institutCurie
Département de Biostatistique
Unité INSERM U900

Analyse des données d'expression de gènes issues de puces à ADN : démarche et outils statistiques

Eléonore GRAVIER

27 Mai 2010

15ème Journées des Statisticiens des Centres de Lutte contre le Cancer
4ème Conférence Francophone d'Epidémiologie Clinique - congrès thématique de l'ADELF
1er congrès joint

Contexte

- Puce « transcriptome » : mesure simultanée du niveau d'expression de dizaines de milliers de gènes d'un échantillon biologique

- **Haute dimensionnalité** des données issues de puces à ADN : problèmes méthodologiques et d'interprétation

- **Abondance et spécificité des méthodes/outils** statistiques

→ **Choix de démarche/méthodes/outils statistiques à utiliser non trivial** pour des statisticiens non spécialistes

institutCurie

Contexte

- Biostatisticiens et Bioinformaticiens de l'Institut Curie proposent :
  - **Démarche statistique**
  - Package R **EMA** (Easy Microarray data Analysis)

Données et Objectif

- Kenneth R. Hess, Keith Anderson et al., JCO, 2006 [DLD30]

- 133 tumeurs du sein (82 + 51)
  - **Réponse Complète pathologique (pCR)** de la tumeur après chimiothérapie néoadjuvante
  - Analyse de l'**expression des gènes sur puce** Affymetrix U133 A (~22000 gènes)

→ Prédire la réponse complète pathologique d'une tumeur à partir de son profil d'expression de gènes (variables explicatives=expression des 22000 gènes)

Contrôle Qualité et prétraitement des données

- Motivations
  - **Sources de variabilité** à chaque étape expérimentale, qui se confondent avec le signal biologique à étudier

- Objectif
  - **Contrôle Qualité** : **Évaluer la qualité des données** afin de supprimer/réhybrider des puces « problématiques »
  - **Normalisation** : **S'affranchir de ces variabilités expérimentales** pour débruiter le signal biologique et rendre les puces comparables entre elles
  - **Filtrage** : **Supprimer les gènes très peu exprimés** le long des puces afin de débruiter les données et d'augmenter la puissance statistique

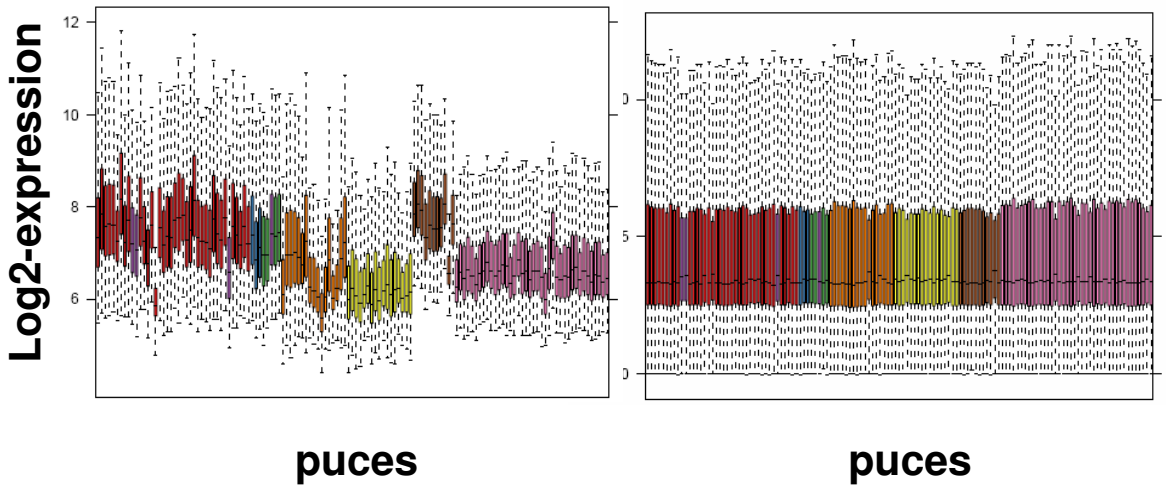
Contrôle Qualité et prétraitement des données

- Contrôle qualité :

Package **arrayQualityMetrics**
(Kauffman et al., Bioinformatics 2009)

- Normalisation GCRMA
(Wu et al, JASA 2004)

→ 22215 gènes



Avant normalisation

Après normalisation

- Filtrage

→ 14125 gènes

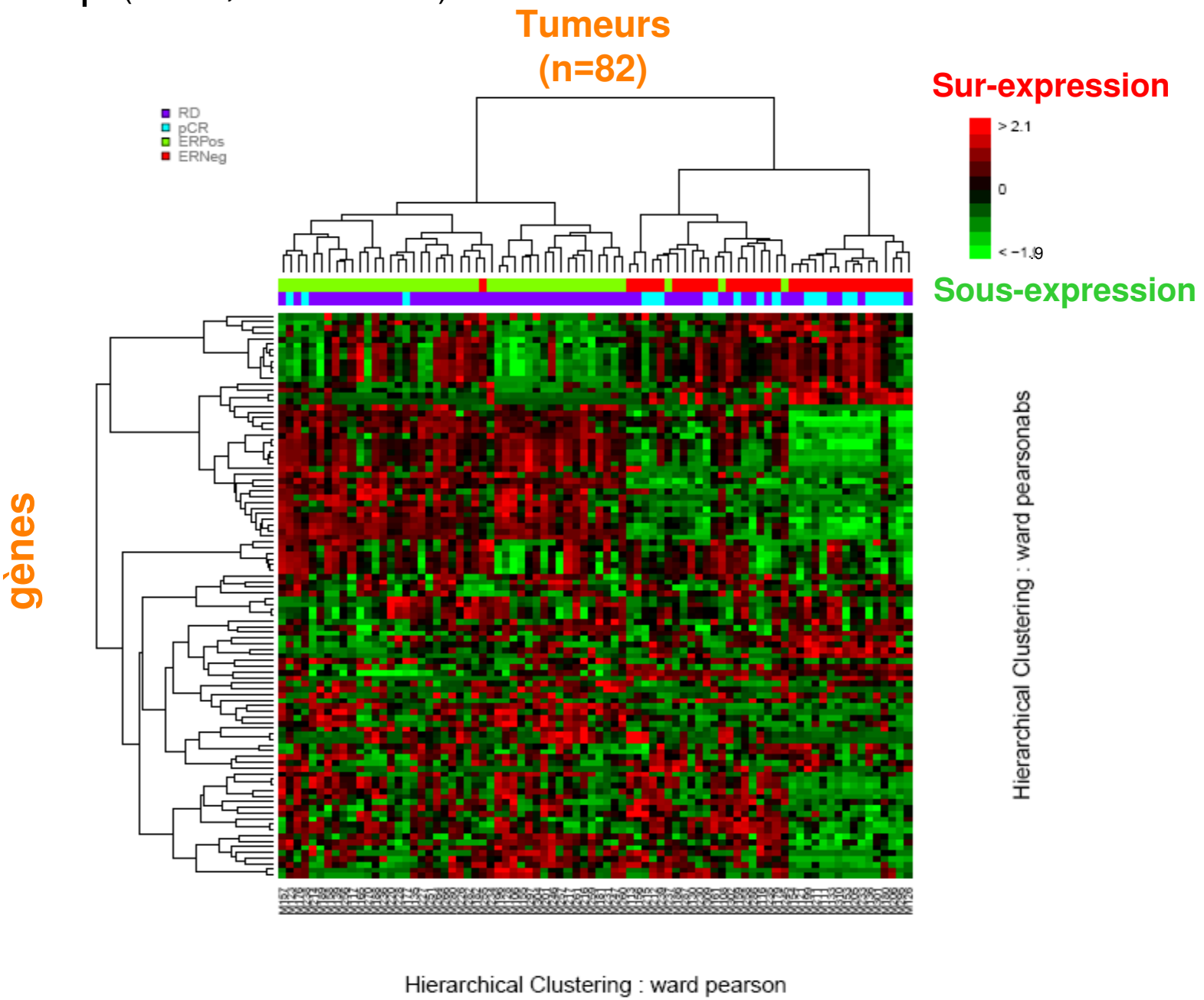
Analyses non supervisées

- Objectif
  - **Décrire les données** indépendamment de toute connaissance *a priori*

  →
  - **Contrôle Qualité** (détection d'outliers, de biais)
  - Identification de groupes de **tumeurs biologiquement homogènes**

- Principe de la classification ascendante hiérarchique (Johnson, 1967)
  - Regroupement successif des tumeurs/gènes aux **profils d'expression « similaires »**

- Heatmap (Eisen, PNAS 1998)



Analyse Différentielle

- Objectif
  - **Identifier les gènes dont l'expression diffère** significativement entre les deux groupes de tumeurs
  - Prendre en compte le problème de **multiplicité des tests**

- Principe Significance Analysis of Microarrays (Tusher PNAS 2001)
  - Estimation du **False Discovery Rate (FDR)** : Espérance du taux de faux positifs parmi les gènes déclarés différentiellement exprimés
  - Utilisation de **permutations aléatoires de la réponse**

Analyse Différentielle

- SAM Wilcoxon
  - **348 gènes différentiellement exprimés** (FDR=0.05)
    - 348*0.05=**18 gènes attendus faux positifs**

institutCurie

Classification supervisée

- Objectif
  - **Prédire la classe d'une tumeur** en fonction de son profil d'expression

- Limites des moindres carrés ordinaires (p>>n)
  - Estimateurs sans biais mais de **grande variance**
  - Risque majeur de **surajustement**

- Principe de la régression pénalisée de type Lasso (Tibshirani, JRSS 1996)
  - Ajout d'une **pénalité sur la norme L_1 des coefficients**

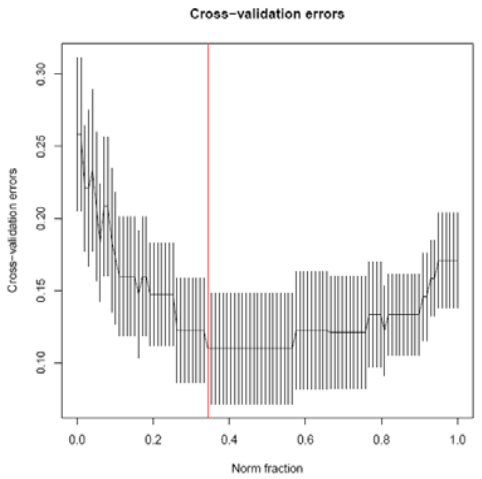
$$\hat{\beta} = \arg\min \left\| y - X\beta \right\|_2^2 \quad tel\ que \sum_{j=1}^{p} \left| \beta_j \right| \leq s\ et\ s \geq 0$$

institutCurie

Classification supervisée

- Lasso : Package R **glmpath**

(Park et al, JRSS 2001)

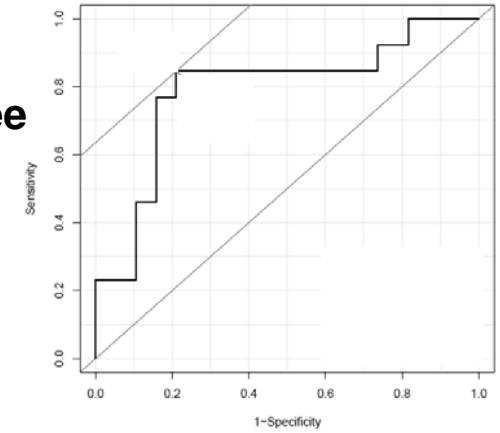


- **Apprentissage (n=82)**
  - Régression **logistique stepwise ascendante** (AIC)

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	28970.4	1526413.4	0.019	0.985
g220	-2363.8	124251.5	-0.019	0.985
g37	454.2	24152.8	0.019	0.985
g67	-951	49901.7	-0.019	0.985
g98	-4211.7	222806.7	-0.019	0.985
g80	-316.3	17822.5	-0.018	0.986

  - **Lasso**
    - Choix du paramètre de pénalité par **validation croisée** (10-folds)
    - **24 gènes** sélectionnés (parmi 348)

$$u = s / \sum_{j=1}^{p} \left| \hat{\beta}_j^{ols} \right|$$



- **Validation (n=51)**
  - Lasso : AUC=**79%**, IC=[0.63;0.94]

institutCurie

Conclusions

- Stratégie d'analyse **validée**, couvrant de **nombreuses problématiques** du domaine de la génomique

- **Implémentation, visualisation et interprétation facilités**

- Package **R EMA + vignette** accessibles gratuitement

http://bioinfo.curie.fr/projects/ema/

Bon point de départ pour des statisticiens non spécialistes des analyses de données de puces d'expression

institutCurie

Collaborations

Unité INSERM U900

« Cancer et génome : Bioinformatique, biostatistique et épidémiologie d'un système complexe »

- Servant Nicolas
- Valet Fabien
- Gestraud Pierre
- Laurent Cécile
- Paccard Caroline
- Biton Anne
- Brito Isabel
- Mandel Jonas
- Asselain Bernard
- Barillot Emmanuel
- Hupé Philippe

institutCurie

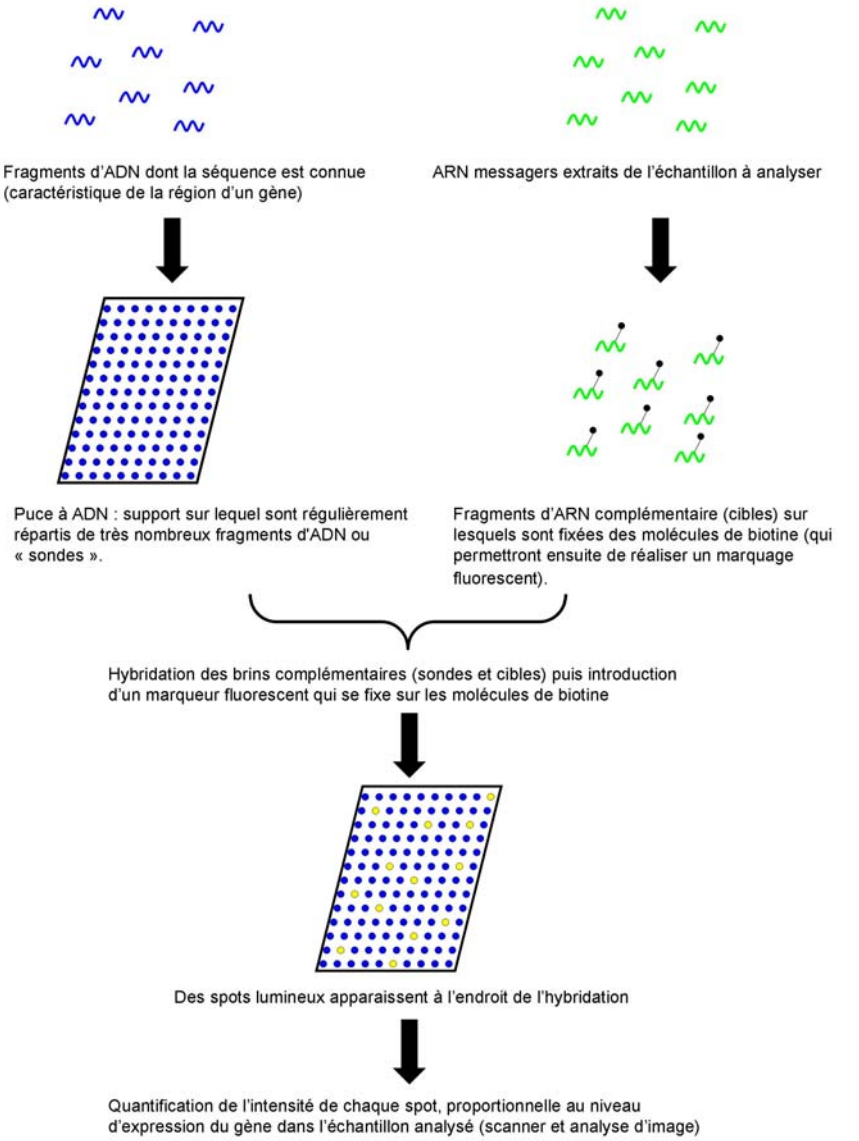
Merci pour votre attention

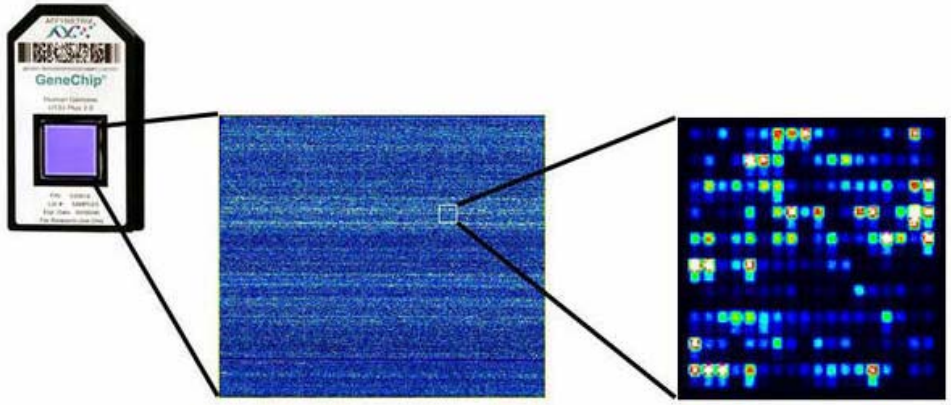
institutCurie

Annexes

institutCurie

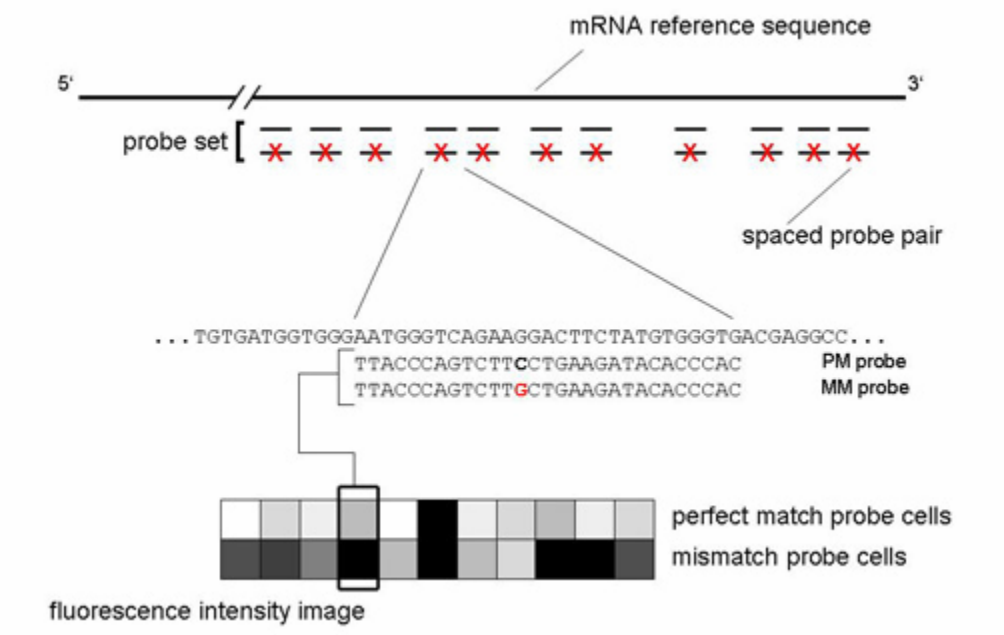
Puce « transcriptome » Affymétrix





Puce « transcriptome » Affymétrix

- Puce U133A
  - ~500 000 sondes
  - 11 à 20 paires de sondes par transcrit
  - ~22000 transcrits mesurés

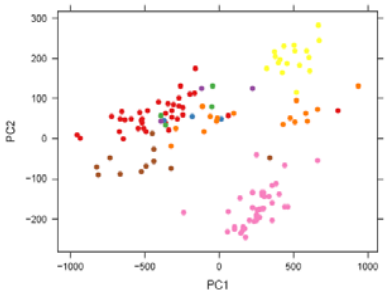


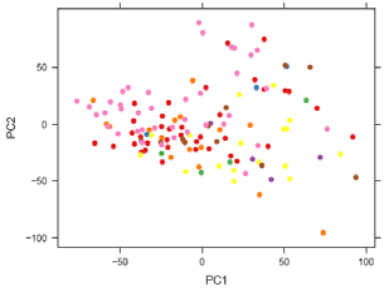
institutCurie

Contrôle Qualité et prétraitement des données

- Contrôle qualité :

Package **arrayQualityMetrics**
(Kauffman et al., Bioinformatics 2009)

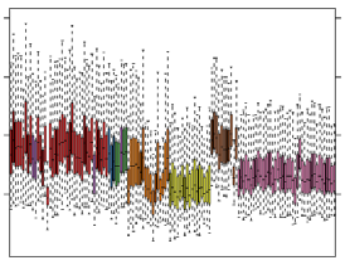




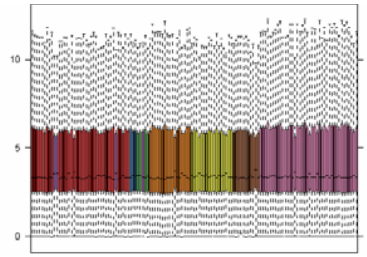
- Normalisation GCRMA (Wu et al, JASA 2004)

library(EMA)
celpath<-"133_fichiers_CEL_MD_Anderson"
MDA<-normAffy(celfile.path=celpath, method="GCRMA")

22215 gènes



Avant normalisation

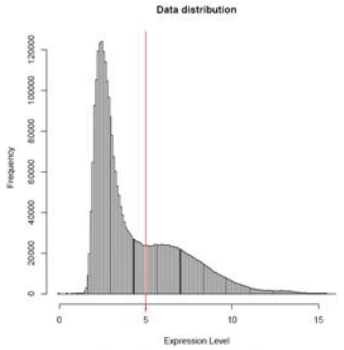


Après normalisation

- Filtrage

MDA.f<-expFilter(MDA, threshold=5)

14125 gènes



Contrôle Qualité

- Potentielles sources de variabilité à chaque étape expérimentale :
  - Extraction et marquage des cibles
  - Hybridation
  - Scanner et analyse d'image
  - Conditions expérimentales (température, technicien…)
  - …
- Ces variabilités expérimentales se confondent avec le signal biologique à étudier

→ Évaluation de la qualité des données afin éventuellement de supprimer/réhybrider des puces « problématiques »

Normalisation

- Objectif : s'affranchir de ces variabilités techniques et expérimentales pour
  - Débruiter le signal biologique
  - Rendre les différentes puces comparables entre elles

- Principe GCRMA (Wu et al, 2004)
  - **Ajustement du bruit de fond**: Correction du bruit optique (scanner) et d'hybridation non spécifique (Capacité d'un probe à s'hybrider spécifiquement dépend de sa séquence)
  - **Normalisation**: Rendre les différentes puces comparables (Normalisation par quantiles)
  - **Résumé**: Combiner les intensités des 11-20 probes pour un probeset donné pour définir une valeur d'expression par probeset (Median polish)

Principe normalisation GCRMA (Wu et al, 2004)

- Normalisation par quantiles
  - Idée : Même distribution des niveaux d'intensité des probes de chaque puce

$$X=\begin{bmatrix} I_{1,1} & \dots & I_{1,n} \\ \vdots & I_{i,j} & \vdots \\ I_{p,1} & \dots & I_{p,n} \end{bmatrix} \rightarrow X_s=\begin{bmatrix} \text{Min}(C_1) & \cdots & \text{Min}(C_n) \\ \vdots & I_{\text{Rang}_i,\, j} & \vdots \\ \text{Max}(C_1) & \dots & \text{Max}(C_n) \end{bmatrix} \rightarrow \begin{bmatrix} \overline{L}_1 \\ \vdots \\ \overline{L}_n \end{bmatrix} \rightarrow X_{norm}=\begin{bmatrix} \overline{L}_{5236} & \dots & \overline{L}_{36971} \\ \vdots & \overline{L}_{158006} & \vdots \\ \overline{L}_{259} & \dots & \overline{L}_{13} \end{bmatrix}$$

p probes (lignes)

n puces (colonnes)

Tri de chaque colonne de X par ordre croissant

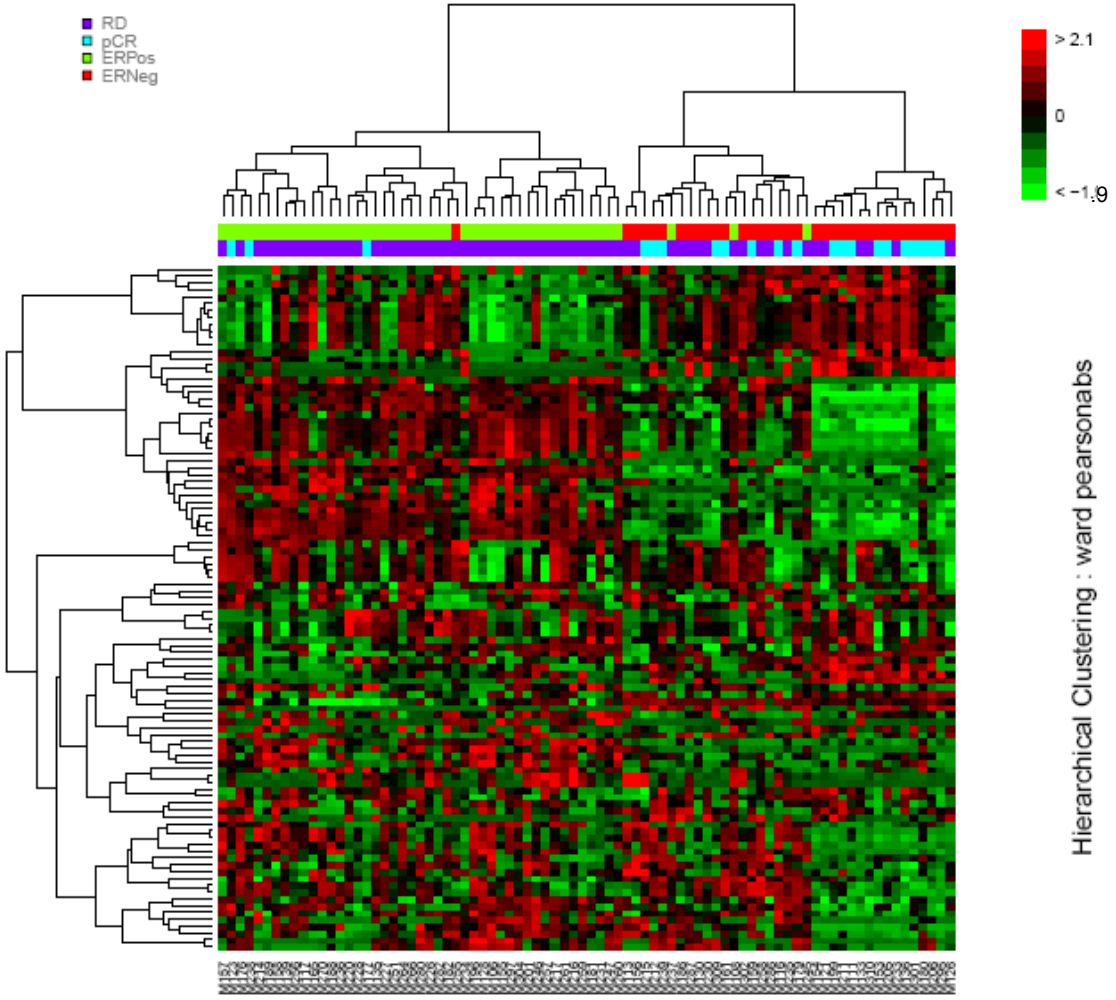
Moyenne de chaque ligne de X_s

Remplace chaque valeur de X par la moyenne correspondante à son rang dans la puce

institutCurie

Heatmap (Eisen, PNAS 1998)

```
mvgenes<-genes.selection(MDA.tr, thres.num=100)
c.sample<-clustering(data=MDA.tr[mvgenes,], metric="pearson", method="ward")
c.gene<-clustering(data=t(MDA.tr[mvgenes,]), metric="pearsonabs", method="ward")
clustering.plot(tree=c.sample, tree.sup=c.gene, data=MDA.tr[mvgenes,],names.sup=FALSE, lab=cl, trim.heatmap=0.99)
```



Analyse Différentielle (SAM)

- Graphe des statistiques de tests observées (d_{obs}) en fonction des attendues (d_{exp})

- Gènes déclarés différentiellement exprimés si $\Delta=|d_{obs}- d_{exp}|>$seuil

- **Proportion médiane de FP (nb_{FP}) sous H_0** : Comptage du nombre médian de gène tel que $\Delta>$seuil sur l'ensemble des permutations

- **Estimation de Π_0** (proportion de gènes vraiment non différentiellement exprimés)

$$\hat{\pi}_0 = \frac{\#\{d_{obs} \in [q_{25}; q_{75}]\}}{0.5 * p}$$

Avec q_{25} et q_{75} les quantiles à 25% et 75% des statistiques attendues d_{exp} sur l'ensemble des permutations

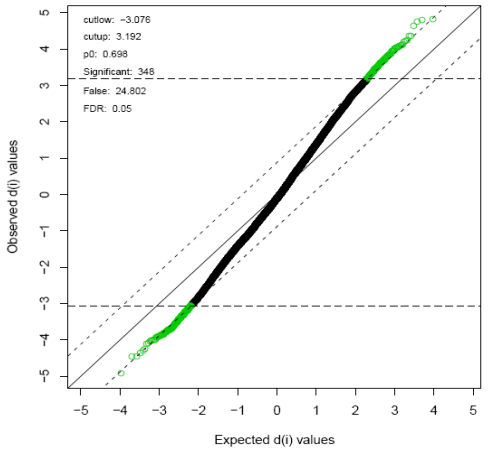
- FDR median : $\hat{\pi}_0 \times nb_{FP}$

True data

0	0	0	1	1	1
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$
⋮	⋮	⋮	⋮	⋮	⋮
$x_{p,1}$	$x_{p,2}$	$x_{p,3}$	$x_{p,4}$	$x_{p,5}$	$x_{p,6}$

Random permutation of sample labels

1	0	1	1	0	0
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$
⋮	⋮	⋮	⋮	⋮	⋮
$x_{p,1}$	$x_{p,2}$	$x_{p,3}$	$x_{p,4}$	$x_{p,5}$	$x_{p,6}$



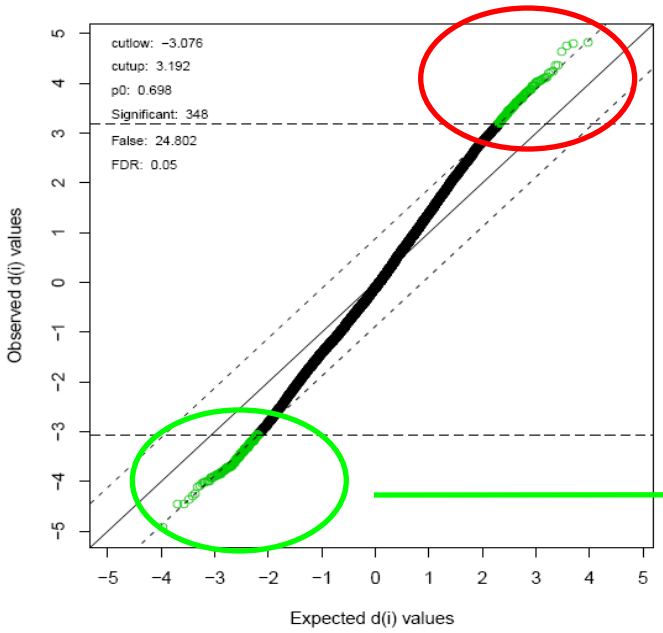
Analyse Différentielle

- SAM Wilcoxon

rs<-runSAM(MDA.tr, labels=tr.num,method ="wilc.stat")

348 gènes différentiellement exprimés (FDR=0.05).

348*0.05=**18 gènes attendus faux positifs**



148 gènes sur-exprimés chez les tumeurs ayant répondu au traitement

200 gènes sous-exprimés chez les tumeurs ayant répondu au traitement

institutCurie

SAM (Significance Analysis of Microarrays)

The data is x_{ij}, $i = 1, 2, \ldots p$ genes, $j = 1, 2, \ldots n$ samples, and response data y_j, $j = 1, 2, \ldots n$ (y_j may be a vector).

Here is the generic SAM procedure:

1. Compute a statistic

$$d_i = \frac{r_i}{s_i + s_0};\ i = 1, 2, \ldots p \tag{17.1}$$

   r_i is a score, s_i is a standard deviation, and s_0 is an exchangeability factor. Details of these quantities are given later in this note.

2. Compute order statistics $d_{(1)} \leq d_{(2)} \cdots \leq d_{(p)}$

3. Take B sets of permutations of the response values y_j. For each permutation b compute statistics d_i^{*b} and corresponding order statistics $d_{(1)}^{*b} \leq d_{(2)}^{*b} \cdots \leq d_{(p)}^{*b}$.

4. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$ for $i = 1, 2, \ldots p$.

5. Plot the $d_{(i)}$ values versus the $\bar{d}_{(i)}$.

6. For a fixed threshold Δ, starting at the origin, and moving up to the right find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_1 are called "significant positive". Similarly, start at origin, move down to the left and find the first $i = i_2$ such that $\bar{d}_{(i)} - d_{(i)} > \Delta$. All genes past i_2 are called "significant negative". For each Δ define the upper cut-point $\text{cut}_{up}(\Delta)$ as the smallest d_i among the significant positive genes, and similarly define the lower cut-point $\text{cut}_{low}(\Delta)$.

7. For a grid of Δ values, compute the total number of significant genes (from the previous step), and the median number of falsely called genes, by computing the median number of values among each of the B sets of $d_{(i)}^{*b}$, $i = 1, 2, \ldots p$, that fall above $\text{cut}_{up}(\Delta)$ or below $\text{cut}_{low}(\Delta)$. Similarly for the 90th percentile of falsely called genes.

8. Estimate π_0, the proportion of true null (unaffected) genes in the data set, as follows:

   (a) Compute $q25, q75 = 25\%$ and 75% points of the permuted d values (if $p = \#$ genes, $B = \#$ permutations, there are pB such d values).

   (b) Compute $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\}/(.5p)$ (the d_i are the values for the original dataset: there are p such values.)

   (c) Let $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$ (i.e., truncate at 1). This estimate of π_0 is analogous to setting $\lambda = 0.5$ in the $\hat{\pi}_0$ proposed in [5]. For *multiclass* data, the scores are all positive, so we use the 0th and 50th percentiles of the permuted values [NOTE: this was corrected in version 2.0].

9. The median and 90th percentile of the number of falsely called genes from step 6, are multiplied by $\hat{\pi}_0$.

10. User then picks a Δ and the significant genes are listed.

11. The False Discovery Rate (FDR) is computed as [median (or 90th percentile) of the number of falsely called genes] divided by [the number of genes called significant].

Analyse de groupes de gènes

- Motivations
  - Meilleure **interprétation biologique** et **reproductibilité** des résultats
  - **Réduction du nombre de tests**

- Objectif
  - Détecter parmi les **groupes de gènes** connus *a priori* ceux dont **l'expression diffère significativement** entre les deux groupes de tumeurs

- Principe du « global test » (Goeman et al, Bioinformatics, 2004)
  - Modèle linéaire généralisé : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$
  - Lorsque m>>n, réécriture du modèle en modèle à effet aléatoire :

  On suppose que les β_j suivent la même loi, d'espérance 0 et de variance τ^2

  $$H_0 : \tau^2 = 0$$ (résolution par test du score)

institutCurie

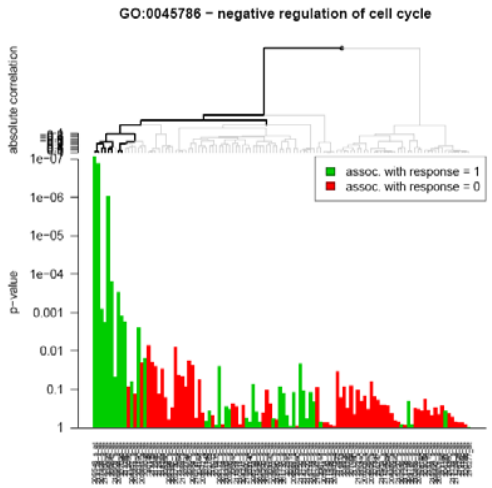
Analyse de groupes de gènes

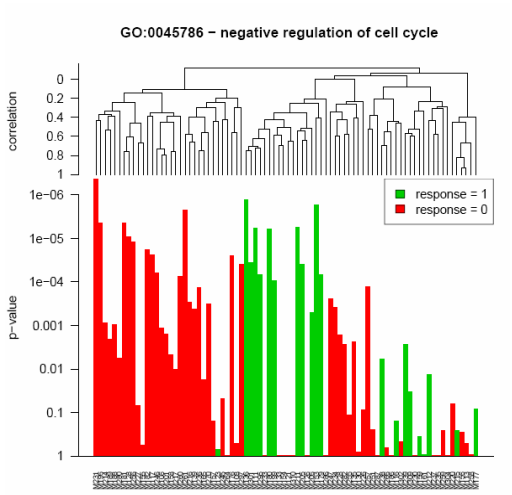
- Package R globaltest (Goeman et al, Bioinformatics, 2004)

4264 termes GO testés, 422 significatifs (FDR<=0.01, Benjamini and Yekutieli)

4264*0.01=**43 groupes de gènes attendus faux positifs**

Goid	p-value	Statistic	Expected	Std.dev	#Cov	BY	alias
GO:0045859	4.94E-08	4.879	1.235	0.336	431	0.001	regulation of protein kinase activity
GO:0043549	5.19E-08	4.942	1.235	0.337	439	0.001	regulation of kinase activity
GO:0051338	6.12E-08	4.879	1.235	0.335	452	0.001	regulation of transferase activity
GO:0007205	1.26E-07	7.456	1.235	0.571	32	0.001	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway
GO:0043405	1.62E-07	5.675	1.235	0.418	196	0.001	regulation of MAP kinase activity
GO:0045664	2.66E-07	7.345	1.235	0.506	132	0.001	regulation of neuron differentiation
GO:0050767	2.75E-07	6.808	1.235	0.471	176	0.001	regulation of neurogenesis
GO:0051960	4.15E-07	6.266	1.235	0.448	198	0.002	regulation of nervous system development
GO:0033674	4.84E-07	5.638	1.235	0.411	262	0.002	positive regulation of kinase activity
GO:0051347	5.17E-07	5.541	1.235	0.406	269	0.002	positive regulation of transferase activity





Calcul des solutions LASSO : algorithme LAR

Least angle regression is like a more "democratic" version of forward stepwise regression. Recall how forward stepwise regression works:

Forward stepwise regression algorithm:

- Start with all coefficients bj equal to zero.
- Find the predictor xj most correlated with y, and add it into the model. Take residuals r= y-yhat.
- Continue, at each stage adding to the model the predictor most correlated with r.
- Until: all predictors are in the model

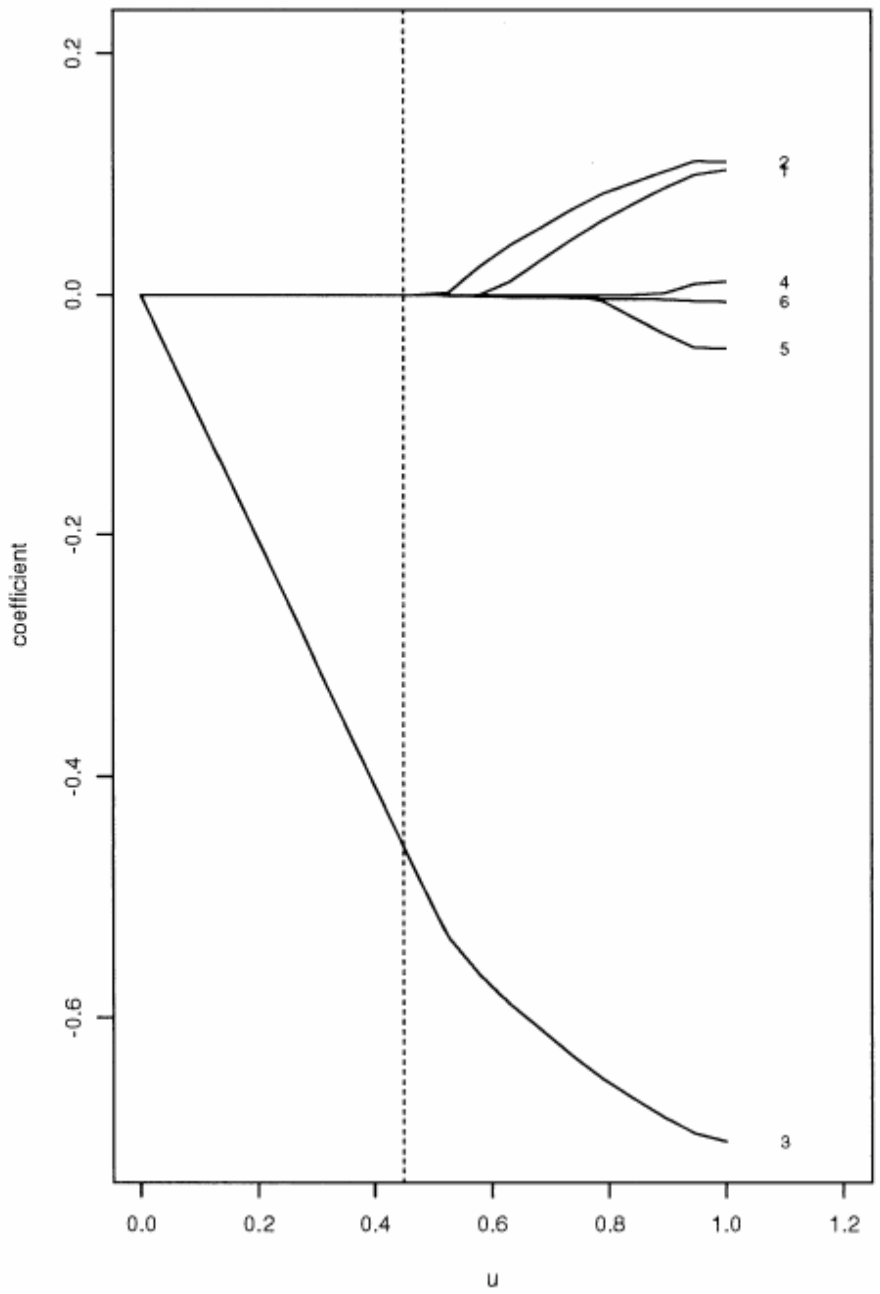
The least angle regression procedure follows the same general scheme, but doesn't add a predictor fully into the model. The coefficient of that predictor is increased only until that predictor is no longer the one most correlated with the residual r. Then some other competing predictor is invited to "join the club".

Least angle regression algorithm:

- Start with all coefficients bj equal to zero.
- Find the predictor xj most correlated with y
- Increase the coefficient bj in the direction of the sign of its correlation with y. Take residuals r=y-yhat along the way. Stop when some other predictor xk has as much correlation with r as xj has.
- Increase (bj, bk) in their joint least squares direction, until some other predictor xm has as much correlation with the residual r.
- Continue until: all predictors are in the model

institutCurie

LASSO



Confrontation aux variables « cliniques »

- Objectif
  - **Évaluer l'apport de la génomique à l'information « clinique »** utilisée habituellement

- Principe
  - Construire les **modèles clinique et clinico-génomique sur le training**
  - Comparer leurs **performances sur le jeu de validation**

Confrontation aux variables « cliniques»

- **Modèle clinique (training)**

	Analyse univariée		Analyse multivariée	
	Odds ratio (IC 95%)	*p*	Odds ratio (IC 95%)	*p*
Taille T3/T4 vs T0/T1/T2	0.66 (0.22-1.94)	$4.52x10^{-1}$		
Envahissement ganglionnaire N1/N2/N3 vs N0	1.41 (0.48-4.16)	$5.33x10^{-1}$		
Grade III vs I/II	9.19 (1.97-42.93)	$4.78x10^{-3}$		
Her2 positif vs negatif	2.11 (0.75-5.94)	$1.58x10^{-1}$		
Statut RP positif vs negatif	0.20 (0.06-0.68)	$9.94x10^{-3}$		
Statut RO positif vs negatif	0.06 (0.02-0.25)	$6.37x10^{-5}$	0.07 (0.02-0.27)	*1.01x10-4*
Age (continue)	0.95 (0.91-0.99)	$4.72x10^{-2}$		

- **Modèle clinico-génomique (training)**
  - Lasso sur RO et 348 gènes différentiels. Pénalisation uniquement sur les gènes.

➡ **RO + 23 gènes sélectionnés**

- **Comparaison clinique et clinico-génomique (validation)**

Clinique

		RO	
		Positif	**Negatif**
realite	**RD**	31	7
	pCR	4	9

Clinico-génomique

		RO + 23 gènes	
		score<=0.5	**score>0.5**
realite	**RD**	**32**	**6**
	pCR	4	9

institutCurie

Autres fonctions du package EMA

- Évaluation de la robustesse du clustering
- Annotation des probesets via BioMart
- Graphe des profils d'expression des probes pour un probeset donné
- Analyse différentielle sur petits échantillons : Rank prod
- Analyse fonctionnelle (loi hypergéométrique, GSA)
- …