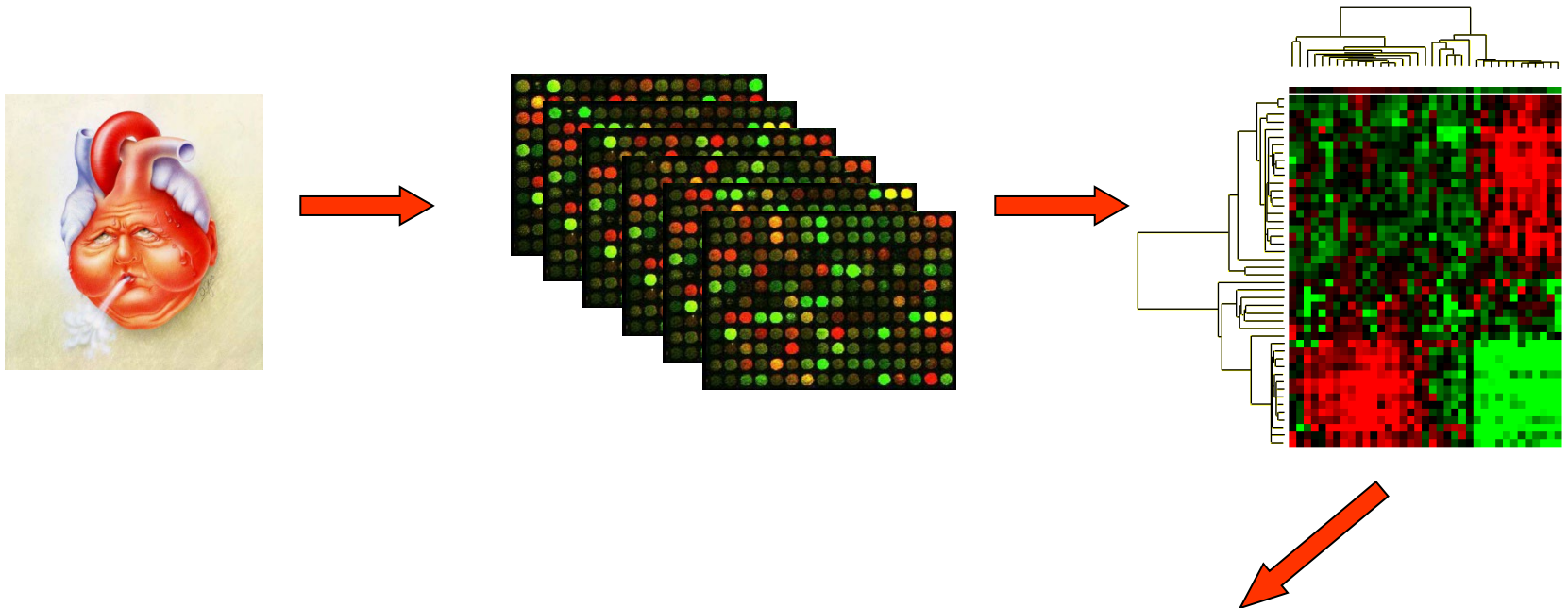


- 66-year-old woman
- HTN, CAD s/p CABG
- NYHA III
- Severe DOE
- 2+ LE edema
- Recent HF hospitalization
- LVEF 50%, 3+ LAE
- Grade III DD
- PASP 45 mmHg at rest
- 2+ MR, 2+ AR



- 59-year-old woman
- HTN, DM2, CKD, obese
- NYHA III
- Severe SOB, DOE
- 3+ edema, ascites
- Frequent HF hospitalizations
- LVEF 65%, 4+ LAE
- Grade II DD
- PASP 60 mmHg at rest
- RVH + RV dysfunction

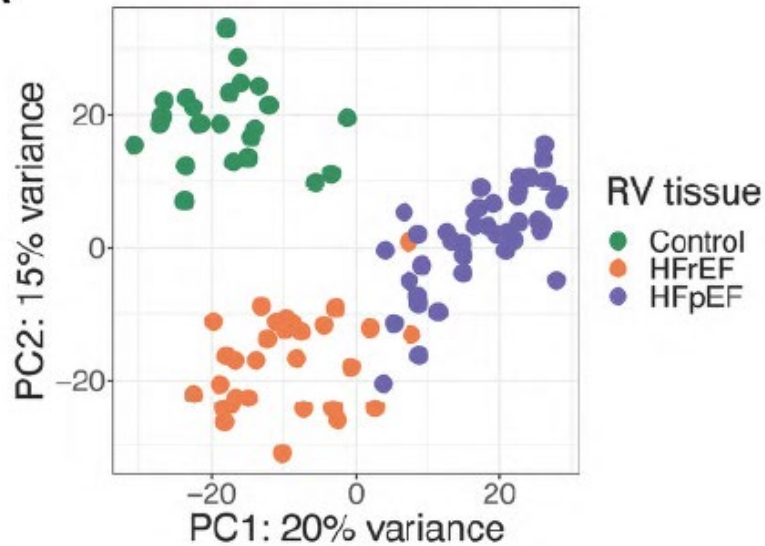
Profils d'expression génique



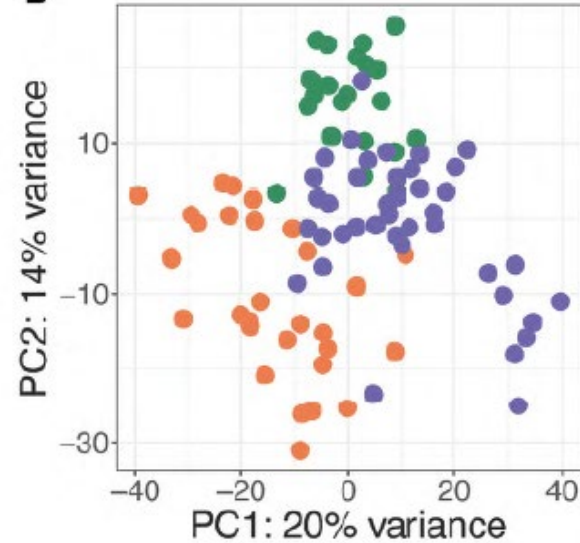
1. Découvertes physiopathologiques
2. Phénotypage moléculaire

Ins. cardiaque: étude du transcriptomique cardiaque

A



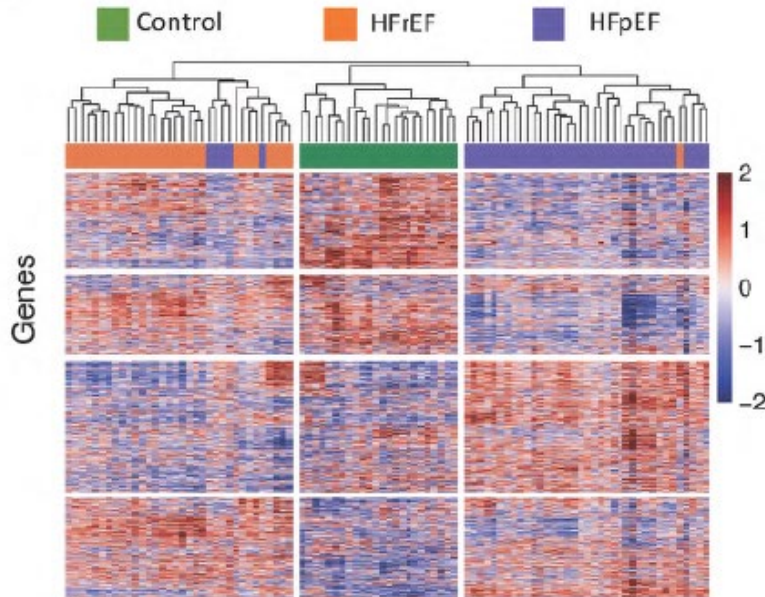
B



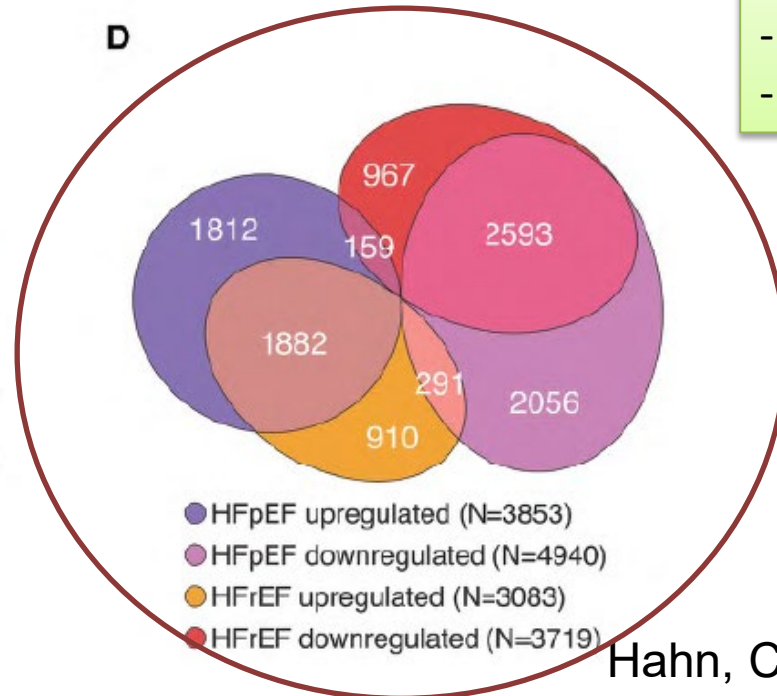
Biopsies de VD:

- Contrôle
- HFpEF
- HFrEF

C



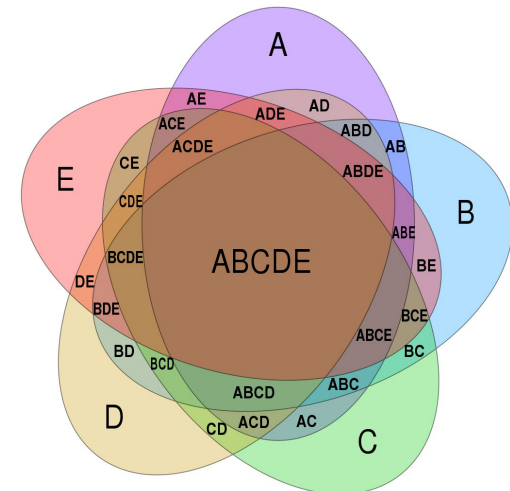
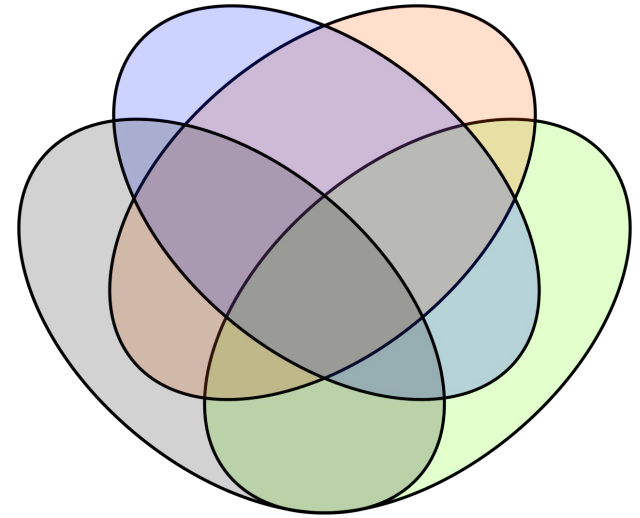
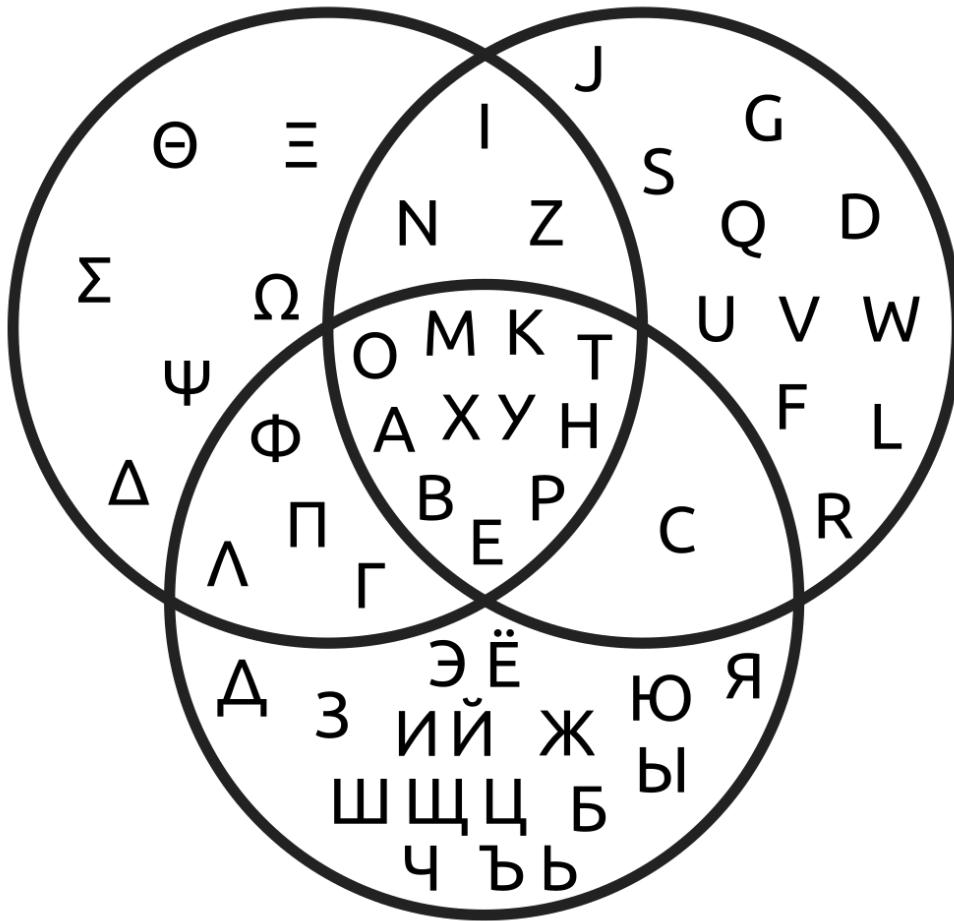
D



Interprétation des données

- **Detection de gènes différentiels**
Méthodes statistiques
- **Detection de groupes de gènes co-regulés**
- **Annotation fonctionnelle**

Diagrammes de Venn



Type I and II Errors

Actual Situation “Truth”

Decision \	Actual Situation “Truth”	
	H_0 True	H_0 False
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Reject H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$$

Why Multiple Testing Matters

Genomics = Lots of Data = Lots of Hypothesis Tests

- Environ 10000-15000 gènes s'expriment en même temps dans une cellule ou un tissu
- Analyser des différences d'expression entre 2 situations (sains vs. contrôle) pour chaque gène fait réaliser 10000 à 15000 tests statistiques
- Avec un seuil de p-value à 0,05, on s'attend donc à avoir 500 à 750 gènes déclarés comme significatifs par erreur.

Why Multiple Testing Matters

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?

$$P(\text{Making an error}) = \alpha$$

$$P(\text{Not making an error}) = 1 - \alpha$$

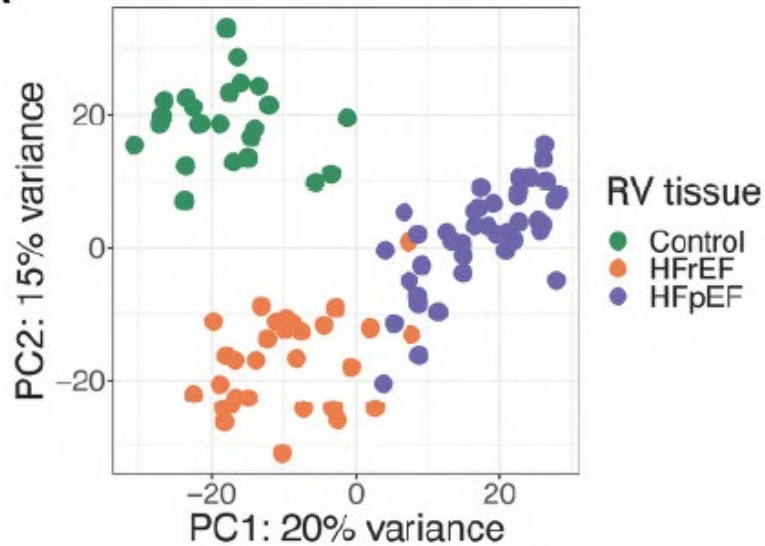
$$P(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

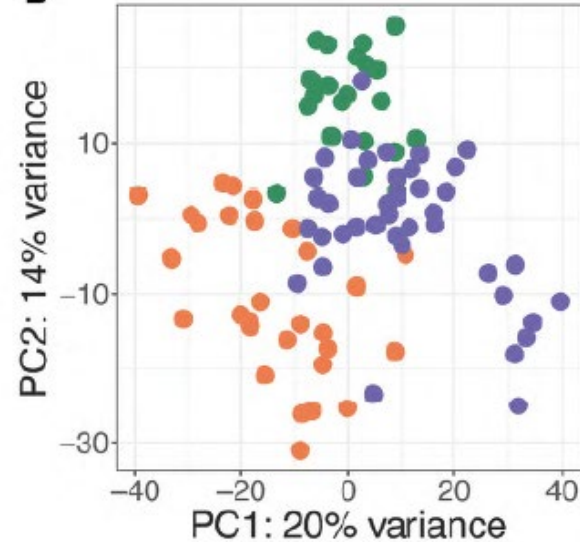
Il faut des méthodes pour contrôler l'erreur de type I

Ins. cardiaque: étude du transcriptomique cardiaque

A



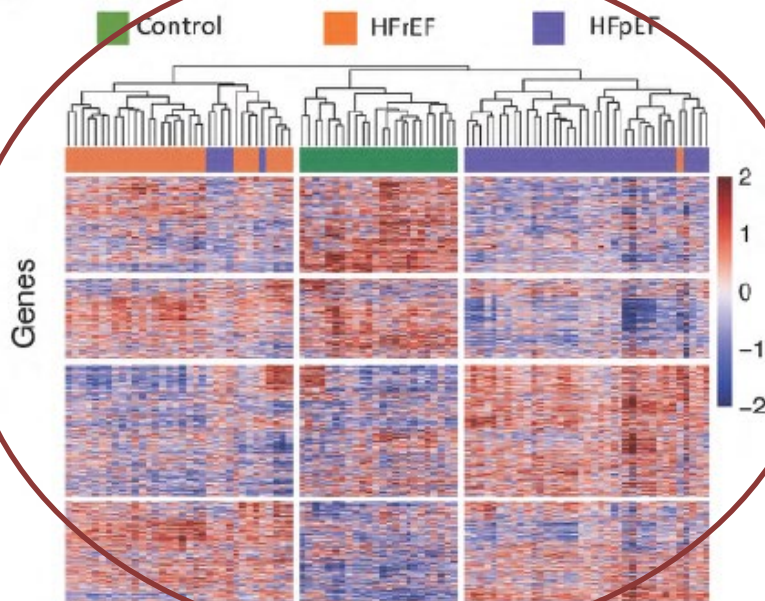
B



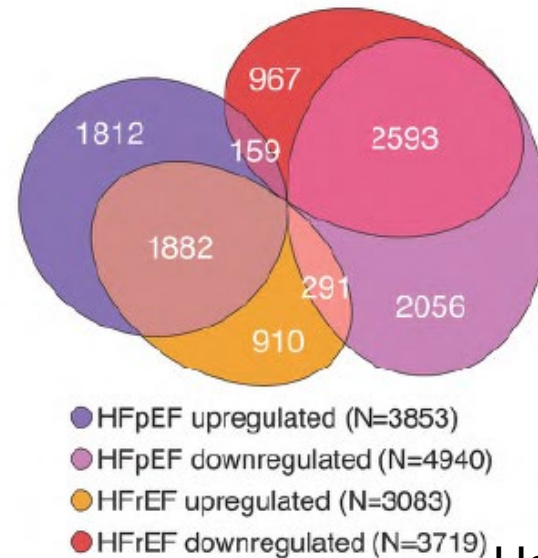
Biopsies de VD:

- Contrôle
- HFpEF
- HFrEF

C



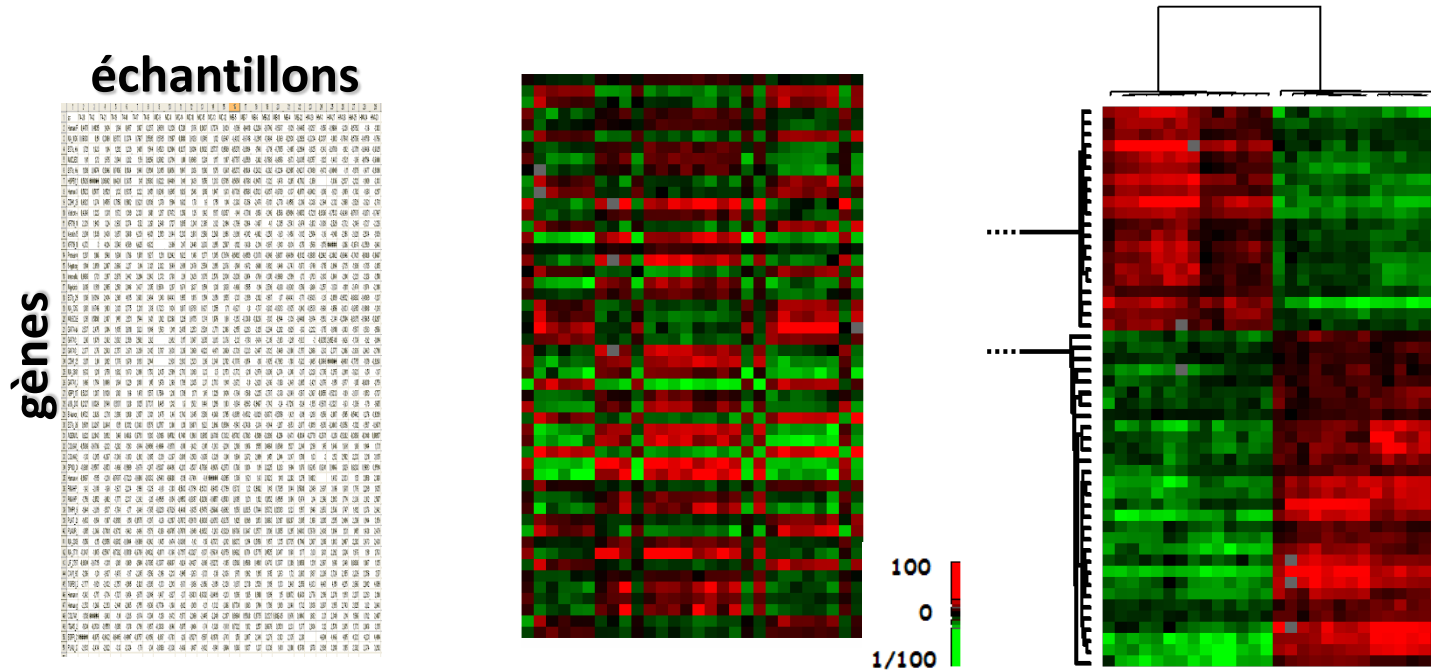
D



Interprétation des données

- **Detection de gènes différentiels**
- **Detection de groupes de gènes co-regulés**
Méthodes de « clustering »
- **Annotation fonctionnelle**

Clustering - Eisen et al. (1998)



**Données
d'expression**

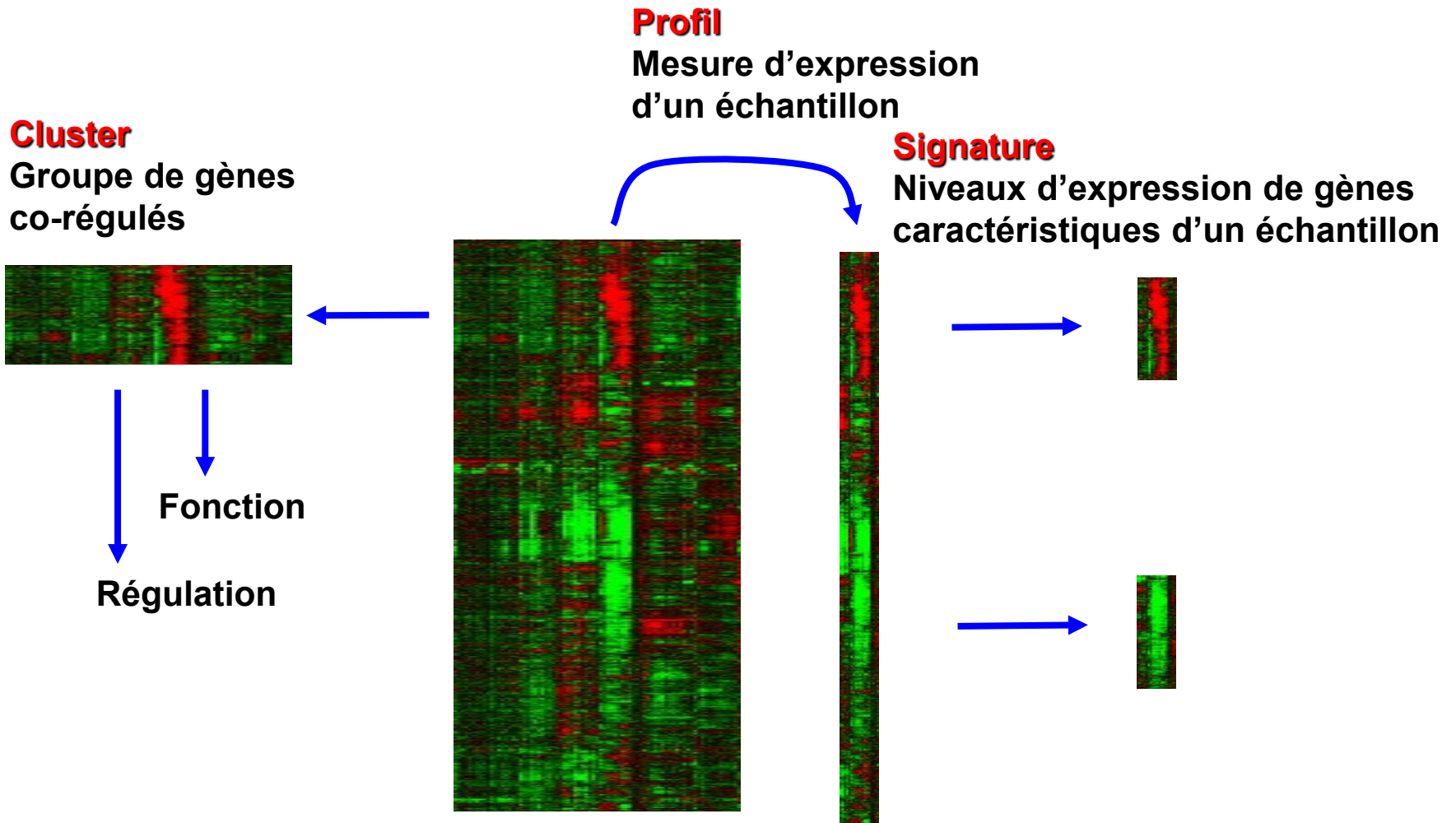


**code
couleur**



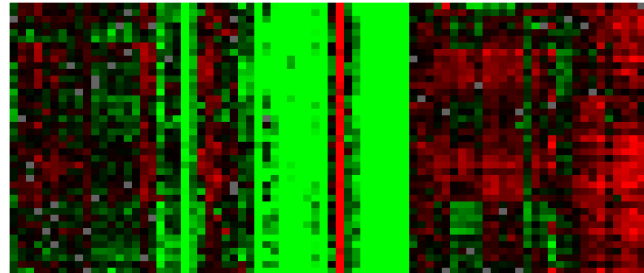
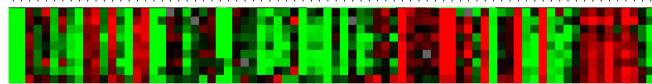
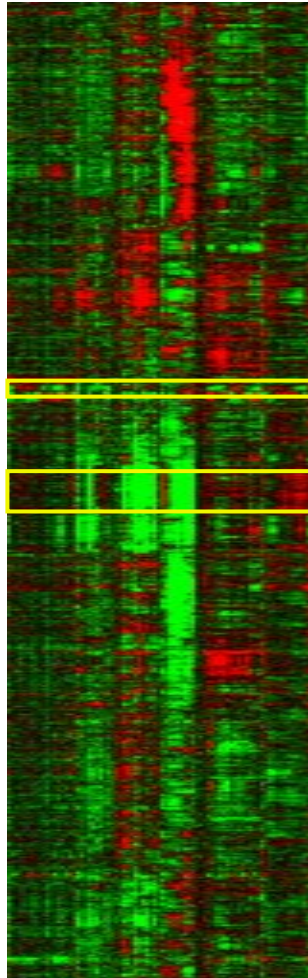
**clustering
hiérarchique**

Définitions



Correlation = Fonction

Eisen et al. (1998) PNAS 95: 14863-14868



Les gènes d'expression
corrélée participent à une
même fonction biologique

HTE2	CHROMATIN	STRUCTURE	HISTONE	H2B
HHT1	CHROMATIN	STRUCTURE	HISTONE	H3
HHE1	CHROMATIN	STRUCTURE	HISTONE	H4
HTE1	CHROMATIN	STRUCTURE	HISTONE	H2A
HHE2	CHROMATIN	STRUCTURE	HISTONE	H2B
HTE2	CHROMATIN	STRUCTURE	HISTONE	H4
HHT2	CHROMATIN	STRUCTURE	HISTONE	H2A
HHT1	CHROMATIN	STRUCTURE	HISTONE	H3
HHT1	CHROMATIN	STRUCTURE	HISTONE	H1

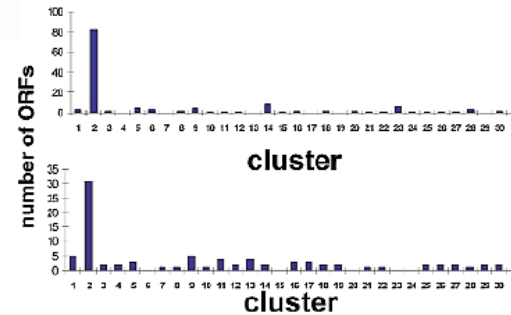
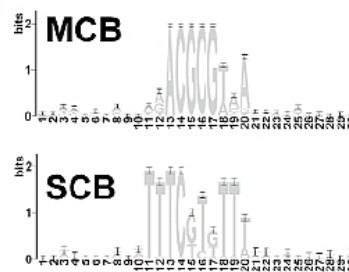
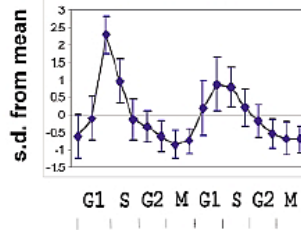
RPS5	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S5
RPS4A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S4A
RPL26B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L26B
RPS7A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S7A
RPS24A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S24A
RPS21A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S21A
RPL14A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L14A
RPL38	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L38
RPL24A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L24A
RPS1A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S1A
RPS1B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S1B
RPL24B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L24B
RPS23A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S23A
RPS23B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S23B
RPS3	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S3
RPS4B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S4B
RPS6B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S6B
RPL5	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L5
RPL17B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L17B
RPS19A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S19A
RPL8B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L8B
RPS19B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S19B
RPS26B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S26B
RPL11A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L11A
RPL11B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L11B
RPL34B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L34B
RPL10	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L10
RPL12B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L12B
RPS15	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S15
RPS20	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S20
RPS0A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S0A
RPL18B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L18B
RPL2B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L2B
RPL16B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L16B
RPS0B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	S0B
RPL7B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L7B
RPL7A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L7A
RPL35B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L35B
RPL6B	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L6B
RPL33A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L33A
RPL13A	PROTEIN	SYNTHESIS	RIBOSOMAL	PROTEIN	L13A

Cluster = Régulation

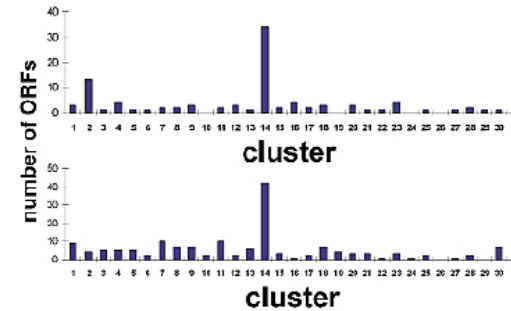
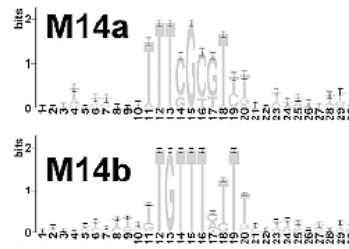
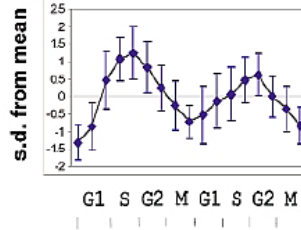
Tavazoie et al. (1998) Nature Genetics



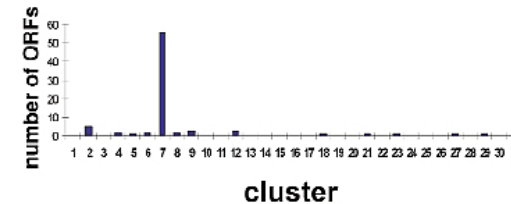
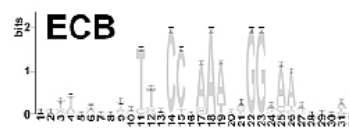
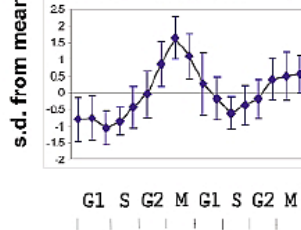
replication & DNA synthesis (2)



organization of centrosome (14)

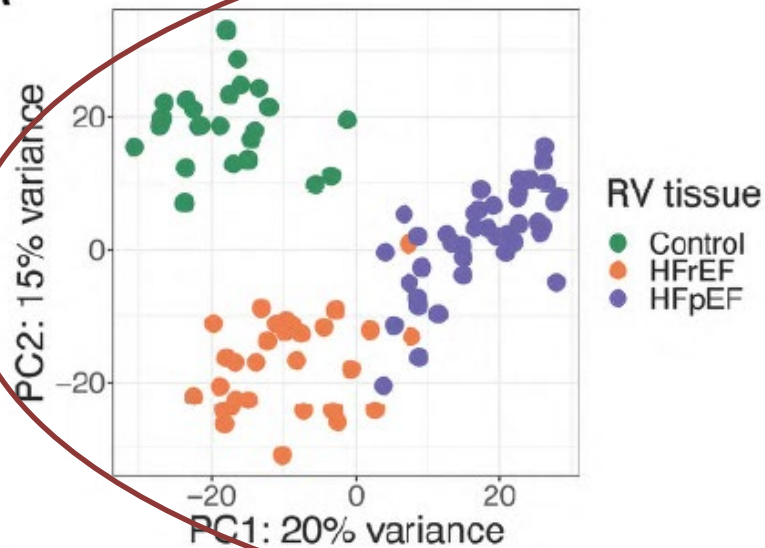


budding and cell polarity (7)

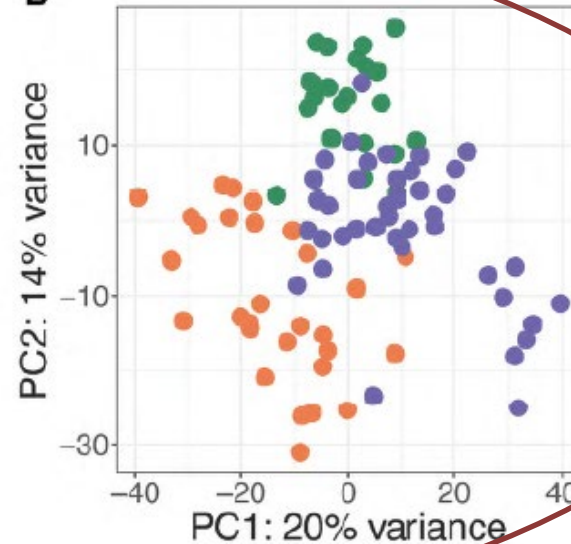


Ins. cardiaque: étude du transcriptomique cardiaque

A



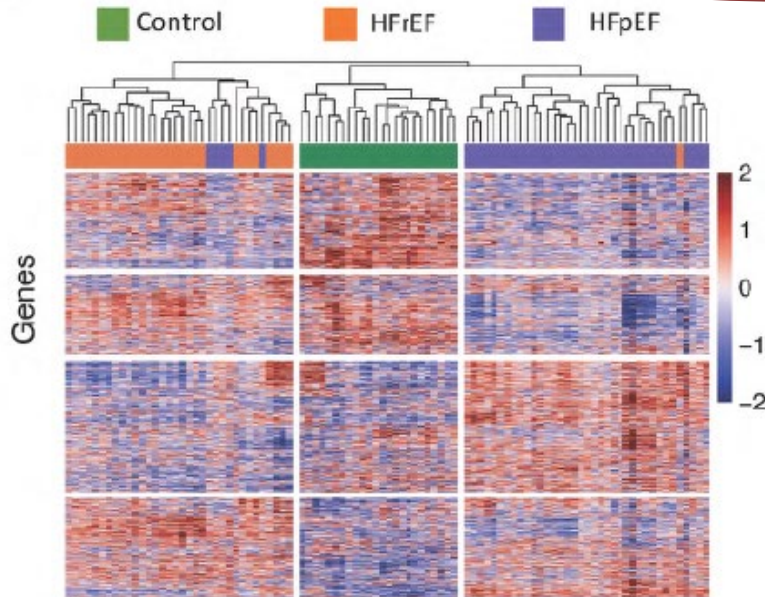
B



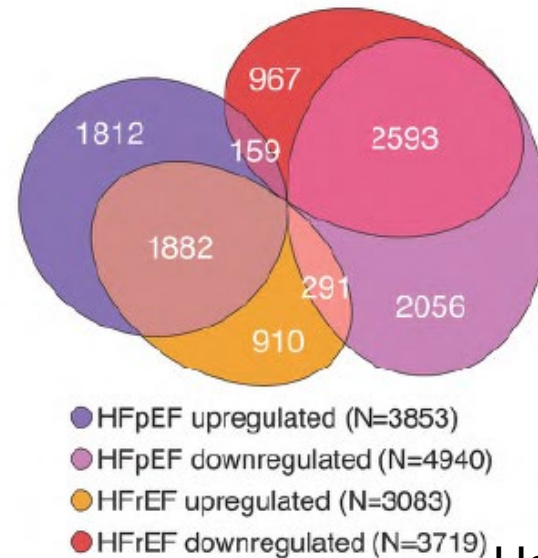
Biopsies de VD:

- Contrôle
- HFpEF
- HFrEF

C



D

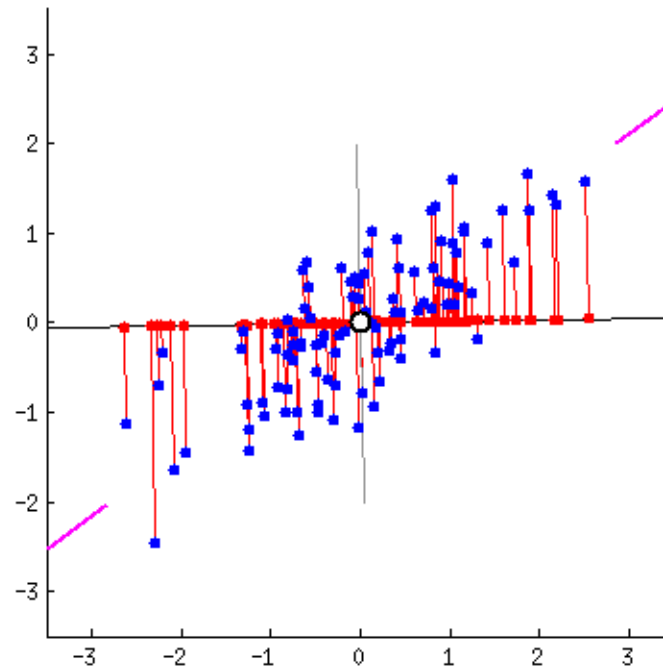


Interprétation des données

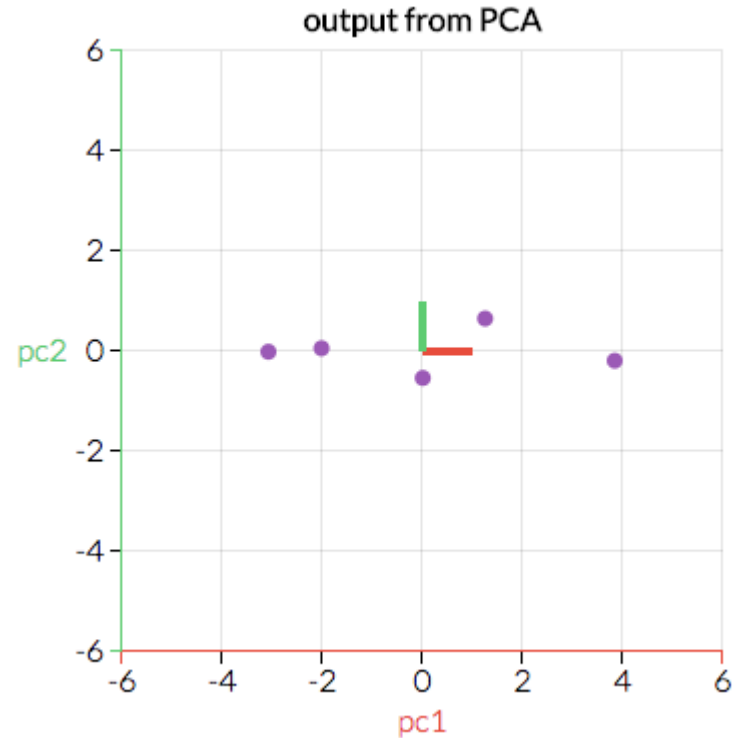
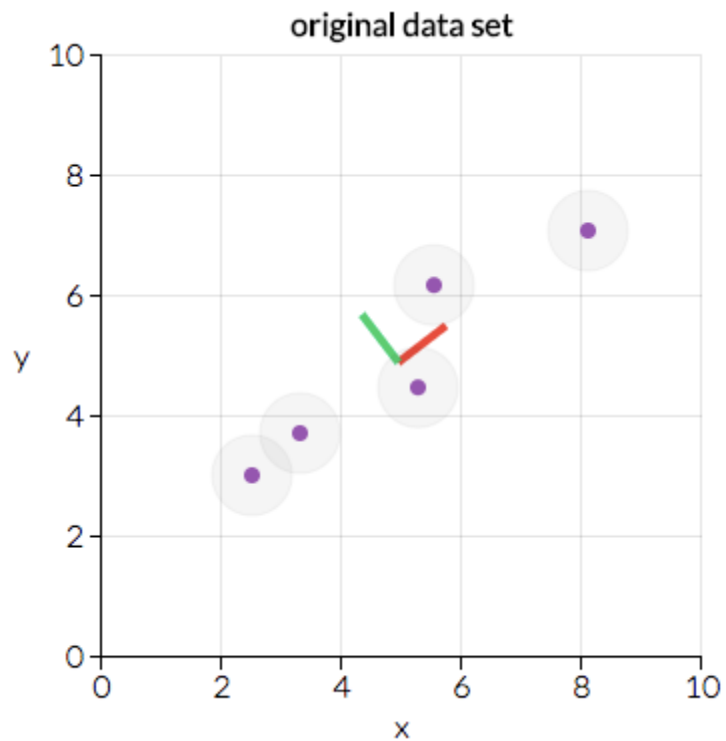
- **Detection de gènes différentiels**
- **Detection de groupes de gènes co-regulés**
Réduction de dimension
- **Annotation fonctionnelle**

Analyse en composante principale

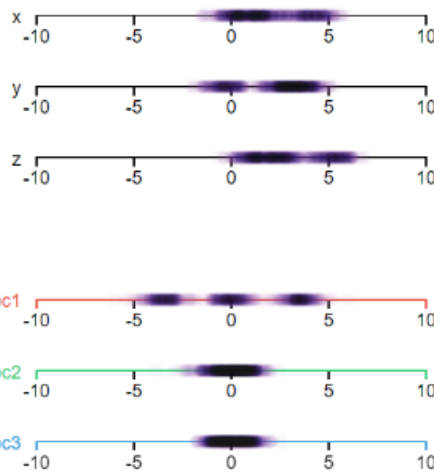
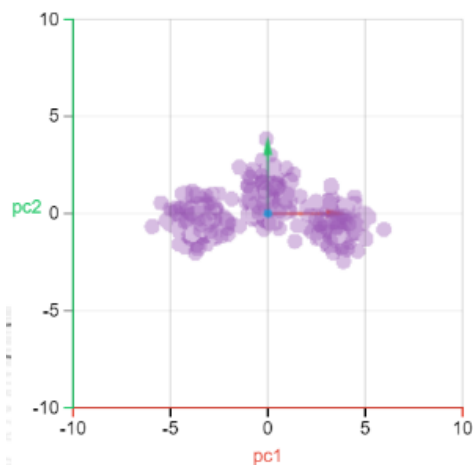
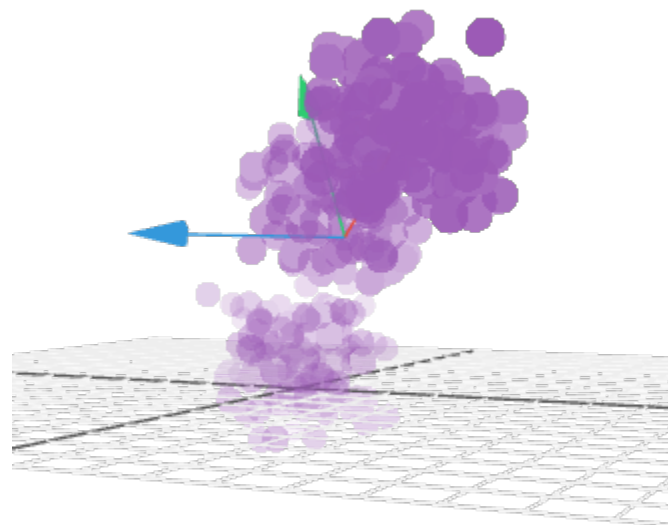
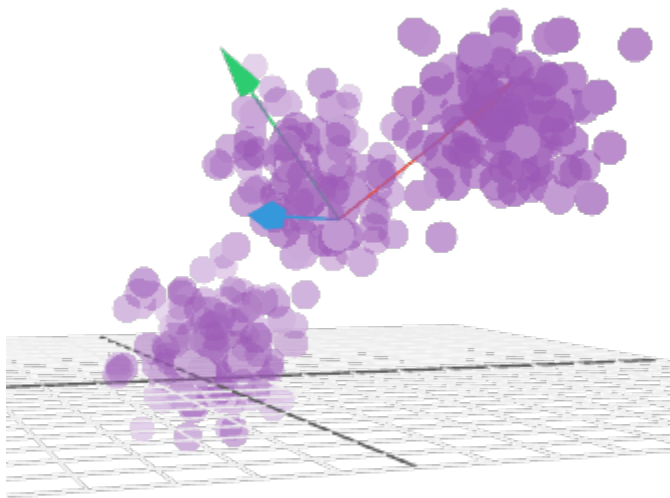
- Une méthode de réduction du nombre de variables utiles en préservant au maximum la quantité d'information
- Création séquentielle de nouvelles variables (« pseudogènes ») qui capturent à chaque fois le plus possible d'information discriminante
- Plus de simplicité/lisibilité en limitant la perte de précision/information



5 échantillons et 2 variables

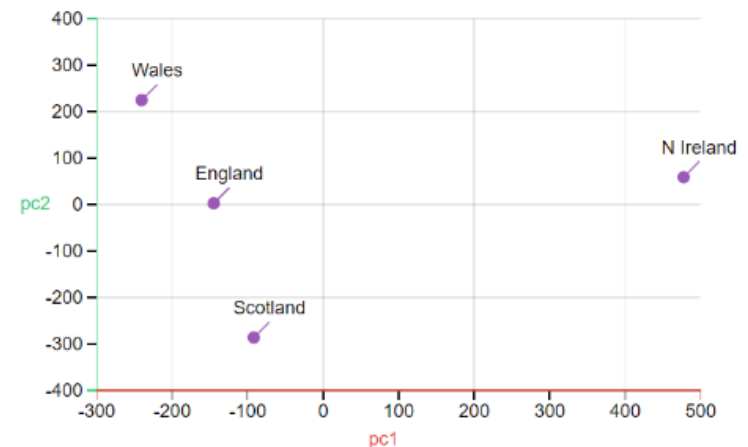
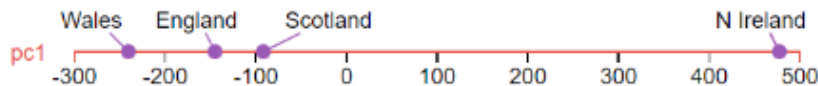


Nombreux échantillons et 3 variables

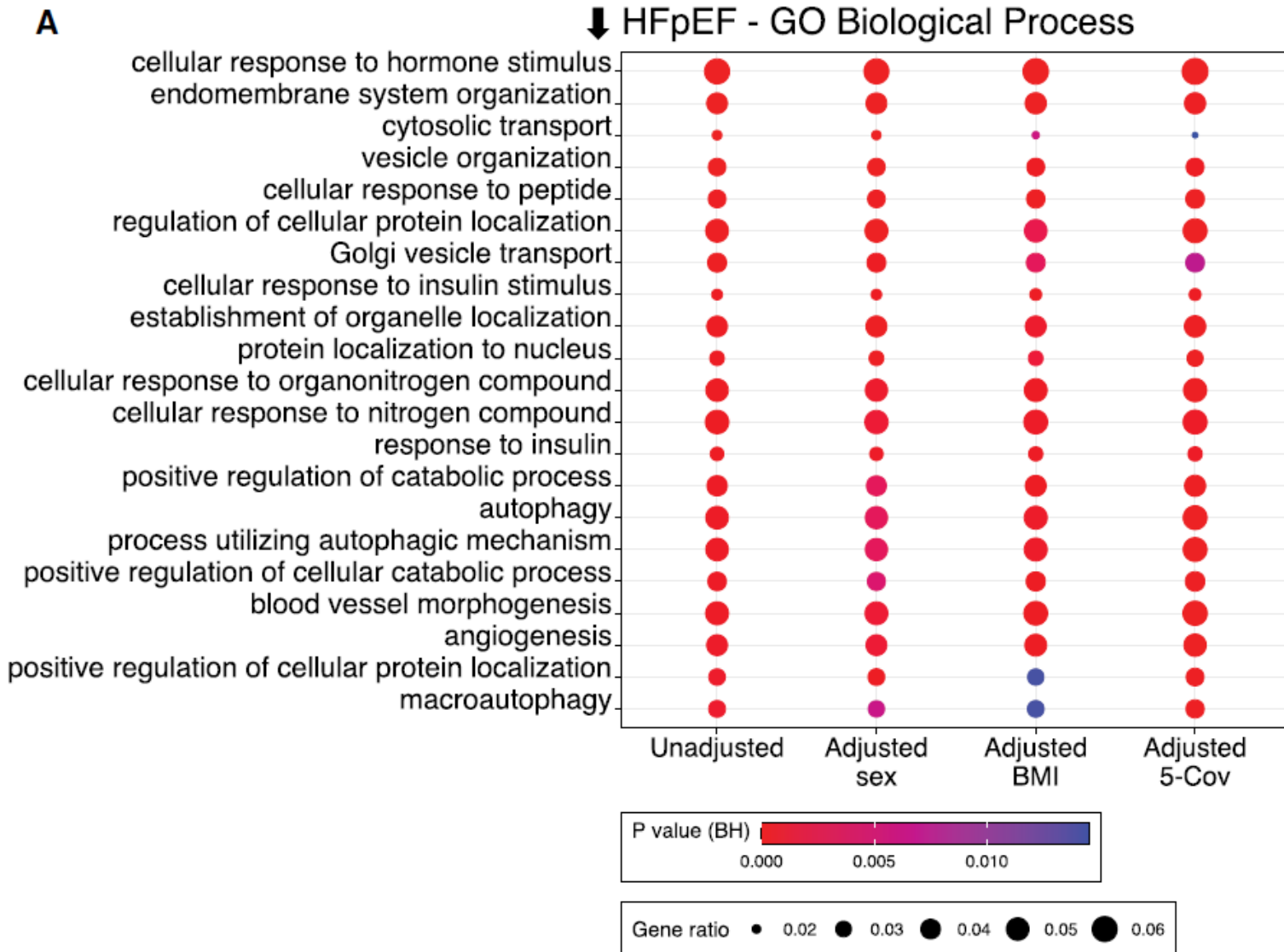


4 groupes et 17 variables

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



A quelles fonctions biologiques participent les gènes sous-exprimés

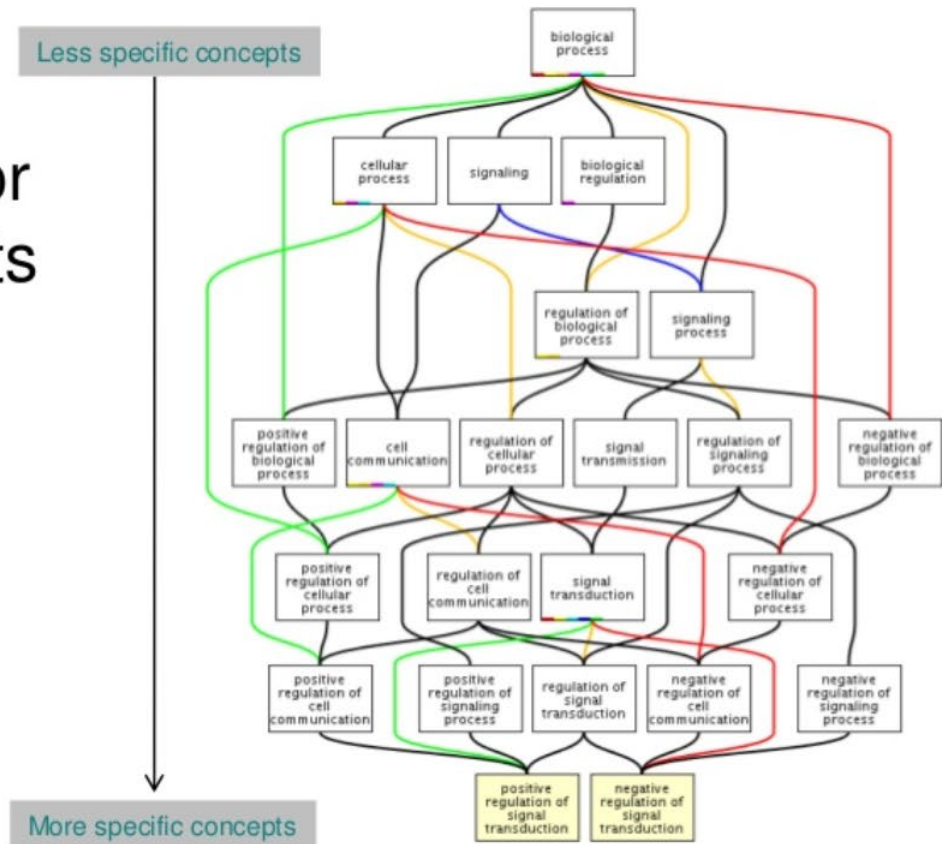


Interprétation des données

- **Détection de gènes différentiels**
- **Détection de groupes de gènes co-regulés**
- **Annotation fonctionnelle**
 - Banques d'information (www)**
Gene Ontology, Reactome,...

The Gene Ontology

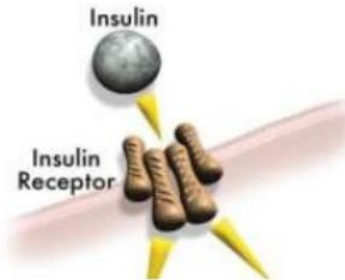
- A way to capture biological knowledge for individual gene products in a written and computable form
- A set of concepts and their relationships to each other arranged as a hierarchy



<http://www.ebi.ac.uk/QuickGO>

1. Molecular Function

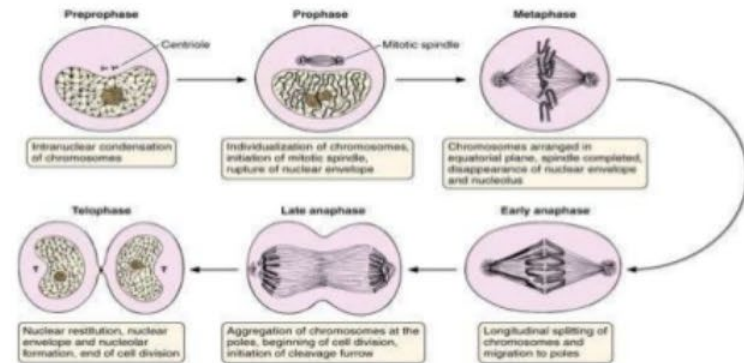
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

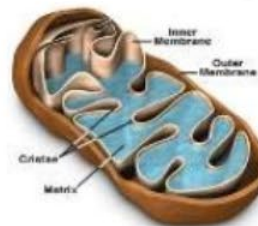
A commonly recognized series of events



- cell division

3. Cellular Component

Where a gene product is located

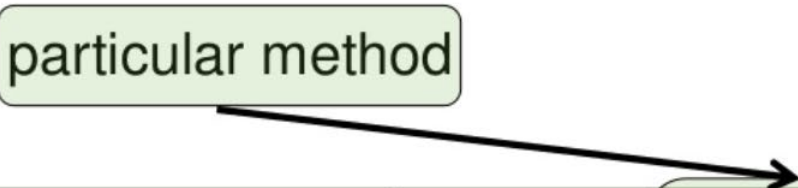


- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

A GO annotation is ...

...a statement that a gene product;

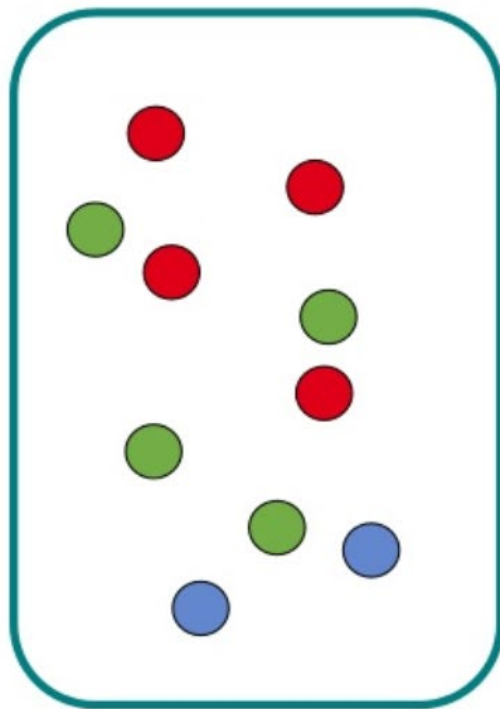
1. has a particular molecular function
or is involved in a particular biological process
or is located within a certain cellular component
2. as described in a particular reference
3. as determined by a particular method



<i>Accession</i>	<i>Name</i>	<i>GO ID</i>	<i>GO term name</i>	<i>Reference</i>	<i>Evidence code</i>
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Enrichment analysis

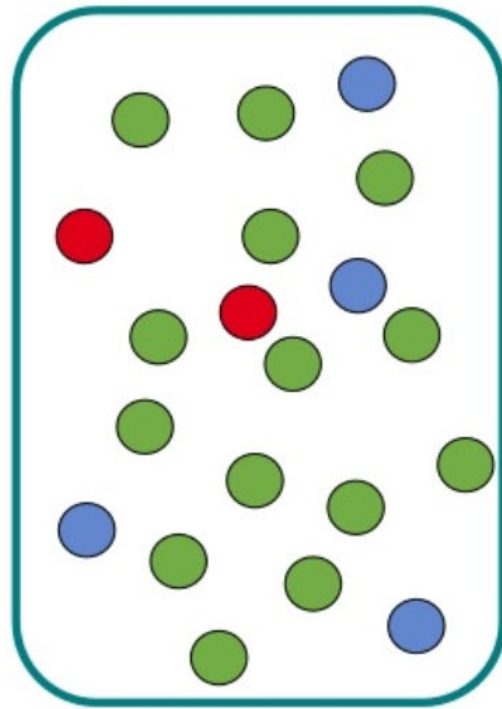
Sample



● 40%

● 20%

Reference



● 20%

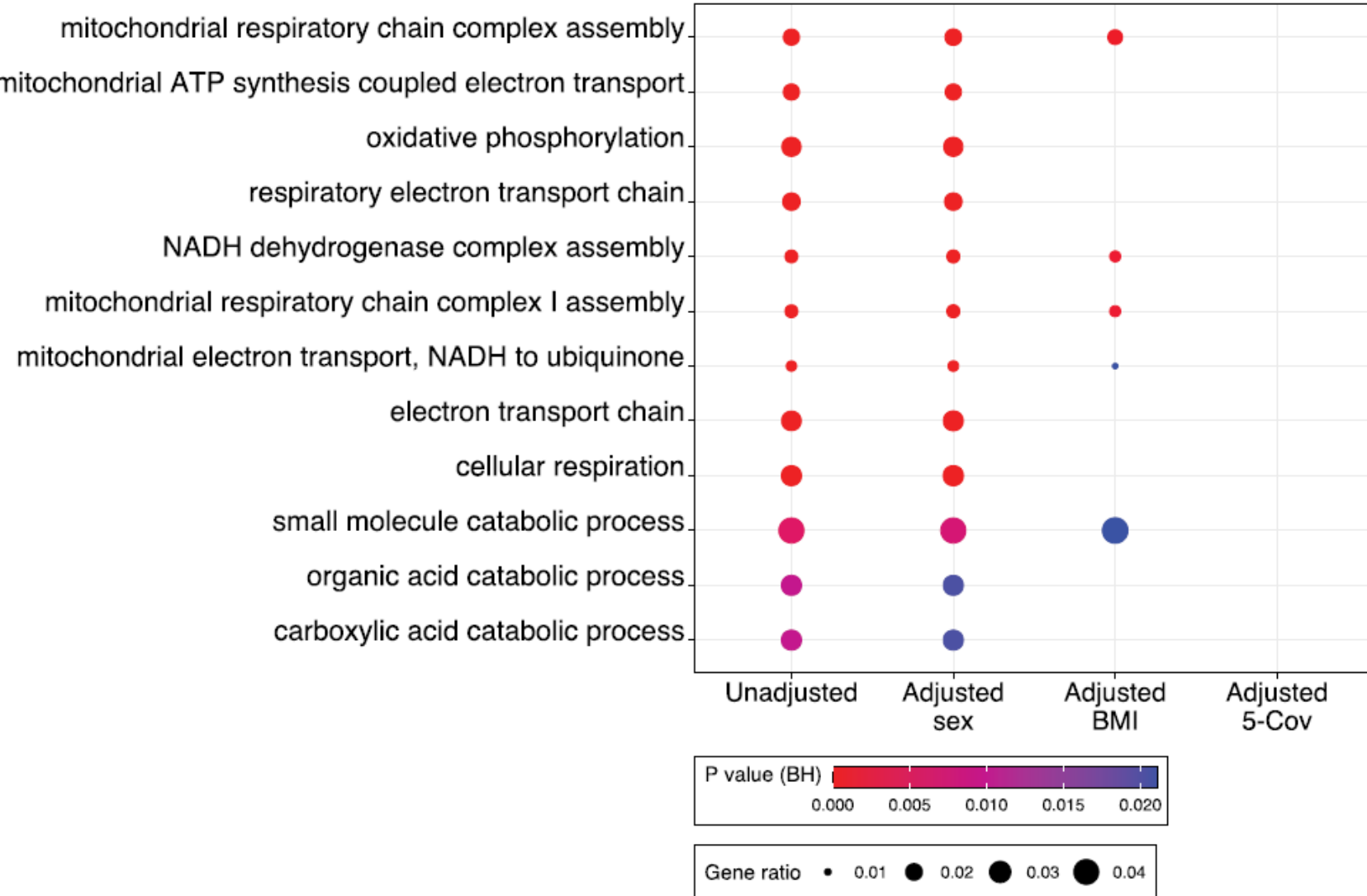
● 20%

=> The sample is over-enriched for ●

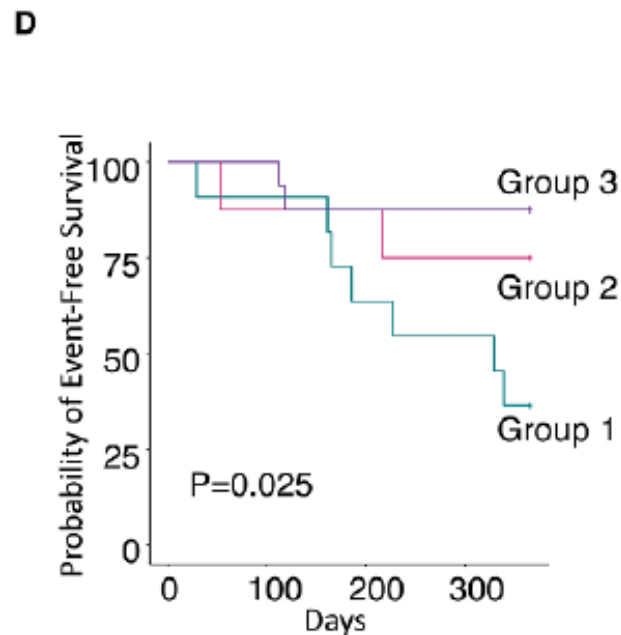
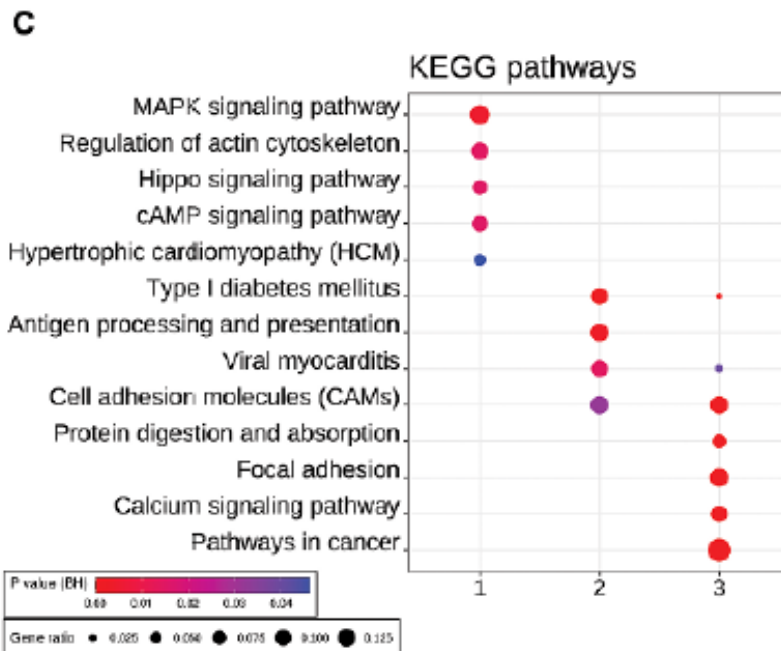
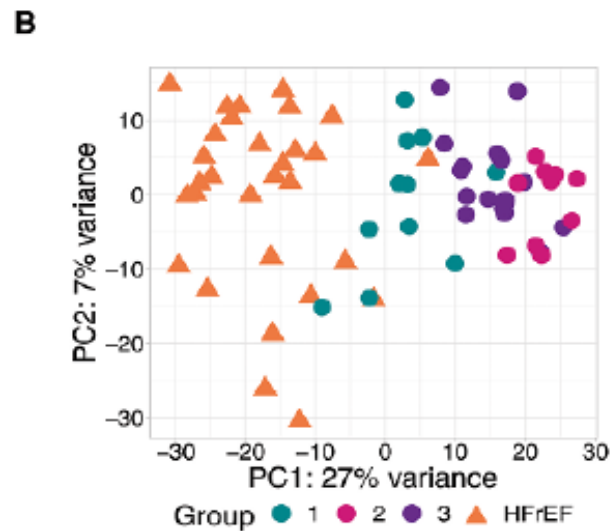
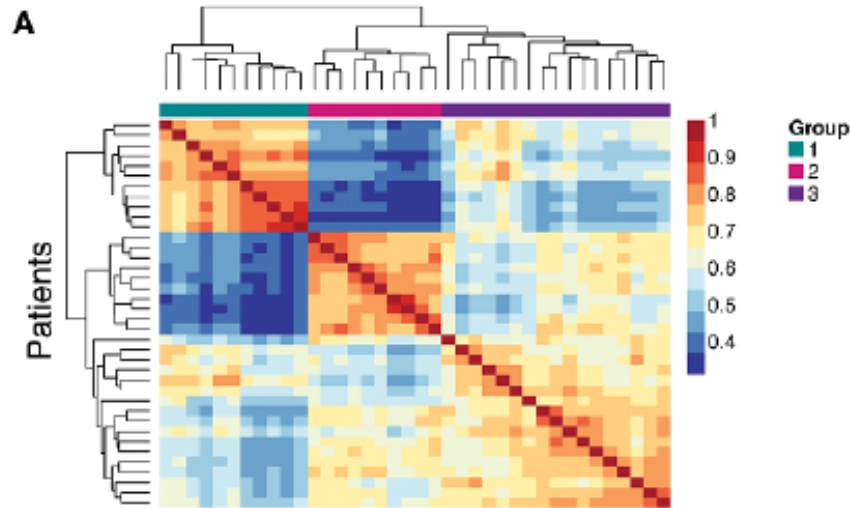
A quelles fonctions biologiques participent les gènes sur-exprimés

B

↑ HFpEF - GO Biological Process

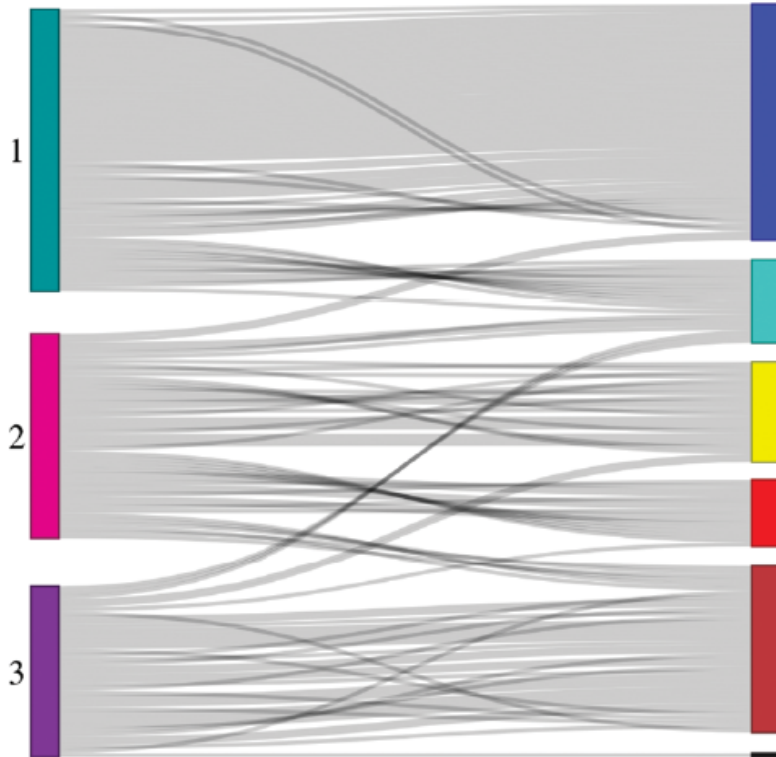


Sous-groupes de patients HFpEF



Sous-groupes de gènes et sous-groupes de patients

B

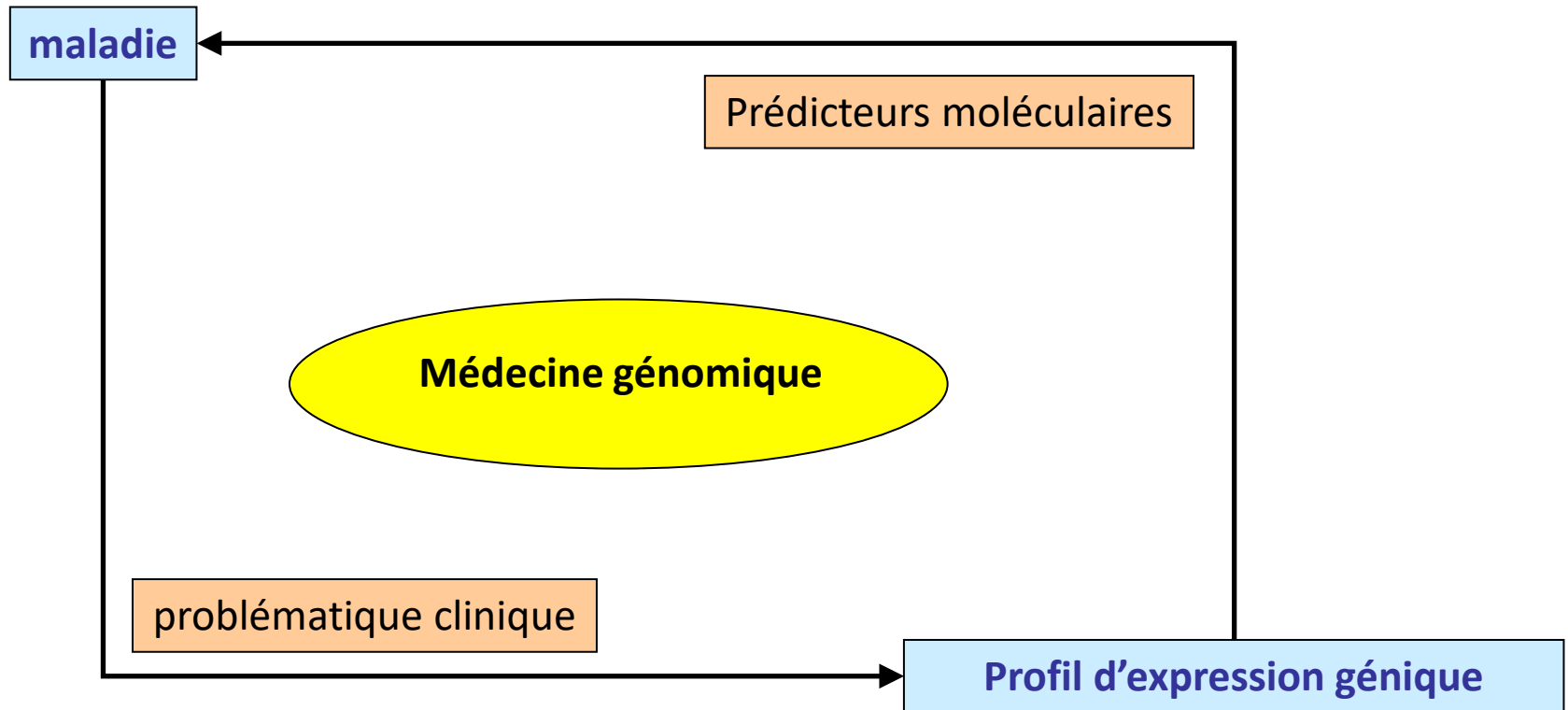


C

	Clinical Characteristics	Gene Ontology Biological Processes
Blue	Higher LV wall thickness and myocyte hypertrophy, Higher RV afterload and pulmonary vascular resistance, higher NTproBNP	Regulation of cytoskeleton organization; Striated muscle cell development; Sarcomere organization; Actin filament organization; Activation of protein kinase activity; Stress-activated protein kinase signaling cascade
Turquoise	Lower BMI, higher LV wall thickness and myocyte hypertrophy	Histone modification; Golgi organization; Regulation of gene expression, epigenetic; Negative regulation of microtubule depolymerization
Yellow	More female, less RV load and hypertrophy	Oxidative phosphorylation; ATP biosynthesis; electron transport chain; mitochondrial translation and membrane organization; cell response to hypoxia; T-cell and antigen receptor signaling
Red	Female and higher +CD68	Protein targeting to endoplasmic reticulum; Negative regulation of protein ubiquitination
Brown	Worse NYHA functional class, smaller heart and lower cardiac output, less diabetes, better renal function, lower NTproBNP, higher +CD68	Extracellular matrix organization; Regulation of actin cytoskeleton organization; Blood vessel morphogenesis; Inflammatory response; Chemotaxis
Black	Lower blood pressure, less LV hypertrophy	mRNA processing; RNA splicing

L'expression des gènes comme pont entre données cliniques et processus biologiques

Perspectives



Merci de votre attention!

Questions?

Guillaume LAMIRAULT

guillaume.lamirault@univ-nantes.fr

L'unité de recherche de l'institut du thorax

Inserm UMR 1087 / CNRS UMR 6291

Nantes, France

www.umar1087@univ-nantes.fr

