Analyse des données de la Biologie Moléculaire Identifications de Biomarqueurs Associés au Diagnostic et au Pronostic des Maladies

Pr Roch Giorgi

roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

Biomarqueur – Introduction (1)

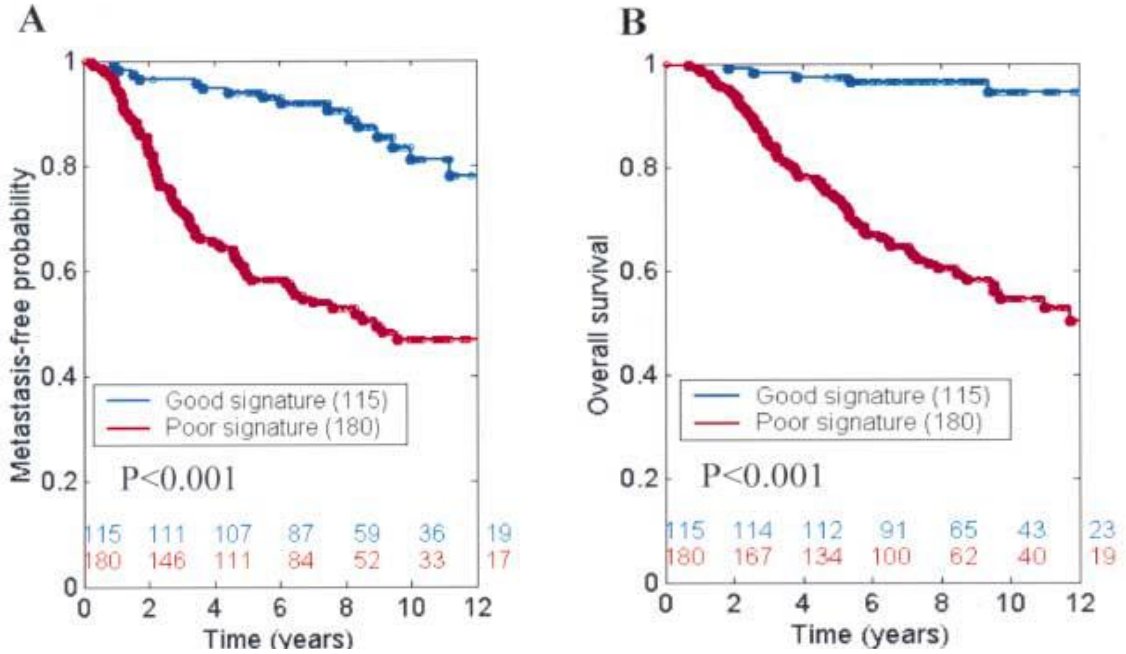
- Elément quantifiable permettant de renseigner sur
  - ✓ Un effet biologique
  - ✓ Une susceptibilité
  - ✓ Une pathologie
- Ses valeurs différentes entre des groupes de patients distincts
- Développement des techniques de biologie moléculaire, progrès dans l'analyse de l'ADN, de l'ARN et des protéines
  - ↗ du nombre de pathologies pouvant être identifiées, suivies avec la mesure de biomarqueurs

Biomarqueur – Introduction (2)

- Types de biomarqueurs
  - ✓ Diagnostique
    - Identification des sujets malades, des sujets sains
    - Associé à la présence de la maladie au stade pré-clinique
      ⇒ intérêt pour le dépistage
    - Associé à une caractéristique de la maladie
      ⇒ intérêt diagnostique
    - Le biomarqueur permet de savoir quoi traiter

Biomarqueur – Introduction (3)

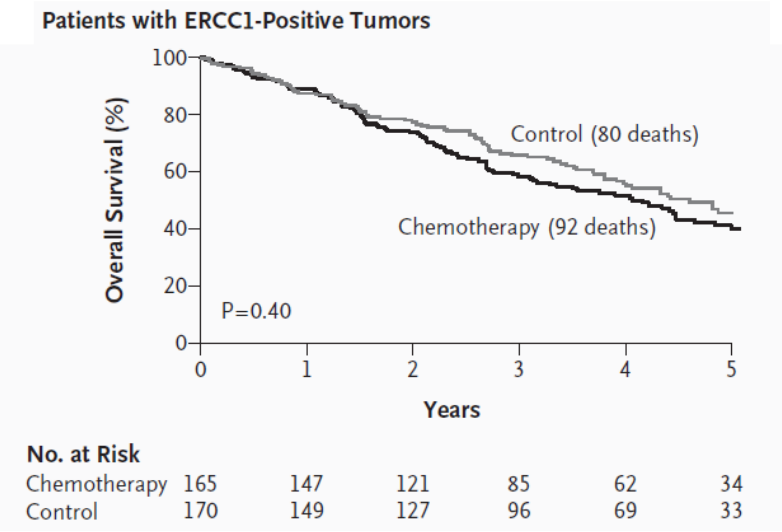
- Types de biomarqueurs
  - ✓ Pronostique
    - Caractérise l'évolution de la maladie i.e. survie des patients
    - Survie globale, survie spécifique, survie sans rechute,…
    - Le biomarqueur permet de savoir qui traiter

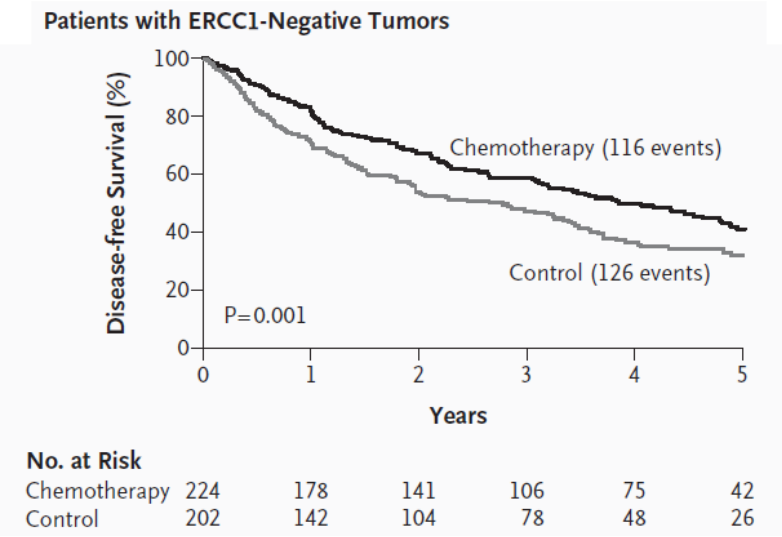


Van de Vijver MJ et al. N Engl J Med 2002;347:1999-2009.

Biomarqueur – Introduction (4)

- Types de biomarqueurs
  - ✓ Prédictif
    - Associé à la réponse à un traitement
    - Survie globale, survie spécifique, survie sans rechute,…
    - Le biomarqueur permet de savoir comment traiter





Olaussen KA et al. N Engl J Med 2006;355:983-91.

Biomarqueur Moléculaire

- Ancien postulat
  - ✓ 1 gène = 1 ARN = 1 protéine
    ⇒ l'étude du génome permet de tout comprendre
- Or
  - ✓ Le génome code pour 20 000 gènes
  - ✓ Plus de 1 million d'ARNm
  - ✓ Plus de 10 millions de protéines

    ⇒ analyse du génome ADN → étude génomique

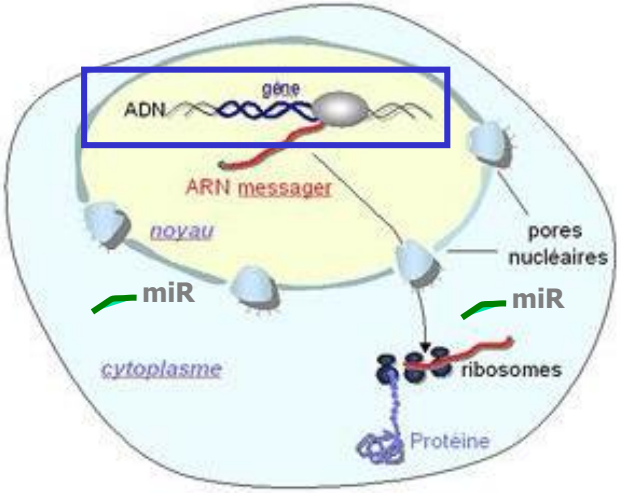
    ⇒ analyse du transcriptome ARN → étude transcriptomique

    ⇒ analyse du protéome protéine → étude protéomique

Le Génome

- Ensemble du matériel génétique d'un individu
- Le même pour toutes les cellules d'un même organisme
- ADN codant / non codant
  - ✓ 80 % du génome humain est fonctionnel
  - ✓ Par la structure 3D de l'ADN, une région n codante peut avoir une action régulatrice
- Altérations
  - ✓ Mutations
  - ✓ Translocations
  - ✓ Amplification ou délétion
  - ✓ Méthylation

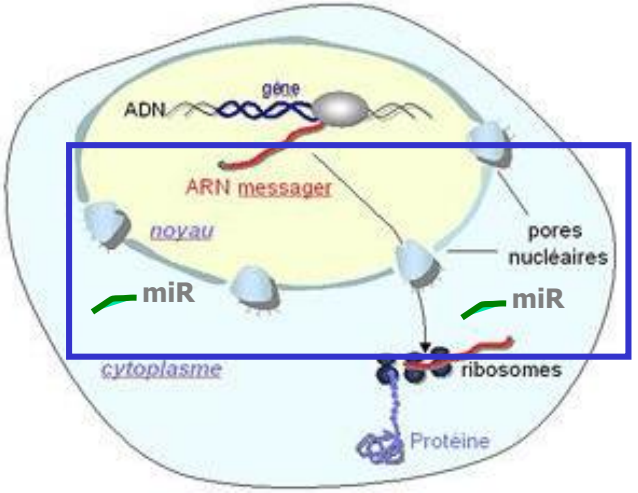
Du gène à la protéine



Le Transcriptome

- Ensemble des ARN issus de la transcription
- Variabilité du transcriptome
- ARN codant (ARNm) / non codant (microARN)
- Altérations
  - ✓ Sur-expression
  - ✓ Sous-expression

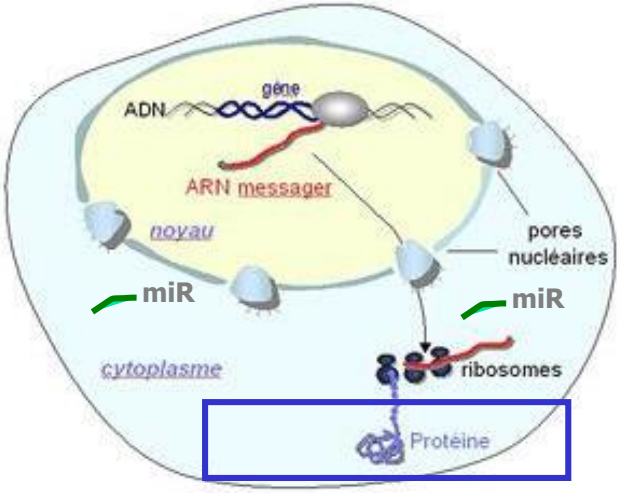
Du gène à la protéine



Le Protéome

- Ensemble des protéines exprimées par un génome, plus particulièrement celui d'une cellule ou d'un tissu
- Etapes
  - ✓ Transcription
  - ✓ Traduction
  - ✓ Modifications post-traductionnelles
- Altérations
  - ✓ Sur-expression
  - ✓ Sous-expression

Du gène à la protéine



Le Métabolome

- Déclinaison des 3 concepts précédents au niveau des métabolites

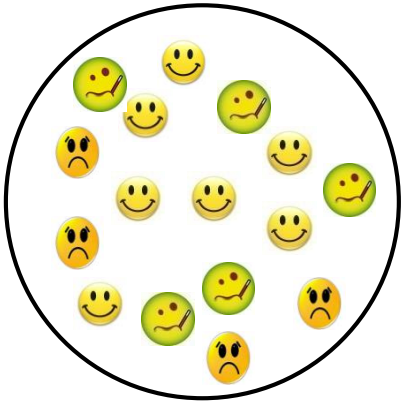
- Etude de l'ensemble des métabolites (sucres, acides aminés, acides gras,…) présents dans une cellule, un organe ou un organisme

Approches « omiques »

- Permettent d'aborder la complexité du vivant dans son ensemble

- Particulièrement utiles pour mieux connaître les maladies héréditaires, adapter les traitements au profil génétique

« Médecine Personnalisée » (1)



Diagnostique de pathologie

Pratique Médicale Actuelle

Molécule X

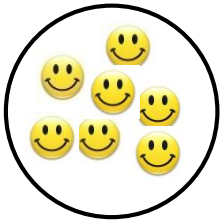
Diagnostique de pathologie

+

Analyse Biomarqueurs

Signatures

Répondeurs tolérants



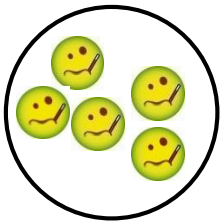
Molécule X

Non Répondeurs

Molécule Y

Nouveau dosage Molécule X

Répondeurs non-tolérants

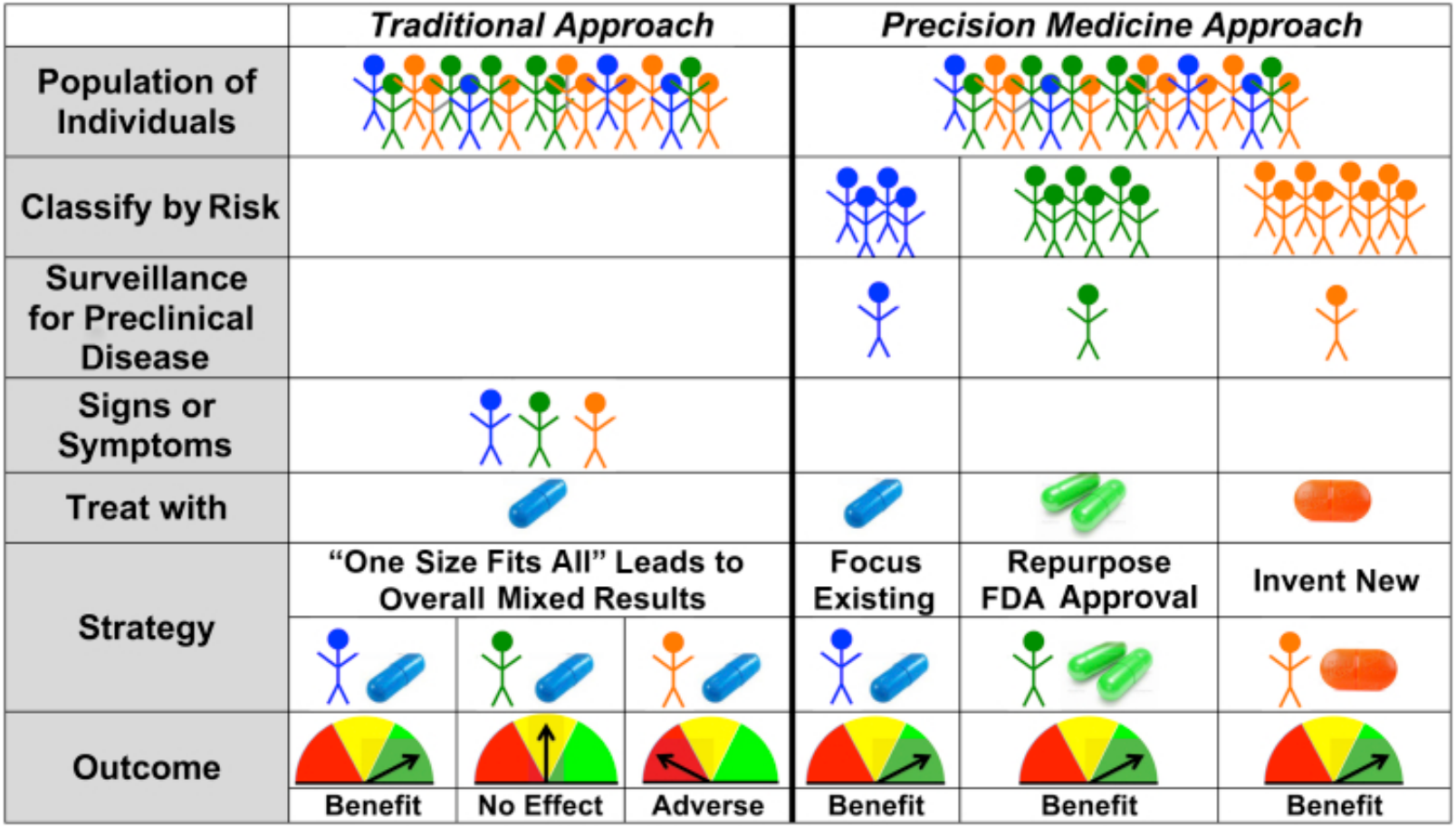


Molécule Z

Nouveau dosage Molécule X

Huriez A, TcLand Expression, LEEM. Bordeaux 2010

« Médecine Personnalisée » (2)



Cholerton B, et al. Precision Medicine: Clarity for the Complexity of Dementia. *Am J Pathol* 2016;186(3):500-6.

Procédure en Etapes

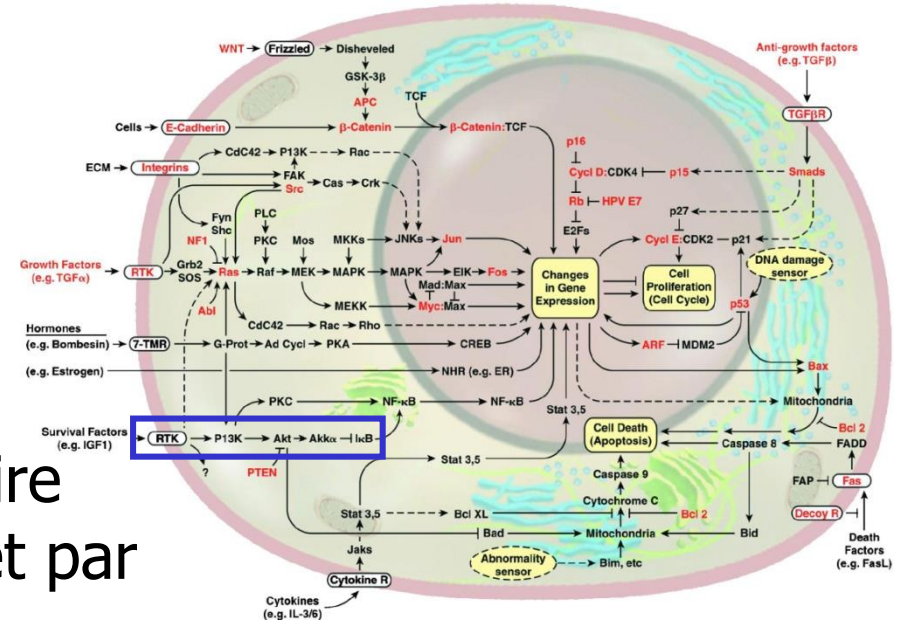
- Etudes d'identification
  - ✓ Biomarqueurs candidats
  - ✓ Estimation des ampleurs d'effets
- Etudes de validation
  - ✓ Biomarqueurs confirmés
  - ✓ Ré-estimation des ampleurs d'effets
- Objectifs
  - ✓ Prendre en compte les sources de variabilités (multiples), les risques de surestimation de l'ampleur de l'effet (biais d'optimisme), de fausse identification,…

Identification de Biomarqueurs (1)

Approche gène candidat

- Classiquement
  - ✓ « gène-à-gène »
  - ✓ « protéine-à-protéine »
  - ✓ « tissu-à-tissu »
- Un seul paramètre moléculaire sélectionné par échantillon et par expérience
- Retombées cliniques…

  mais progrès insuffisants

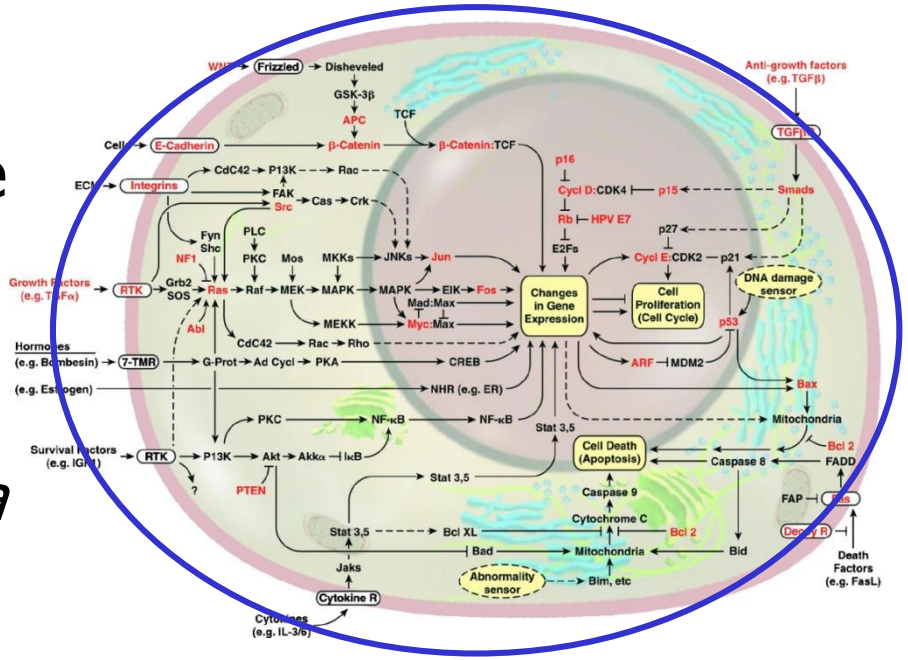


Hanahan D, Weinbegr RA. Cell 2000;100(1):57-70.

Identification de Biomarqueurs (2)

Approche haut débit

- Typage moléculaire à grande échelle
- Analyse d'un très grand nombre de paramètres moléculaires sans sélection *a priori*
- Adapté à la complexité biologique



Hanahan D, Weinbegr RA. Cell 2000;100(1):57-70.

⇒ Puces ADN et ARN

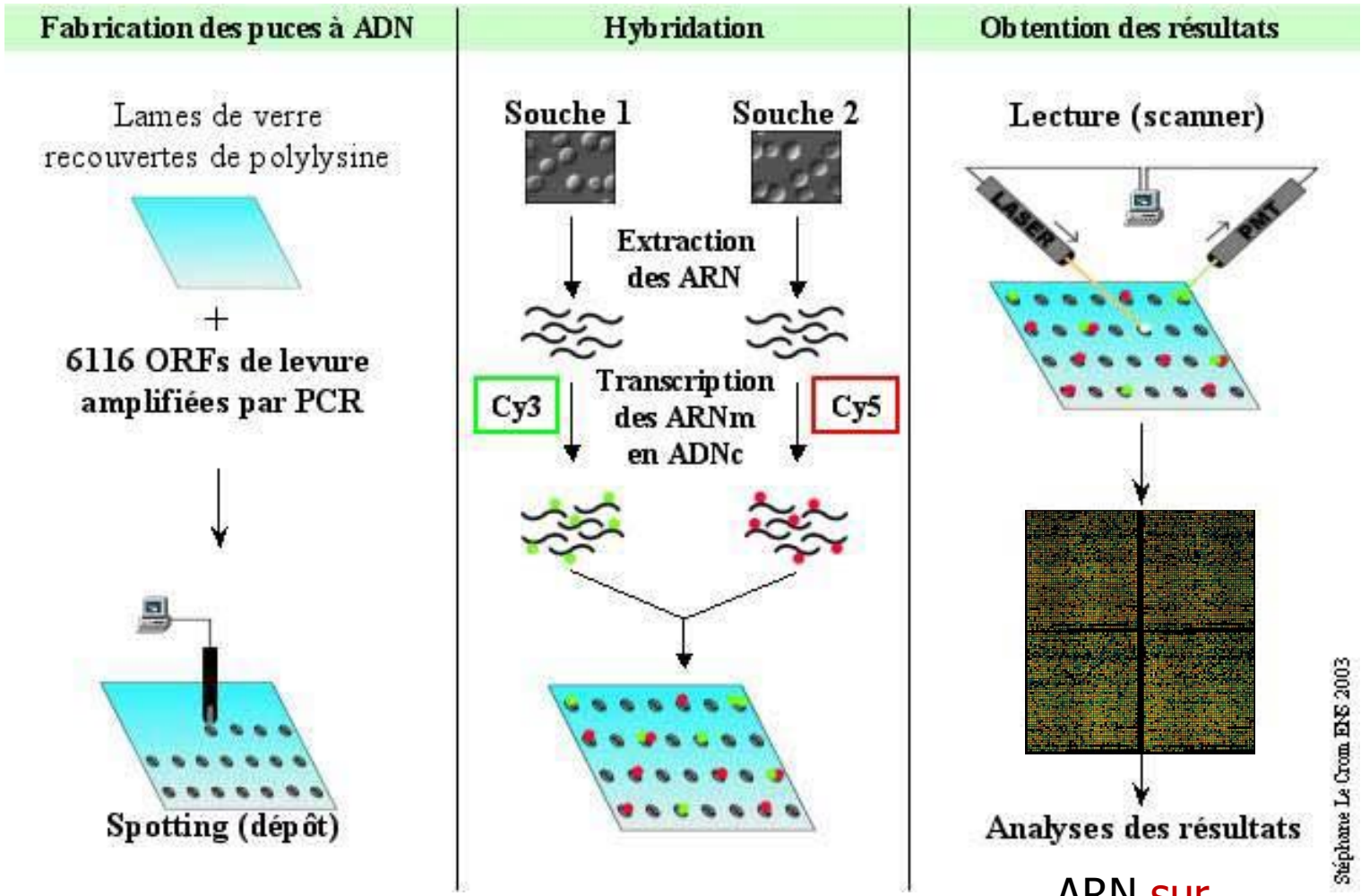
Mesure du Transcriptome par Biopuces (1)

- Puces à ADN (microarrays)
- Analyse des séquences d'ADN ou d'ARN avec un très haut débit
- But
  - ✓ Détecter, quantifier chacune des différentes espèces d'ARNm présents dans la préparation
  - ✓ Analyser le transcriptome c'est identifier, à un temps *t*, et/ou dans une condition donnée, les séquences codantes du génome effectivement exprimées (transcrits, ou ARNm)

Mesure du Transcriptome par Biopuces (2)

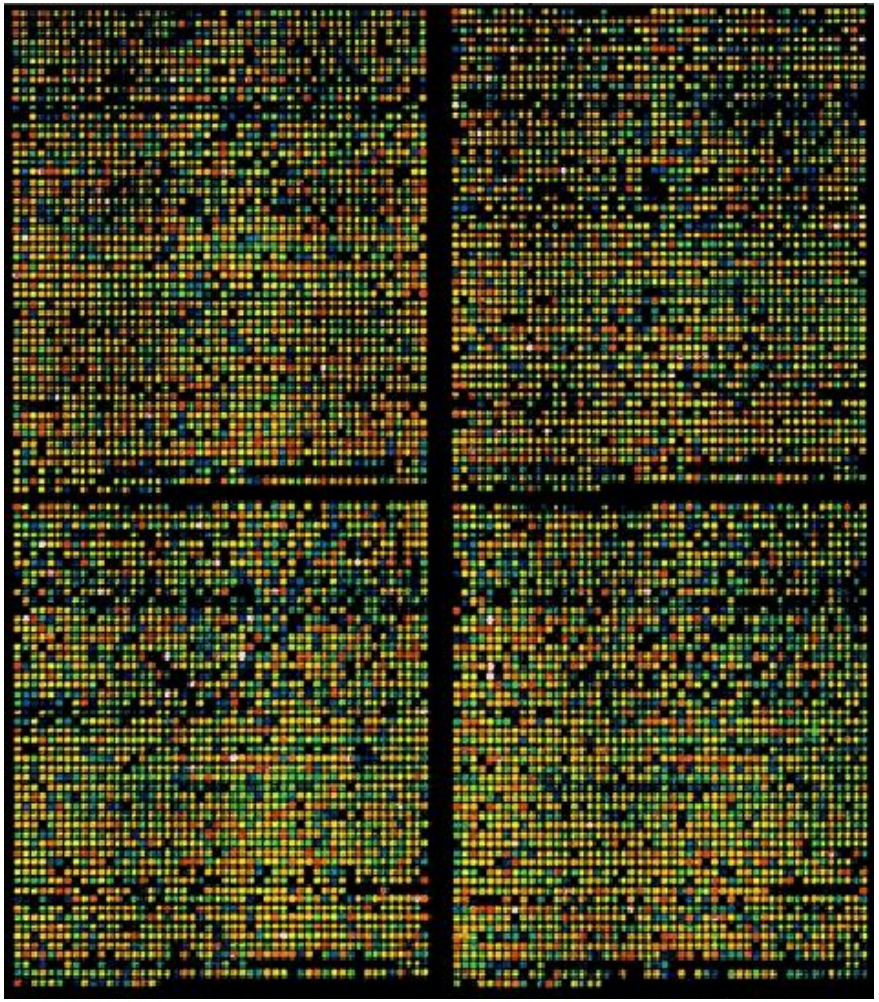
- Principe
  - ✓ Puces sur lesquelles sont fixées des sondes
  - ✓ On y applique l'ADN ou l'ARN d'un sujet pour en connaître son transcriptome
  - ✓ Exploitent la propriété d'hybridation des séquences nucléiques
- Détectent, quantifient simultanément plusieurs centaines de milliers de séquences
- Différents types de puces
  - ✓ Expression – Cible = transcrits
  - ✓ Génotypage – Cible = SNP (Single Nucleotide Polymorphism)
  - ✓ …

Mesure du Transcriptome par Biopuces (3)



ARN sur
ou sous-exprimé ?

Mesure du Transcriptome par Biopuces (4)



ARN sur
ou sous-exprimé ?

Analyse des Données

- Nombreuses sources d'erreur et de variabilité
  - ✓ Variabilité biologique
    - Population de cellules ou patients/tissus différents
  - ✓ Variabilité technique
    - Etape d'amplification
    - Incorporation de fluorochromes
    - Bruit (artefacts, bruit de fond)
    - Données manquantes
  - ✓ Erreur (spot sur la puce, plaque,…)

⇒ Augmenter les réplicats biologiques et techniques

Traitement d'Images

- Quantification
  - ✓ Intensités en rouge et vert
  - ✓ Nécessité de corriger le bruit de fond
  - ✓ …

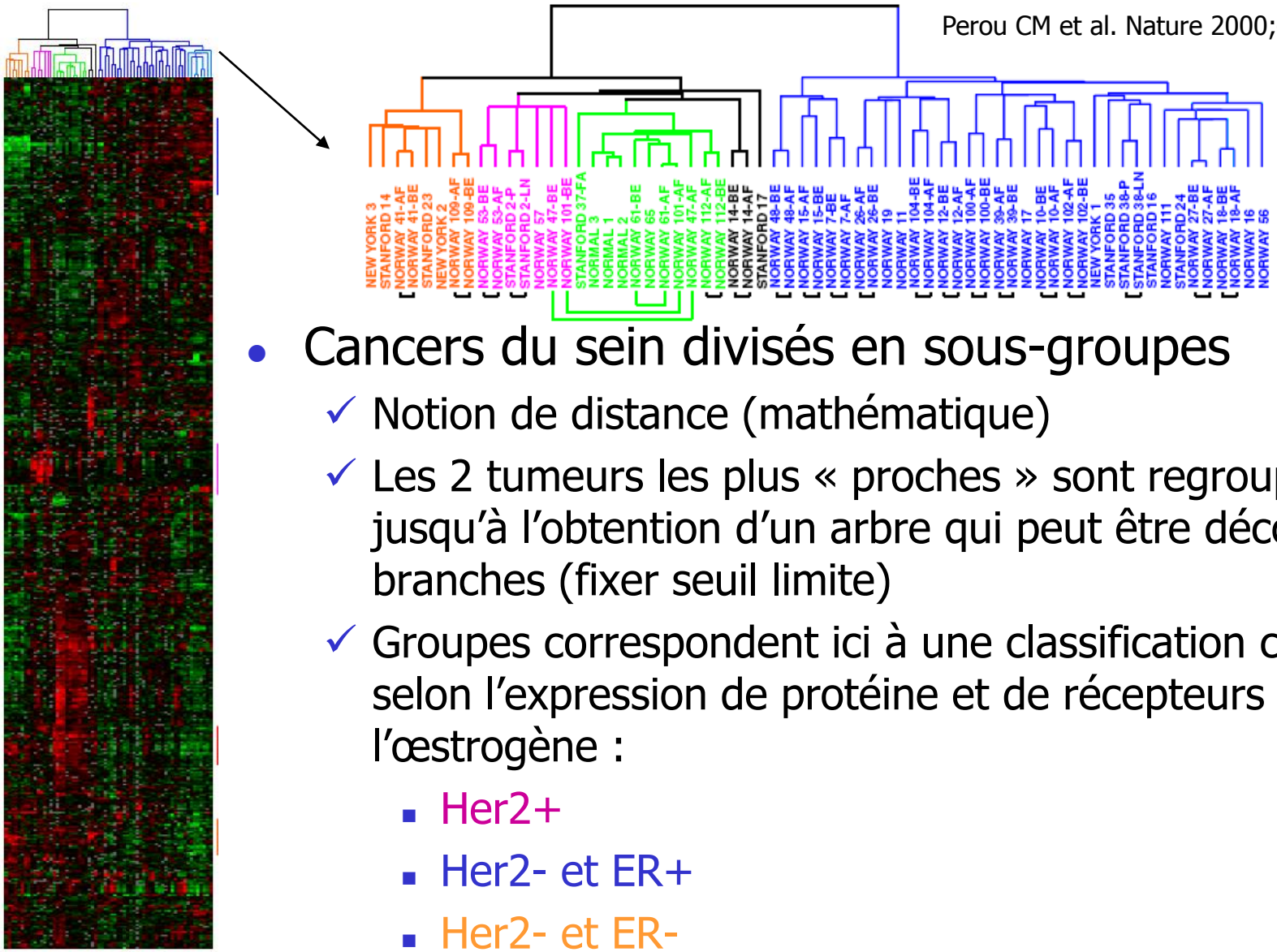
⇒Extraction des gènes d'intérêts

Analyse des Données

- Identifier les gènes ayant une variation significative par rapport à l'expérience
- Trouver les gènes ayant une variation recherchée
- Regrouper les gènes ayant des variations similaires
- Trouver les gènes expliquant les variations d'autres gènes

$\Rightarrow$ Identification d'une signature moléculaire

Exemple : Signature Moléculaire Cancer du Sein



Perou CM et al. Nature 2000;406(6797):747-52.

- Cancers du sein divisés en sous-groupes
  - ✓ Notion de distance (mathématique)
  - ✓ Les 2 tumeurs les plus « proches » sont regroupées jusqu'à l'obtention d'un arbre qui peut être découpé en branches (fixer seuil limite)
  - ✓ Groupes correspondent ici à une classification connue selon l'expression de protéine et de récepteurs à l'œstrogène :
    - Her2+
    - Her2- et ER+
    - Her2- et ER-

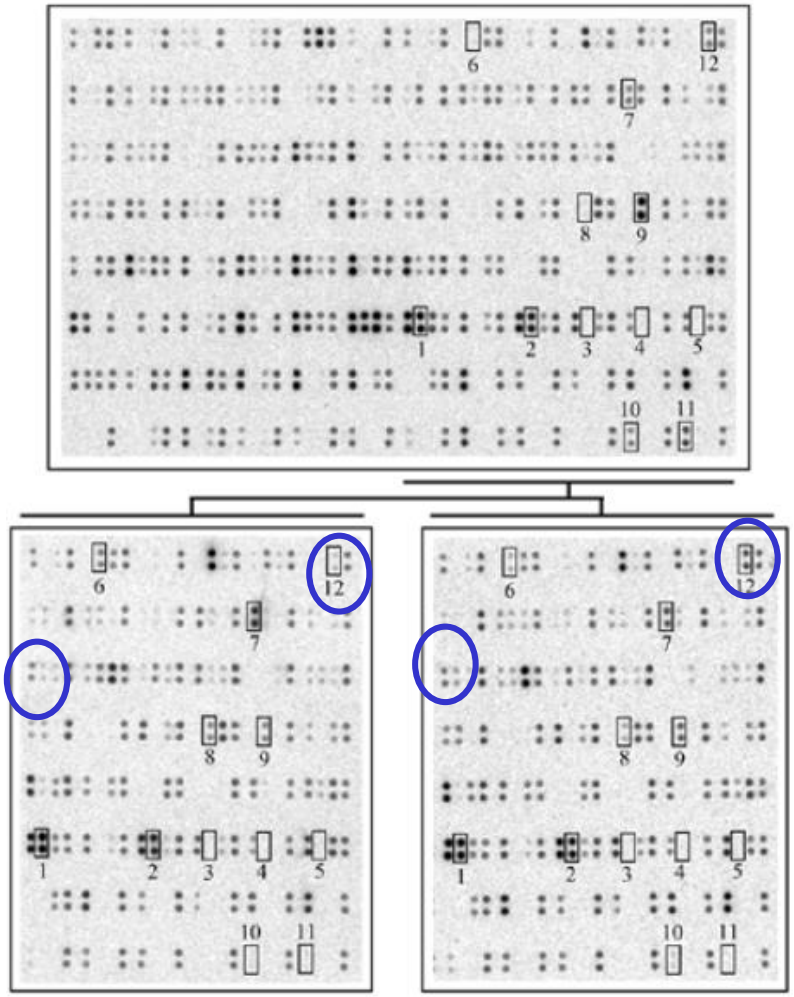
Approches Analyse

- Approche différentielle
  - ✓ Comparaisons gène par gène
  - ✓ Ne prend pas en compte la corrélation entre les gènes

- Approche profil d'expression
  - ✓ Prise en compte dans la globalité

Approche Différentielle : Exploration (1)

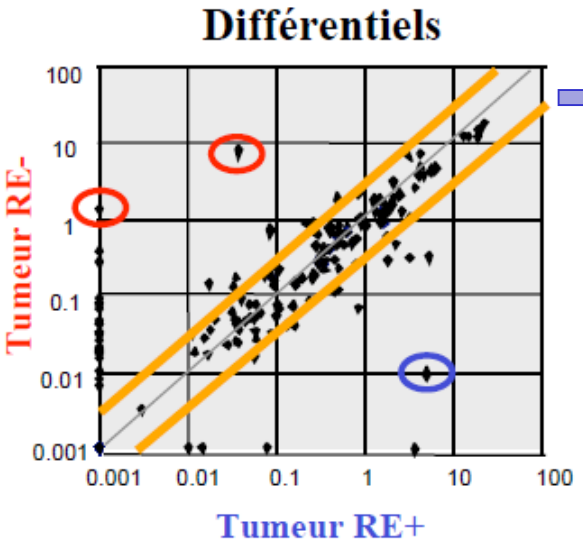
Sein Normal



Tumeur RE- Tumeur RE+

Bertucci F, et all. Human Molecular Genetics 2000;9(20):2981-91.

Approche Différentielle : Exploration (2)



	Gene/protein identity	Gene symbol	ER⁺:ER⁻
	GATA-binding protein 3	GATA3	28.6
	Granzyme A	GZMA	5.7
	MYB proto-oncogene	MYB	3.4
	KIAA1075 protein	KIAA1075	3.3
	Stromelysin 3	STMY3	3.1
	Macrophage stimulating 1	MST1	2.8
	Cellular retinoic acid binding protein 2	CRABP2	2.7
	X-box binding protein 1	XBP1	2.7
205314	*Tumor protein p53*	TP53	2.5
126233	*Insulin-like growth factor 2*	IGF2	2.4
66322	*CD3G antigen, gamma*	CD3G	0.0
195022	*Interleukin 2 receptor gamma chain*	IL2RG	0.0
111461	*SOX4 protein*	SOX4	0.4
151475	*Epidermal growth factor receptor*	EGFR	0.5
195022	*Interleukin 2 receptor beta chain*	IL2RB	0.5
130788	*Topoisomerase (DNA) II beta* (180 kDa)	TOP2B	0.6
323948	*SOX9 protein*	SOX9	0.6
183641	*S100 calcium-binding protein beta*	S100B	0.6
246620	EST N53133	EST	0.6
231424	*Glutathione S transferase Pi*	GSTP1	0.6

Bertucci F, et all. Human Molecular Genetics 2000;9(20):2981-91.

Approche Différentielle : Analyse

- Analyse statistique bivariée « classique » sur données de densités
  - ✓ Coefficient de corrélation (Pearson, Spearman)
  - ✓ Comparaisons (test t de Student, test des rangs de Mann-Whitney

- Hypothèse nulle $H_0 : m_1 = m_2$

Décision

- Hypothèse nulle $H_0 : m_1 = m_2$

		Réalité		
		H_0 vraie	H_0 fausse	
Conclusion	Acceptation H_0	VN	FN	m - R
	Rejet H_0	FP	VP	R
		m_0	m_1	m

  - FP : nb de tests faussement positifs
  - R : nb total de fois où H_0 est rejetée
  - Erreur de type I : faux positifs
  - Erreur de type II : faux négatifs

Problèmes Spécifiques

- Données de « grande dimension »
  - ✓ Nombre de sujet n : petit
    - $\Rightarrow$ Choix de la statistique de test crucial
  - ✓ Nombre de gène m : grand
    - $\Rightarrow$ Problèmes de tests statistiques multiples
    - Un test par gène $\rightarrow$ m p-values (p_1,…, p_m)
    - Avec un seuil α fixé, on attend mα faux positifs
    - $\rightarrow$ nombre d'erreurs dépend du nombre de tests
    - $\rightarrow$ Nécessité d'ajuster le seuil α en fonction de m

Erreur de Type I et Nombre de Tests

Nombre de tests	Erreur de type I
1	0,05
2	0,08
3	0,11
4	0,13
5	0,14
20	0,25
50	0,32
100	0,37
1 000	0,53
Infini	1,00

Sur 20 tests, effectués au seuil α = 0,05, 25% différences considérées à tort comme statistiquement significatives

Contrôle de l'Erreur de Type I (1)

		Réalité		
		H_0 vraie	H_0 fausse	
Conclusion	Acceptation H_0	VN	FN	m - R
	Rejet H_0	FP	VP	R
		m_0	m_1	m

- Notations
  - ✓ FP : nb de tests faussement positifs
  - ✓ R : nb total de fois où H_0 est rejetée
- Mesures de risque
  - Taux d'erreur global (Family-Wise Error Rate) : FWER = P(FP > 0) (au moins 1 test signif. à tort)
  - Taux de fausses découvertes (False Discovery Rate) : FDR = E(FP/R) si R>0, 0 sinon (% de tests signif. à tort)

Contrôle de l'Erreur de Type I (2)

		Réalité		
		H_0 vraie	H_0 fausse	
Conclusion	Acceptation H_0	VN	FN	m - R
	Rejet H_0	FP	VP	R
		m_0	m_1	m

- Contrôle du FDR
  - ✓ FDR < seuil α^* (fixé a priori)
  - ✓ q-value : proportion attendue de FP parmi toutes les combinaisons possibles, sous H_0
- Contrôle du FWER au niveau α

Auteur	Seuil	Type de contrôle
Sidak	$1-(1-\alpha)^{1/m}$	Suppose indépendance des gènes calculé sous H_0
Bonferroni	α/m	Pas d'hypothèse d'indépendance des gènes Très conservateur

Contrôle de l'Erreur de Type II

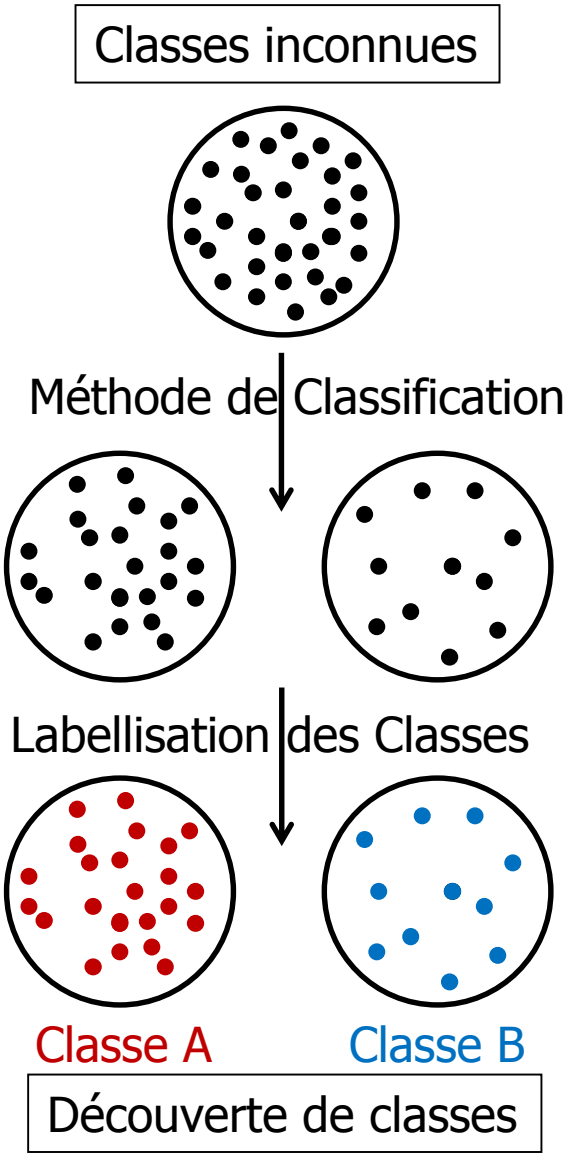
- Puissance dans le cadre de tests multiples
  - ✓ VP : vrais positifs
  - ✓ m_1 : nombre de fois où H_0 rejetée de manière correcte

$$\frac{E(VP)}{m_1}$$

Approche profil d'expression

- Vise à identifier une « signature moléculaire »
- Caractérisation de nouvelles classes d'échantillons
  - ✓ Approches non supervisées
    - Découverte de classes
- Caractérisation de classes phénotypiques connues et prédiction d'appartenance
  - ✓ Approches supervisées
    - Prédiction de classes

Approches Non Supervisées (1)

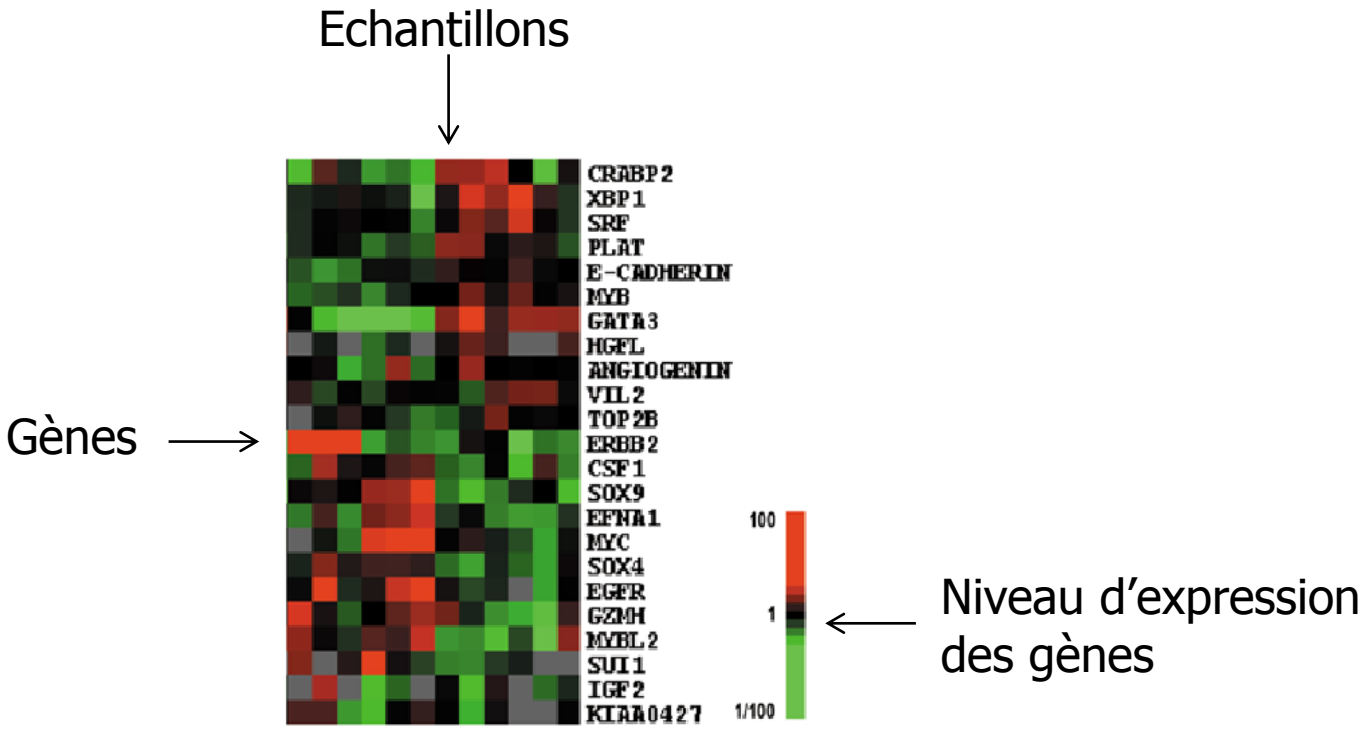


Approches Non Supervisées (2)

- Vise à rechercher des corrélations
  - ✓ Classes de tumeurs et facteurs histo-cliniques
  - ✓ Classes de gènes et fonction
  - ✓ …
- Classification hiérarchique
  - ✓ Calcul d'une matrice de similarité entre gènes
  - ✓ Regroupement des gènes « les plus proches »
  - ✓ Remplacement des deux gènes par un « gène moyen »
  - ✓ Processus itératif
  - ⇒ Arbre phylogénique représentant une hiérarchie de groupes de gènes selon leur similarité

Approches Non Supervisées (3)

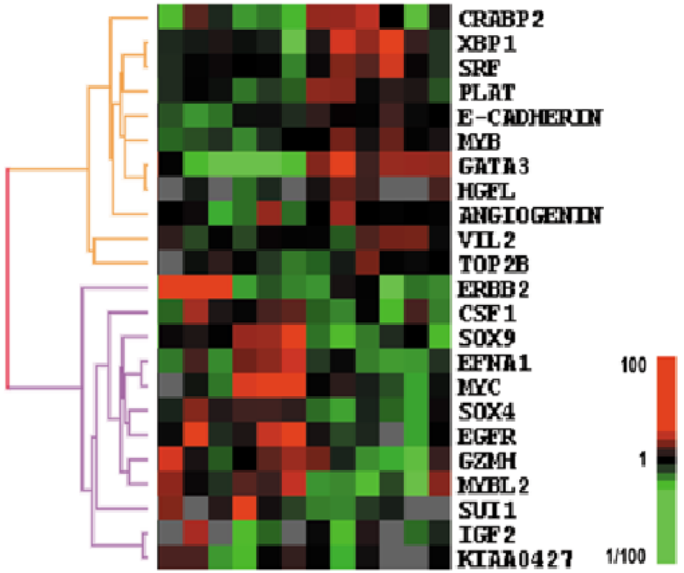
Bertucci F, et all. Human Molecular Genetics 2000;9(20):2981-91.



Approches Non Supervisées (4)

Bertucci F, et all. Human Molecular Genetics 2000;9(20):2981-91.

Classification des gènes selon leur similarité →



Approches Non Supervisées (5)

Bertucci F, et all. Human Molecular Genetics 2000;9(20):2981-91.

Classification des échantillon selon leur similarité →

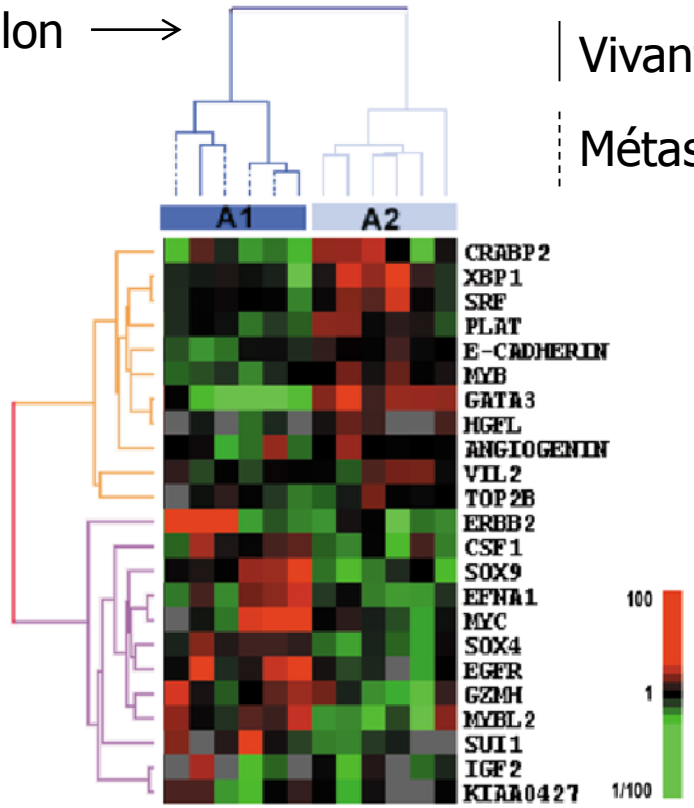
Patients Groupe A1 ou A2

Vivants

Métastase, décès

ET

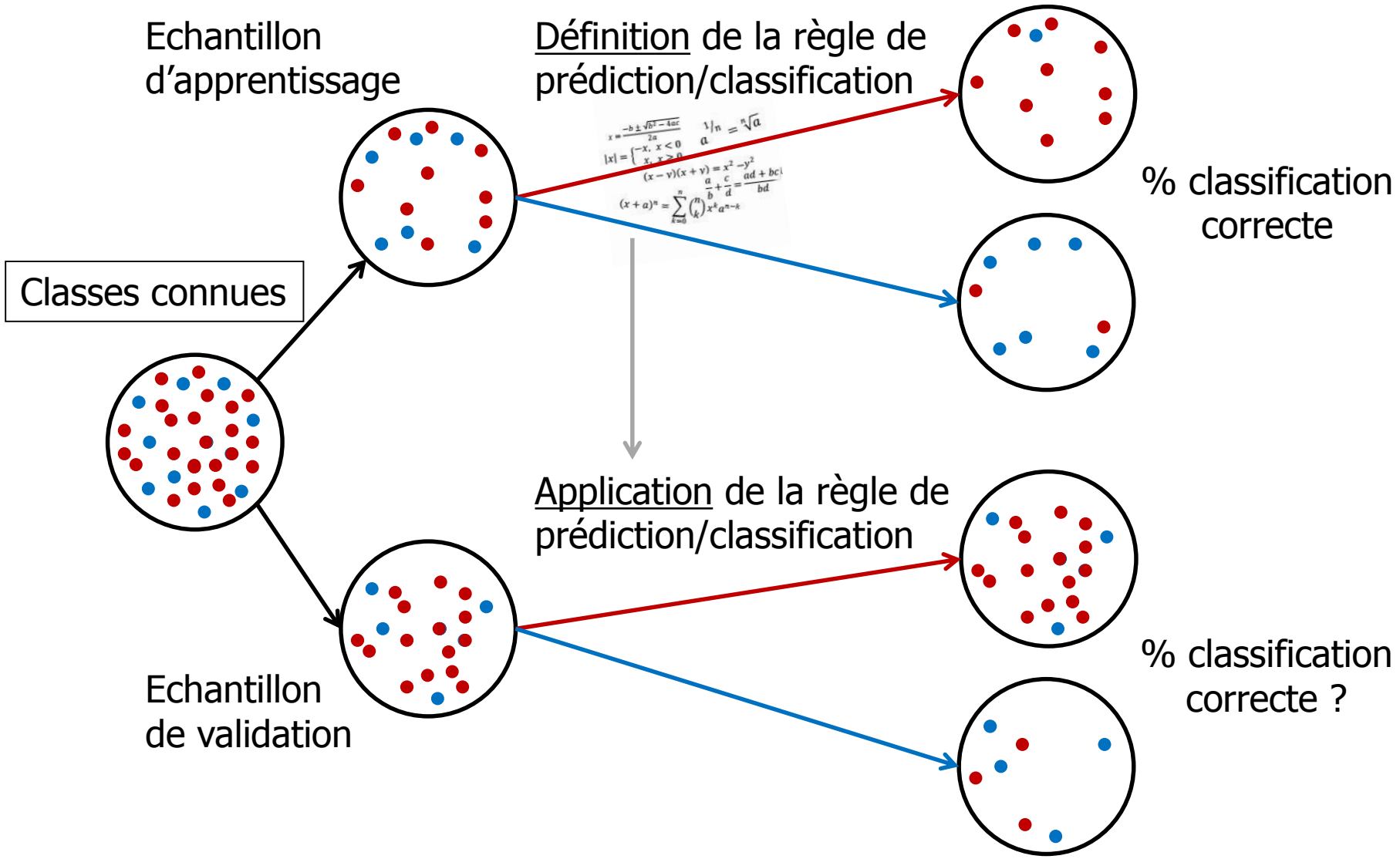
Classification des gènes selon leur similarité →



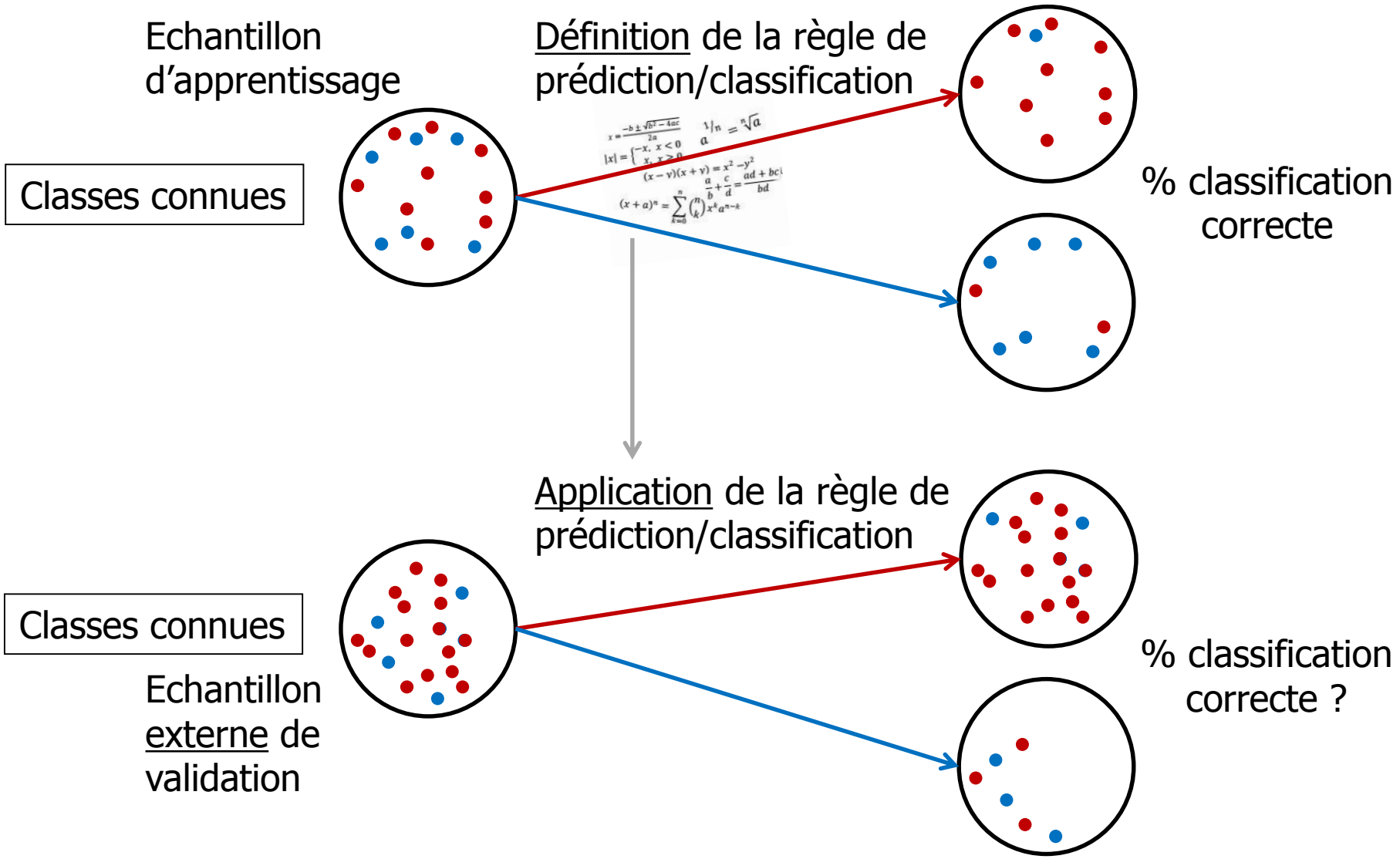
Approches Supervisées

- Vise à prédire une variable clinique ou biologique à partir de données « omiques »
  - ✓ Distinction entre individus sains et atteints d'une maladie
  - ✓ Prédiction de la réponse thérapeutique ou la survie
- Principe
  - ✓ Apprentissage d'une règle sur un échantillon d'entrainement
  - ✓ Validation de cette règle sur un échantillon indépendant
    - Validation interne : données non utilisées de l'échantillon apprentissage
    - Validation externe : données provenant d'une autre étude

Approches Supervisées – Validation Interne



Approches Supervisées – Validation Externe



Approches Supervisées

- Propriétés souhaitables d'un prédicteur/classifieur
  - ✓ Performances
  - ✓ Stabilités/robustesse
  - ✓ Interprétabilité
- Caractéristiques des données
  - ✓ Nombre de variables : grand
    - ⇒ Besoin de sélectionner des variables
  - ✓ Nombre de sujet n : petit
    - ⇒ Nécessité de méthodes de validations appropriées
  - ✓ Structure
    - ⇒ Prise en compte des dépendances entre gènes

Approches Supervisées - Méthodes

- Classification
  - ✓ Analyse discriminante, en composantes principales
  - ✓ k plus proches voisins
  - ✓ Arbres de décision (CART), forêts aléatoires
  - ✓ …
- Régression
  - ✓ Linéaire, logistiques,…
- Réduction de dimension pour la classification
  - ✓ Construction de nouvelles variables synthétiques
  - ✓ Sélection de variables avant / pendant la classification

Exemple (1)

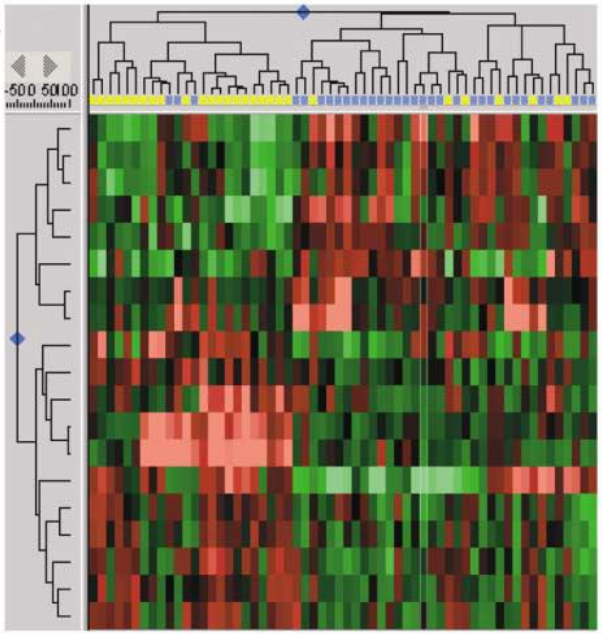
Ma XJ, et all. Cancer Cell 2004;5(6):607-163

- Cancer du sein traité RE+
- Echantillon 1 - Apprentissage
  - ✓ 22 000 gènes, 60 tumeurs localisées RE+
  - ✓ Signature génétique de la survie sans rechute
    - Ratio HOXB13 / IL17BR
    - « élevé » versus « faible » déterminé par régression logistique ajustée sur la taille de la tumeur, l'expression de récepteurs à l'œstrogène et la progestérone
    - Discrimine la survie sans rechute

Exemple (2)

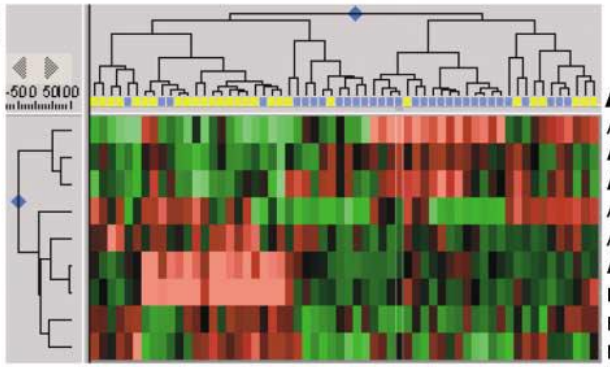
Ma XJ, et all. Cancer Cell 2004;5(6):607-163

Echantillon 1



Accession	P	Description
AF208111	**9.9e-5**	**IL17BR | interleukin 17B receptor**
AI240933	**2.4e-5**	**EST**
AJ272267	0.0002	CHDH | choline dehydrogenase
AW613732	0.0006	Homo sapiens cDNA FLJ31137 fis, clone I.
AK027250	0.001	Homo sapiens cDNA: FLJ23597 fis, clone L.
AL050227	0.0006	Homo sapiens mRNA; cDNA DKFZp586M0.
N30081	0.0002	ESTs, Weakly similar to I38022 hypothetical.
X81896	0.0009	C11orf25 | chromosome 11 open reading fr.
M92432	0.0005	GUCY2D | guanylate cyclase 2D, membrane.
AB000520	9.6e-5	APS | adaptor protein with pleckstrin homol.
BC004960	0.0009	MGC10955 | hypothetical protein MGC10955
AI700363	**0.0004**	**HOXB13 | homeo box B13**
BC007092	**0.0004**	**HOXB13 | homeo box B13**
AL117406	0.0004	ABCC11 | ATP-binding cassette, sub-family.
BC007783	0.0009	SCYA3 | small inducible cytokine A3
AW006861	2.0e-5	SCYA4 | small inducible cytokine A4
X59770	8.1e-5	IL1R2 | interleukin 1 receptor, type II
AI820604	0.0002	ESTs
AI087057	0.0002	DOK2 | docking protein 2, 56kD

Echantillon 2



Accession	P	Description
AI688494	0.0005	FLJ13189 | hypothetical protein FLJ13189
AI240933	**0.0005**	**EST**
AF208111	**0.0004**	**IL17BR | interleukin 17B receptor**
AF033199	0.0004	ZNF204 | zinc finger protein 204
AL157459	0.0007	Homo sapiens mRNA; cDNA DKFZp434B0.
AI700363	**0.0003**	**HOXB13 | homeo box B13**
BC007092	**2.7e-5**	**HOXB13 | homeo box B13**
BC002480	0.0008	FLJ13352 | hypothetical protein FLJ13352
NM_014298	0.0004	QPRT | quinolinate phosphoribosyltransfera.

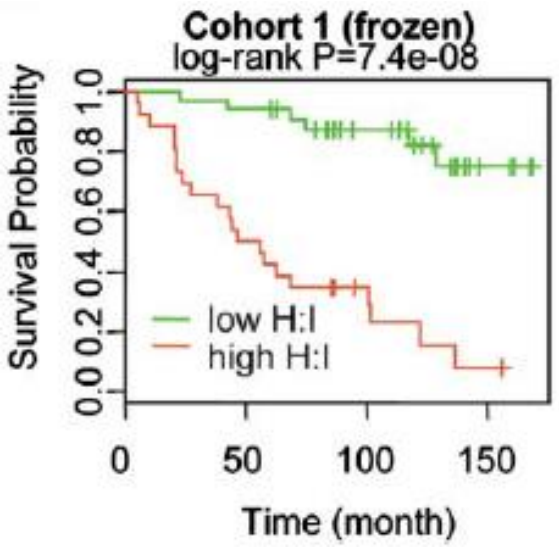
Exemple (3)

Ma XJ, et all. Cancer Cell 2004;5(6):607-163

Multivariate model

Predictors	Odds ratio	95% CI	p value
Tumor size	1.5	0.7–3.5	0.3289
PGR	0.8	0.3–1.8	0.5600
ERBB2	1.7	0.8–3.8	0.1620
HOXB13:IL17BR	7.3	2.1–26.3	0.0022

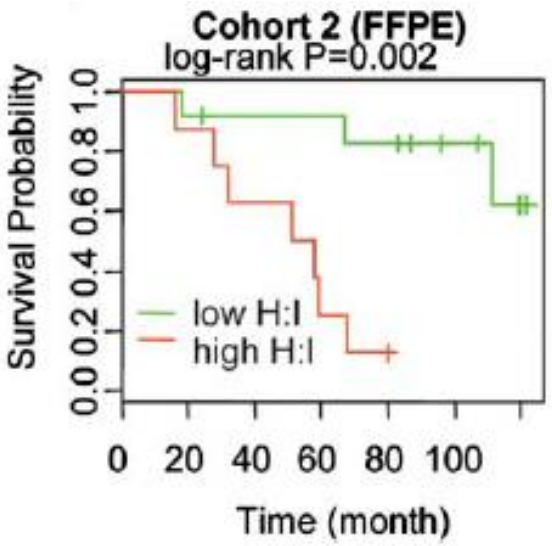
All predictors are continuous variables. Gene expression values were from microarray measurements. Odds ratio is the interquartile odds ratio, based on the difference of a predictor from its lower quartile (0.25) to its upper quartile (0.75); CI, confidence interval.



Exemple (4)

Ma XJ, et all. Cancer Cell 2004;5(6):607-163

- Echantillon 2 - Validation
  - ✓ 20 tumeurs localisées RE+
  - ✓ Classification en Ratio HOXB13 / IL17BR « élevé » versus « faible » avec le seuil identifié sur l'échantillon d'apprentissage
  - ✓ Etude de la survie sans rechute



Conclusion

- Développement considérable des techniques d'analyse de la biologie moderne
- Echantillons biologiques doivent être de qualité
- Contrôle qualité important sur toute la chaîne de production
- Complexité de l'analyse statistique pour prendre en compte
  - ✓ La grande dimension des données
  - ✓ Leur structure corrélée
  - ✓ Le maximum d'information, omique et en même temps clinique, biologiques autres,…