

AVNER BAR-HEN

JEAN-JACQUES DAUDIN

STÉPHANE ROBIN

Comparaisons multiples pour les microarrays

Journal de la société française de statistique, tome 146, n° 1-2 (2005),
p. 45-62

http://www.numdam.org/item?id=JSFS_2005__146_1-2_45_0

© Société française de statistique, 2005, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

COMPARAISONS MULTIPLES POUR LES MICROARRAYS

Avner BAR-HEN, Jean-Jacques DAUDIN *, Stéphane ROBIN **

RÉSUMÉ

Les technologies de haut débit (imagerie médicale, détection de sources en astrophysique, transactions commerciales électroniques et biopuces) produisent des masses importantes de données. Les questions posées amènent à une problématique de tests multiples avec plus de mille voire plus d'un million de tests simultanés. Cet article présente une partie des résultats nouveaux obtenus depuis 1995 sur ce sujet, et s'attache en particulier à l'estimation et au contrôle du *False Discovery Rate (FDR)*, au *FDR local* et aux modèles de mélange sur les probabilités critiques. Le problème est posé dans le cadre de l'analyse des données issues du transcriptome mais les méthodes sont générales.

Mots clés : Biopuces, modèle de mélange, taux de faux positifs, tests multiples.

ABSTRACT

Current scientific and management techniques in data and image processing routinely produce hypothesis testing problems with thousands of cases to consider simultaneously. This review about multiple testing focuses on *False Discovery Rate (FDR)*, *local FDR* and mixture models on *p-values* with many examples from genomic data.

Keywords : FDR, local FDR, microarrays, mixture model, multiple testing.

1. Introduction

La question des tests multiples est un problème classique de statistique depuis l'époque de Fisher : les améliorateurs des plantes comparaient plusieurs dizaines de variétés et cherchaient à détecter celles qui étaient meilleures que la variété témoin. Se pose alors le problème du contrôle du risque de première espèce d'un ensemble de tests (par opposition au risque associé à un seul test). Comme on prend simultanément plusieurs décisions il faut contrôler le nombre de rejets à tort de l'hypothèse H_0 .

Les technologies de haut débit ont donné une nouvelle jeunesse à cette vieille question : imagerie médicale, détection de sources en astrophysique, transactions commerciales et biopuces produisent des masses importantes de données. Les questions posées amènent à une problématique de tests multiples avec un saut qualitatif du nombre d'hypothèses testées : on passe de quelques

* Auteur correspondant : Jean-Jacques Daudin daudin@inapg.fr

** UMR INAPG/ENGREF/INRA-MIA 518, INAPG, 16 rue Claude Bernard, 75231 Paris Cedex 05.

dizaines à plus de mille voire plus d'un million de tests simultanés. Comme on le verra dans la suite de cet article, ce changement modifie profondément la problématique initiale, car on peut tirer parti de ce grand nombre pour poser des modèles statistiques sur les probabilités critiques («*p-values*») des tests eux-mêmes.

Dans la première partie nous présentons les éléments et les notations de base. Le point de vue du statisticien consiste à chercher des procédures qui contrôlent le risque de première espèce (au sens large). Il faut au préalable définir précisément les quantités à contrôler. Parmi ces dernières le «*Family Wise Error Rate*» (FWER) et le «*False Discovery Rate*» (FDR) se sont imposés comme des critères de référence. Pour des raisons de place il était impossible de traiter tous les critères de façon détaillée. Nous avons choisi dans cet article de privilégier le FDR. Les lecteurs intéressés par les nouveaux développements sur le FWER pourront consulter avec profit Dudoit *et al.* [9] et l'article de ces mêmes auteurs dans le présent volume. La troisième partie traite de l'estimation et du contrôle du FDR. La quatrième partie est consacrée au FDR local et la dernière partie traite de l'estimation de la probabilité que l'hypothèse nulle soit vraie pour un item ou test donné, par l'intermédiaire d'un modèle de mélange sur les probabilités critiques.

On pose ici le problème dans le cadre de l'analyse des données du transcriptome : on cherche à savoir quels sont les gènes qui sont différentiellement exprimés entre deux conditions expérimentales. On considère m gènes, $i = 1, m$. L'hypothèse H_{0i} correspond à un gène i non différentiellement exprimé et H_{1i} correspond à un gène i différentiellement exprimé. On note m_0 le nombre de gènes dans H_0 et m_1 le nombre de gènes dans H_1 . On note p_i la probabilité critique associée à une statistique qui permet de tester H_{0i} . On admet que le test est exact, c'est à dire que sous H_{0i} , p_i est distribuée selon une loi uniforme. Dans la suite, sauf indication contraire, on admettra que les probabilités critiques sont de plus indépendantes sous H_0 . Des propositions permettant de s'affranchir de cette hypothèse sont présentées dans ce volume par Pollard, Birkner van der Laan et Dudoit.

Cette méthodologie, présentée à partir du domaine particulier du transcriptome, s'étend à toute problématique de tests multiples.

2. Procédures de tests multiples : éléments de base

2.1. Règle de décision

On dispose des probabilités critiques pour chaque gène p_i , $i = 1, m$. Dans la suite de ce texte, on suppose que les p_i sont ordonnées : $p_1 \leq p_2 \leq \dots p_m$. Une procédure de test multiple est une règle de décision qui consiste à définir un seuil t et à rejeter l'hypothèse H_0 pour tous les gènes qui sont tels que $p_i \leq t$.

Il faut noter que t peut être fonction de $(p_1, \dots p_m)$ ou être définie sans prendre en compte ces probabilités critiques observées.

La question posée est donc de trouver une méthode permettant de choisir t . Pour cela, il faut d'abord définir un critère explicitant ce qu'est une «*bonne règle de décision*».

2.2. Faux positifs et faux négatifs

Pour fixer la terminologie nous appellons

- *positif* un gène déclaré différentiellement exprimé (c'est-à-dire pour lequel on rejette l'hypothèse H_0),
- *négatif* un gène déclaré non-différentiellement exprimé (c'est-à-dire pour lequel on accepte l'hypothèse H_0),
- *gène dans H_1* un gène différentiellement exprimé (c'est-à-dire pour lequel l'hypothèse H_1 est vraie)
- *gène dans H_0* un gène non-différentiellement exprimé (c'est-à-dire pour lequel l'hypothèse H_0 est vraie).

On définit les processus de comptages suivants :

1. $V(t) = \#\{i : (p_i \leq t) \cap H_0\}$, le nombre de faux positifs au seuil t
2. $S(t) = \#\{i : (p_i \leq t) \cap H_1\}$ le nombre de vrais positifs au seuil t
3. $R(t) = S(t) + V(t)$ le nombre de gènes positifs
4. $U(t) = \#\{i : (p_i > t) \cap H_0\}$, le nombre de vrais négatifs au seuil t
5. $T(t) = \#\{i : (p_i > t) \cap H_1\}$ le nombre de faux négatifs au seuil t
6. $W(t) = U(t) + T(t)$ le nombre de gènes négatifs

Le tableau suivant résume la situation :

	Acceptation de H_0	Rejet de H_0	Total
H_0	$U(t)$	$V(t)$	m_0
H_1	$T(t)$	$S(t)$	m_1
Total	$W(t)$	$R(t)$	m

On a donc six processus empiriques à valeurs dans \mathbb{N} et indexés par t . Que sait-on sur ces quantités ?

1. $U(t) + V(t) = m_0$, $T(t) + S(t) = m_1$ et $R(t) + W(t) = m$.
2. Seuls $R(t)$ et $W(t)$ sont observés.
3. Ces processus sont liés. Si on connaît m_0 et m_1 , $V(t)$ suffit pour connaître les cinq autres. En général on ne connaît pas m_0 ni m_1 , et il suffit de connaître $V(t)$ et $U(t)$ pour connaître les quatre autres.
4. On connaît la loi marginale des processus $V(t) : V(t) \sim Bin(m_0, t)$ et $U(t) : U(t) \sim Bin(m_0, 1 - t)$.
5. On connaît aussi certaines lois conditionnelles associées. Par exemple, $V(t_1)/V(t_2) \sim Bin(V(t_2), \frac{t_1}{t_2})$ pour $t_2 > t_1$.

Pour illustrer le comportement de certains de ces processus, considérons un cas simple avec $m_0 = 20$, $m_1 = 10$. On suppose que les probabilités critiques des gènes dans H_1 sont distribuées selon une loi uniforme entre 0 et 0.1. La

figure 1 donne un exemple de trajectoires pour $V(t)$, $S(t)$ et $R(t)$. De ces 3 trajectoires, seule celle de $R(t)$ est observable en pratique.

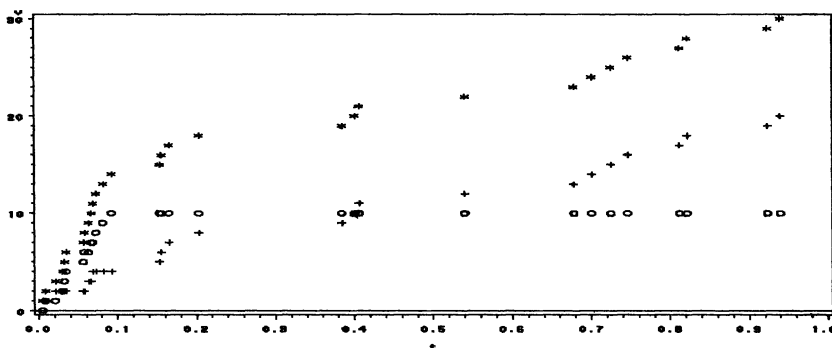


FIG 1. — $V(t)(+)$, $S(t)(o)$ et $R(t)(*)$ en fonction de t en abscisse, $m_1 = 10$, $m_0 = 20$.

2.3. Critères pour le choix de la règle de décision

Le point de vue du statisticien met l'accent sur le contrôle des faux positifs : il s'agit de trouver des procédures qui contrôlent le risque de première espèce (ou le nombre ou la proportion de faux positifs). Ensuite, parmi ces procédures on cherche celle qui maximise $S(t)$ (c'est-à-dire en pratique on cherche à prendre t le plus grand possible).

Il y a plusieurs façons de contrôler le risque de première espèce. On cherche t tel que :

1. Contrôle large

$$E[V(t)] < m\alpha$$

ou $E[V(t)] < m_0\alpha$ si on connaît m_0 . On retrouve ici le « test non corrigé » ou PFER « per-family error rate », classique dans le domaine des comparaisons multiples. $\frac{E[V(t)]}{m}$ est appelé PCER, « per-comparison error rate ». Une variante possible est :

$$P(V(t) > v) < \alpha.$$

Cette quantité est appelée gFWER(v) (« generalized family-wise error rate ») par Dudoit et al. [9]

2. Contrôle intermédiaire

On contrôle le FDR (« False Discovery rate ») défini par

$$FDR(t) = E \left[\frac{V(t)}{\max(1, R(t))} \right].$$

Le FDR est une espérance. Il ne contrôle donc pas le taux réel de faux positifs mais seulement sa moyenne dans un nombre infini d'expériences.

Si on veut assurer le contrôle du taux réel il vaut mieux utiliser le critère suivant :

$$P \left[\frac{V(t)}{\max(1, R(t))} > \alpha \right] < \beta.$$

3. Contrôle strict

$$P(V(t) > 0) < \alpha$$

appelé FWER (« Family-wise error rate »). La solution la plus simple consiste à utiliser l'inégalité de Bonferroni qui conduit à la règle de décision $p_i \leq t$ où $t = \frac{\alpha}{m_0}$ si on connaît m_0 et $t = \frac{\alpha}{m}$ sinon. Ce contrôle est valide même quand les tests sont dépendants. Il existe plusieurs variantes, légèrement plus puissantes en prenant en compte certaines hypothèses de dépendance entre les probabilités critiques. Certaines procédures utilisent des permutations ou une méthode bootstrap pour générer une loi conjointe des statistiques de test sous H_0 . Dudoit *et al.* [8] et [9] détaillent ces différentes méthodes.

La nouveauté du domaine est le *FDR*. Il est plus difficile à estimer et à contrôler que les deux autres parce que c'est l'espérance du rapport de deux variables aléatoires, ce qui donne lieu à des développements mathématiques nouveaux.

3. Résultats sur le FDR

Benjamini et Hochberg [3] ont défini le FDR et montré comment on peut le contrôler. Ils ont en particulier démontré le résultat suivant :

Soit $r = \max\{i : \frac{mp_i}{i} \leq \alpha\}$ et $t = p_r$. Si les statistiques de test sont indépendantes, on a

$$FDR(t) \leq \frac{m_0}{m} \alpha \leq \alpha.$$

Ce travail a été à l'origine de beaucoup d'autres. En particulier, Storey *et al.* [19] ont redémontré par d'autres méthodes les résultats de Benjamini et Hochberg et nous suivons ici leur présentation. Il y a deux classes de résultats : ceux qui concernent l'estimation du FDR et ceux qui concernent son contrôle.

3.1. Estimation du FDR

On suppose que les probabilités critiques sont indépendantes et on ne fait pas d'hypothèse sur la distribution des probabilités critiques sous H_1 . Storey *et al.* [19] ont établi deux théorèmes, l'un pour m_0 connu et l'autre pour m_0 inconnu :

THÉORÈME 3.1. — Soit $F\widehat{DR}(t) = \frac{m_0 t}{R(t)}$ si $R(t) > 0$ et $F\widehat{DR}(t) = 0$ si $R(t) = 0$. On a

$$E(F\widehat{DR}(t)) \geq FDR(t).$$

Idée de démonstration. — La fonction

$$f(x) = \frac{a-x}{x+b}$$

est convexe ($f'' = \frac{2(a+b)}{(x+b)^3} > 0$) pour $x > 0$ et $b > 0$. Donc $E(f(X)) \geq f(E(X))$ grâce à l'inégalité de Jensen.

Application :

- $a = m_0 t$
- $b = S(t)$
- $X = V(t)$

$V(t) \sim \text{Bin}(m_0, t)$ donc $E(V(t)) = m_0 t$ donc $f(E(V(t))) = 0$. La seule difficulté est que $S(t)$ est aléatoire. Le raisonnement précédent est vrai pour tout s conditionnellement à $S(t) = s$ ($V(t)$ et $S(t)$ sont indépendantes à t fixé). Il faut faire attention au cas $R(t) = 0$ qu'il faut traiter à part. \square

Dans le cas où m_0 est inconnu, on l'estime par $\widehat{m}_0(\lambda) = \frac{W(\lambda)}{1-\lambda}$ où $\lambda \in [0, 1[$.

THÉORÈME 3.2. — Soit $\widehat{FDR}(t, \lambda) = \frac{\widehat{m}_0(\lambda)t}{R(t)}$ si $R(t) > 0$ et $\widehat{FDR}(t, \lambda) = 0$ si $R(t) = 0$. On a

$$E(\widehat{FDR}(t, \lambda)) \geq FDR(t).$$

Idée de démonstration. — C'est la même technique sauf qu'il faut ajouter l'élément suivant : $W(\lambda) \geq U(\lambda)$ donc

$$E(W(\lambda)) \geq E(U(\lambda)) = m_0(1-\lambda)$$

donc

$$m_0 \leq \frac{E(W(\lambda))}{1-\lambda}.$$

Finalement $\widehat{m}_0(\lambda) = \frac{W(\lambda)}{1-\lambda}$ surestime m_0 . \square

3.2. Contrôle du FDR

Si on cherche maintenant à contrôler le FDR et non pas seulement à l'estimer, on utilise l'estimateur précédent qui majore en espérance le FDR. Il est alors naturel de choisir le seuil

$$t_\alpha = \sup\{t : \widehat{FDR}(t) \leq \alpha\}.$$

où $\widehat{FDR}(t)$ est donné dans le théorème 3.1. Cependant la propriété attendue

$$FDR(t_\alpha) \leq \alpha.$$

n'est pas évidente et nécessite une démonstration.

Storey et al. [19] ont montré les résultats suivants :

LEMME 3.3. — Si les p_i sont indépendantes pour $i \in H_0$, $\frac{V(t)}{t}$ est une martingale à temps renversé (« reverse martingale » ou « time running backward ») par rapport à la filtration $F_t = \sigma\{p_i \leq s\}, t < s < 1$ quand t varie de 1 vers 0 : pour $s \leq t$,

$$E \left[\frac{V(s)}{s} / F_t \right] = \frac{V(t)}{t}$$

Idée de démonstration. — Conditionnellement à $V(t)$, $V(s) \sim \text{Bin}(V(t), \frac{s}{t})$ □

LEMME 3.4. — t_α est un temps d'arrêt par rapport à F_t avec t allant de 1 vers 0, (martingale à temps renversé).

THÉORÈME 3.5. — Si on connaît m_0 ,

$$FDR(t_\alpha) = \alpha.$$

Idée de démonstration. — On a $R(t_\alpha) = \frac{t_\alpha m_0}{\alpha}$, et donc

$$\frac{V(t_\alpha)}{R(t_\alpha)} = \frac{\alpha}{m_0} \frac{V(t_\alpha)}{t_\alpha}$$

On utilise le « optional stopping theorem », qui dit que

$$E \left[\frac{V(t_\alpha)}{t_\alpha} \right] = E(V(1)) = m_0$$

□

THÉORÈME 3.6. — Si on ne connaît pas m_0 ,

$$FDR(t_\alpha, \lambda) = (1 - \lambda^{m_0})\alpha \leq \alpha$$

où $t_\alpha = \sup\{t : \widehat{FDR}(t, \lambda) \leq \alpha\}$ et $\widehat{FDR}(t, \lambda)$ est donné dans le théorème th2. Storey et al. [19] ont également obtenu des résultats asymptotiques analogues aux résultats précédents mais un peu plus généraux dans la mesure où une faible dépendance entre les probabilités critiques est acceptable.

3.3. Estimation de π_0

L'estimation de $\pi_0 = \frac{m_0}{m}$ est un point clé pour l'estimation et le contrôle du FDR. Plusieurs méthodes ont été proposées. Toutes sauf une sont basées sur l'argument suivant : les probabilités critiques sous H_0 sont distribuées selon une loi uniforme et les probabilités critiques sous H_1 se situent plutôt à proximité de 0. Par suite π_0 est égal à l'ordonnée de la densité de probabilité de toutes les probabilités critiques pour l'abscisse 1. Il ne reste plus qu'à estimer cette dernière :

1. Storey & Tibshirani [20] proposent de modéliser $\hat{\pi}_0(\lambda)$ en fonction de λ par une fonction spline cubique et d'en déduire $\lim_{\lambda \rightarrow 1} \pi_0(\lambda)$.
2. Pounds & Cheng [16] utilisent un modèle de mélange de lois beta (cf section 5) (méthode Beta Uniform Mixture «BUM»).
3. Pounds & Morris [17] lissent la densité des probabilités critiques au voisinage de 1 à l'aide d'une méthode de régression locale LOESS (méthode SPLOSH).
4. Dalmaso *et al* [5] proposent une méthode différente, nommée LBE («Location Based Model») (voir également l'article de ces auteurs dans ce numéro). Il est facile de vérifier que $\frac{2}{m} \sum p_i$ est un estimateur biaisé de π_0 , ayant un biais positif. Ils proposent une famille d'estimateurs du même type mais ayant un biais plus faible :

$$\pi_0(n) = \frac{\sum_{i=1}^m [-\ln(1 - p_i)]^n}{mn!}$$

Dalmaso *et al.* ont comparé BUM, SPLOSH et LBE à l'aide de simulations. D'après leurs résultats, BUM et SPLOSH fournissent une bonne estimation de π_0 au sens de l'erreur quadratique moyenne. Cependant ils peuvent sous-estimer π_0 et ne sont donc pas utilisables pour contrôler le FDR, alors que l'estimateur de Storey et Tibshirani et LBE surestiment π_0 et peuvent donc être utilisés pour contrôler le FDR.

4. FDR local

Le contrôle du FDR est utile, mais on connaît les inconvénients liés à la fixation d'un seuil arbitraire α pour le FDR. Ce sont les mêmes que pour un test unique. Pourquoi 5 % et pas 3 ou 7 % ? Que penser d'un gène très proche du seuil de signification qu'il soit au dessus ou au dessous de la barre fatidique ? Dans le cadre d'un test unique, la notion de probabilité critique répond à ces questions en attribuant à chaque gène son propre degré de signification statistique. D'où la question suivante : peut-on définir un équivalent de la probabilité critique dans le cadre des tests multiples ?

Storey et Tibshirani [20] ont tenté de répondre à cette question en définissant la « Q-value » : $q_i = FDR(p_i)$, le FDR estimé associé au seuil t le plus faible qui rejette le gène i . Cette idée n'est pas complètement satisfaisante car elle associe dans un même ensemble le gène i et tous les gènes qui ont une probabilité critique plus faible que lui. D'ailleurs Storey et Tibshirani [20] disent explicitement « *We stress that a "Q-value" is not the probability that the feature is a false positive* ». D'où l'idée de « FDR local » défini comme la probabilité qu'un gène de probabilité critique p_i soit dans H_0 . Il existe une version bayésienne du FDR local, définie par Efron *et al.* [10]. Nous suivons ici la démarche purement fréquentiste proposée dans Aubert *et al.* [2].

4.1. Définition et propriétés du FDR local

Soit $0 \leq t_1 < t_2 \leq 1$, et

$$FDR(t_1, t_2) = E \left(\frac{V(t_1, t_2)}{\max(1, R(t_1, t_2))} \right) = E \left(\frac{\#\{i : (t_1 < p_i \leq t_2) \cap H_0\}}{\#\{i : (t_1 < p_i \leq t_2)\}} \right).$$

Le FDR local au point t_2 est $FDRL(t_2) = FDR(t_1, t_2)$ quand $t_2 - t_1$ est petit. Pour obtenir un estimateur de $FDRL(t_2)$, on procède en 2 étapes. Dans une première étape on définit une estimation brute du FDR local, $FDRL_B$. Dans une deuxième étape l'estimation du FDR local est obtenue par un lissage des valeurs brutes. Pour calculer les valeurs brutes, on considère l'estimation la plus locale possible, tout en ayant $R(t_1, t_2) > 0$. Si on choisit $t_2 = p_i$ et $t_1 = p_{i-1}$, on obtient les valeurs brutes suivantes :

$$FDRL_B(i) = \frac{\widehat{m}_0(p_i - p_{i-1})}{1}.$$

On obtient des résultats très proches avec

$$FDRL_B(i) = \frac{\widehat{m}_0(p_{i+1} - p_{i-1})}{2}.$$

Comme ces valeurs sont extrêmement variables, on utilise une méthode de lissage des valeurs brutes par les méthodes loess, spline ou moyennes mobiles pour obtenir des estimations plus stables. Les figures 2, 3 et 4 montrent les valeurs obtenues avant et après lissage par la méthode loess avec un « span » de 0.2 et celles obtenues par une méthode de moyenne mobile adaptative définie dans [2].

En pratique la méthode loess n'est pas toujours adaptée au problème posé (voir la figure 4(c)). La fenêtre de lissage doit elle aussi être adaptative : il ne faut pas de lissage pour les très petites valeurs de $FDRL$ et il faut un fort lissage pour les valeurs de $FDRL$ qui sont proches de 1 ou le dépassent. En effet le $FDRL$ associé à un gène dans H_1 lui est propre alors que le $FDRL$ associé à un gène dans H_0 est théoriquement le même pour tous les gènes dans H_0 . De plus la variance de $FDRL_B$ pour gène très fortement différentiellement

exprimé est très faible alors que la variance de $FDRL_B$ pour un gène dans H_0 est très grande. Il est d'ailleurs facile de calculer cette dernière.

THÉORÈME 4.1. — Si $m_0 = m$, on a

$$E(FDRL_B(i)) = FDRL(i) = 1$$

$$V(FDRL_B(i)) = \frac{m^3}{(m+1)^2(m+2)} \simeq 1.$$

Idée de démonstration. — Comme les p_i sont distribués selon une loi uniforme $U([0, 1])$, les différences successives entre les statistiques d'ordre, h_i sont distribuées selon une loi beta($m, 1$). \square

Comme la loi beta($m, 1$) est très dissymétrique les écarts à la moyenne positifs sont plus grands que les écarts négatifs. On peut donc obtenir des valeurs brutes de $\widehat{FDRL}(i)$ comprises entre 0 et 6.

En pratique on n'a pas $m = m_0$ et on ne connaît pas m_0 . Donc le théorème précédent ne dit rien sur la loi des estimateurs de $FDRL$ en pratique. Par contre, il donne une loi de référence si on connaît m_0 .

4.2. Exemples de calcul du FDR local

4.2.1 Données de Golub

Dudoit *et al.* [8] ont analysé les données de Golub *et al.* qui cherchent à identifier les gènes qui sont différentiellement exprimés dans deux types de leucémies (ALL, AML). Les expressions ont été mesurées avec des puces Affimetrix contenant 6 817 gènes humains avec 27 patients « ALL » et 11 patients « AML ». Les données sont disponibles dans la librairie « multtest » du logiciel R. Nous avons utilisé les données après prétraitement, soit un fichier de 3 051 gènes et 38 patients. Nous avons utilisé les probabilités critiques calculées par permutations des étiquettes AML/ALL sur la statistique de Welch pour chaque gène (Dudoit *et al.* [8]). m_0 est estimé par bootstrap comme il est indiqué dans Storey et Tibshirani [20] et implémenté dans la librairie GeneTS de R.

La figure 2(a) représente les valeurs de $FDRL_B(i)$ et la figure 2(b) représente les valeurs lissées par la méthode loess (avec un « span » de 0.2) et par une moyenne mobile adaptative.

On peut voir qu'il y a une accélération rapide du FDR local autour de $t = 0.15$. Cela donne une indication pour choisir la règle de décision.

La Figure 2(c) présente un zoom de la Figure 2(b) pour les 600 premières probabilités critiques.

Les probabilités critiques ont été obtenues avec des permutations. Donc les valeurs des probabilités critiques sont discrètes avec des ex-aequos. Ainsi certaines différences entre probabilités critiques successives peuvent être nulles.

COMPARAISONS MULTIPLES POUR LES MICROARRAYS

On n'a donc pas une distribution uniforme continue des probabilités critiques. Ceci explique les discontinuités obtenues avec le lissage par moyennes mobiles (Figure 2(c)).

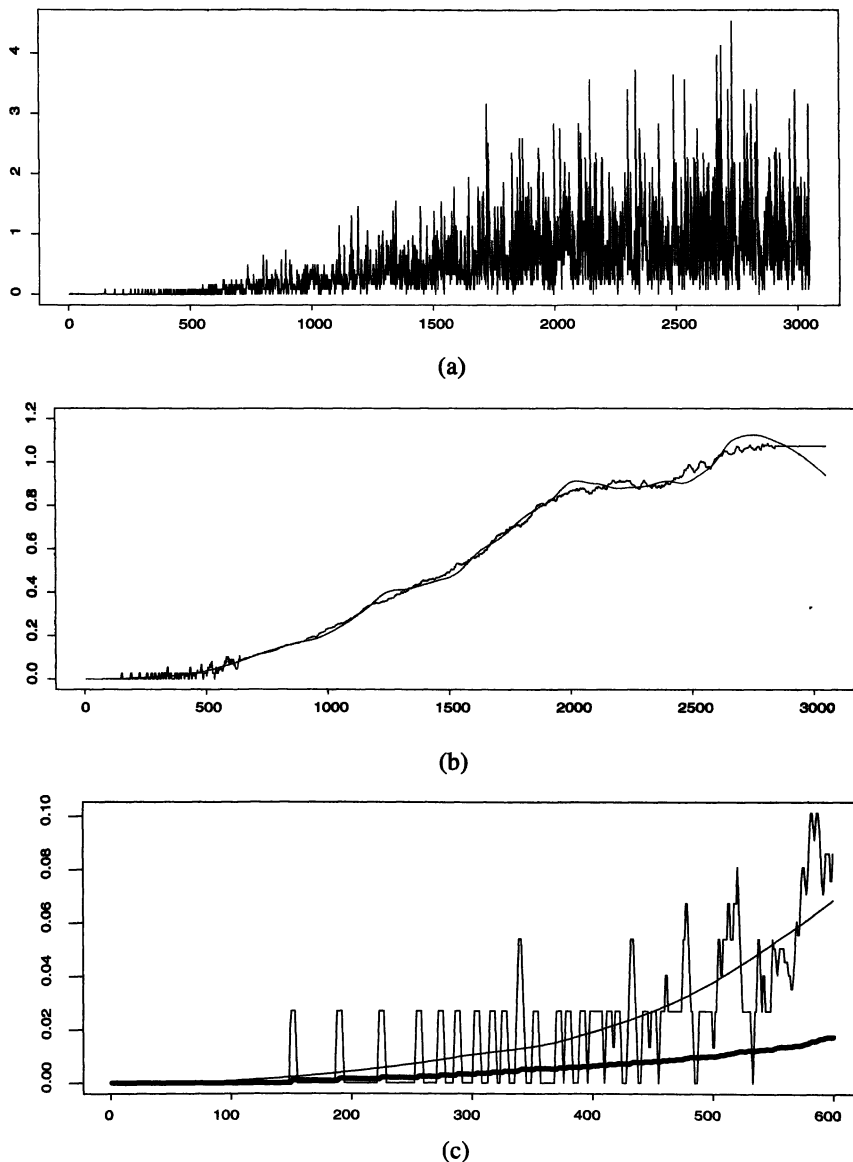


FIG 2. — Estimation du FDR local pour les données de Golub. Abscisse : numéro du gène (ordonné selon la probabilité critique), ordonnée : estimation du FDR local. Fig a : valeurs brutes, Fig b : valeurs lissées, Fig c : zoom sur les 600 premiers gènes de la Fig b, moyenne mobile (avec discontinuités), loess (courbe fine), « Q-value » (courbe épaisse).

La Figure 3(a) présente $FDRL_B(i)$ et la Figure 3(b) présente le FDR local lissé.

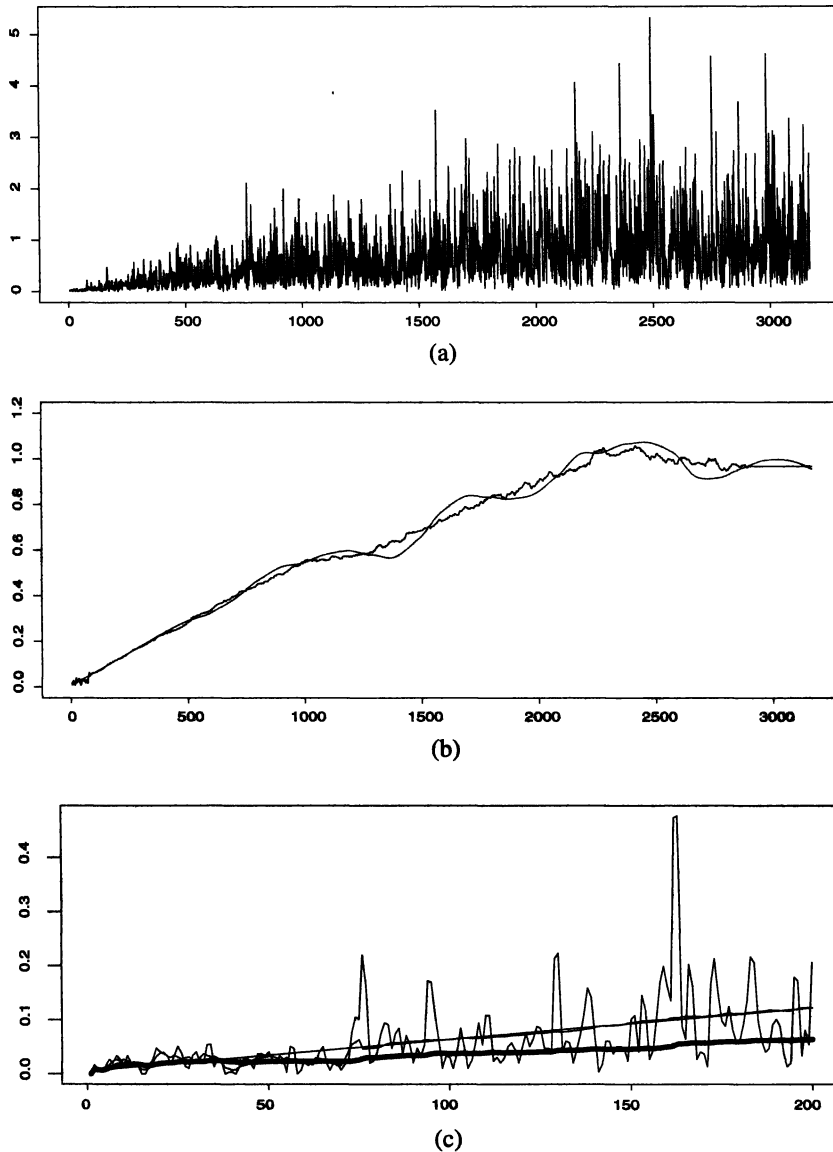


FIG 3. — Estimation du FDR local pour les données d'Hedenfalk. Abscisse : index des gènes ordonnés selon la probabilité critique, ordonnée : FDR local estimé. Fig a : valeurs brutes, Fig b : FDR local estimé avec lissage : moyenne mobile (avec sauts), loess (courbe lissée), Fig c : zoom sur les 200 premiers gènes de la Fig b, valeurs brutes, moyenne mobile et loess (courbes lissées), « Q-value » (courbe lissée inférieure en gras).

4.2.2 Cancer du poumon

Storey et Tibshirani [20] ont analysé en détail les données issues d'une étude sur le cancer du poumon [14] avec 15 microarrays et 3170 gènes. À partir des mêmes probabilités critiques nous avons estimé le FDR local. Nous commentons ici les 3 gènes relevés par les auteurs (cf tableau 1).

TABLEAU 1. – Probabilité critique, « Q-value » et estimation du FDR local pour 3 gènes de l'étude de Hedenfalk

gène	probabilité critique	« Q-value »	FDR local
MSH2	0.00005	0.0013	0.010
PDCD5	0.00048	0.022	0.033
CTGF	0.0036	0.049	0.098

On peut remarquer que l'estimation du FDR local est supérieure à la « Q-value » et donne une meilleure idée de la probabilité que le gène considéré soit un faux positif. Dans certains cas la différence est suffisamment importante pour remettre en cause la conclusion scientifique.

On peut utiliser la courbe pour choisir la valeur de m_0 : à partir d'une certaine valeur de t , la courbe a rejoint son asymptote à 1.

4.2.3 ApoAI

C'est un jeu de données très utilisé par l'équipe de T. Speed dans leurs présentations. On compare l'expression de 8 souris témoins à celle de 8 souris pour lesquelles le gène ApoAI a été annihilé. L'analyse statistique est décrite dans Dudoit *et al.* [7]. On s'attend à trouver différentiellement exprimés les deux clones de ApoAI et six autres clones de gènes co-régulés avec ApoAI sur les 6226 gènes testés. Ces 8 clones sont effectivement les 8 ayant les plus faibles probabilités critiques. Les résultats sont dans les Figures 4(a)(b)(c). Le lissage par la méthode loess n'est pas satisfaisant pour les 50 premiers clones.

On peut calculer le FDR associé au gène ApoAI : c'est la moyenne des FDR locaux de ses deux clones : $\frac{0 + 0.00048}{2} = 0.00024$.

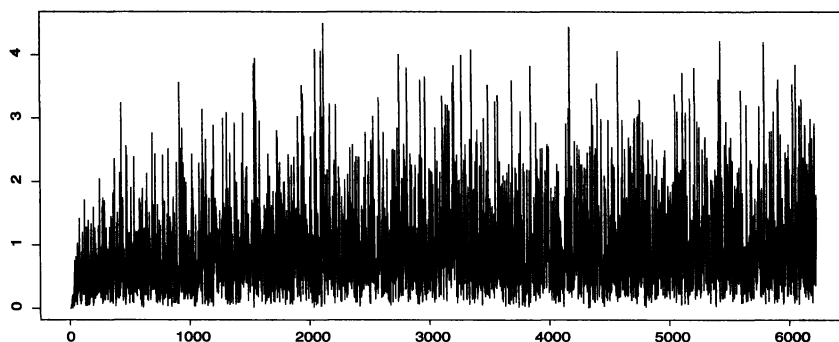
4.3. Conclusion sur le FDR local

La « Q-value » donne une estimation de la proportion de faux positifs dans la liste de gènes dont le gène considéré est le dernier.

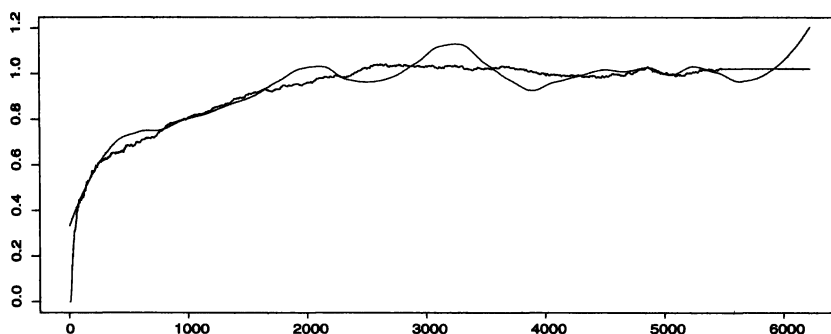
Le FDR local donne une estimation de la proportion de faux positifs dans le voisinage du gène considéré.

Cette dernière quantité peut être utilisée comme une approximation de la probabilité que le gène considéré soit un faux positif compte tenu de la multiplicité des gènes étudiés.

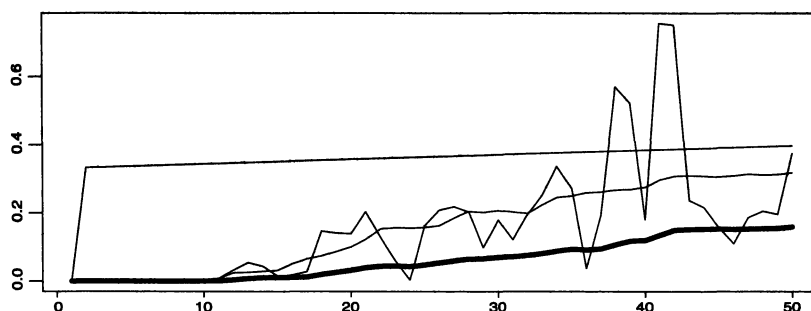
COMPARAISONS MULTIPLES POUR LES MICROARRAYS



(a)



(b)



(c)

FIG 4. — Estimation du FDR local pour les données Apo-AI. Abscisse : index des clones ordonnés selon la probabilité critique, ordonnée : FDR local estimé. Fig a : valeur brute, Fig b : valeurs lissées : moyenne mobile (courbe avec micro irrégularités), loess (courbe lissée), Fig c : zoom sur les 50 premiers gènes de la Figure b, valeurs brutes (courbe avec sauts), moyenne mobile (courbe lissée) loess (droites par morceaux), « Q-value » (courbe lissée en gras).

Un avantage du FDR local est qu'on peut estimer le FDR associé à un groupe de gènes ou de clones : clones d'un même gène, groupe de gènes co-régulés, groupe de gènes appartenant à une même région d'un chromosome.

5. Modèles de mélange

Le problème des tests multiples est un problème de classification : prédire une variable 0/1 connaissant une variable continue (la probabilité critique). Comme on ne dispose pas d'échantillon d'apprentissage, il s'agit de classification non supervisée avec connaissance du nombre de groupes (2). Il est naturel de poser un modèle de mélange sur les probabilités critiques ou sur une transformation continue et monotone $\phi(\cdot)$ des probabilités critiques :

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$$

où

- f est la densité des probabilités critiques (transformées par ϕ),
- $\pi_0 = \frac{m_0}{m}$,
- f_0 est une densité connue, qui est la densité uniforme sur $[0, 1]$ si $\phi = Id$,
- f_1 est la densité des probabilités critiques (transformées par ϕ) sous H_1 .

On a un modèle analogue sur les fonctions de répartition :

$$F(p) = \pi_0 F_0(p) + (1 - \pi_0) F_1(p).$$

À partir de ce modèle (et après avoir estimé f_1 et π_0), on obtient la probabilité *a posteriori* qu'un gène soit dans H_0 :

$$\tau(x) = p(g \in H_0 \mid p_i = x) = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + (1 - \pi_0) f_1(x)}.$$

On a $\tau(x) \equiv FDR_L(x)$ c'est-à-dire que $\tau(x)$ est par définition égal au FDR local. De plus $\tau(x)$ permet d'estimer

$$FDR(t) = \sum_{p_i \leq t} \tau(p_i)$$

mais aussi

$$P \left[\frac{V(t)}{\max(1, R(t))} > \alpha \right] < \beta.$$

Cette modélisation est donc à la fois naturelle et très utile. La plupart des travaux actuels sur le sujet empruntent donc cette voie.

Efron *et al.* [10] et [11] ont été parmi les premiers à utiliser ce type de modèle dans un contexte bayésien, avec une modélisation non paramétrique de f_1 .

Le contexte (variables aléatoires à valeurs dans $[0, 1]$ et importance de la loi uniforme) ont incité beaucoup de chercheurs à utiliser un mélange de lois beta.

Allison *et al.* [1] ont été les premiers à explorer cette voie, suivis par Pounds et Cheng [16] qui imposent de plus que l'une des lois beta du mélange soit une loi uniforme. Pounds et Morris [17], ont aussi exploré la voie du mélange de lois beta.

Genovese et Wasserman [12], [13] se placent dans un contexte plus large : ils supposent que la loi F_1 existe et définissent 3 classes de loi F_1 :

1. $F_S = \{F_1 : F_1 \geq U \text{ (loi uniforme) et } F_1 \text{ absolument continue}\}$
2. $F_C = \{F_1 \in F_S \text{ et } F_1 \text{ concave}\}$
3. F_1 appartient à une famille paramétrique

Ils donnent des résultats pour F_1 connue puis pour F_1 estimée. Ils définissent deux risques :

1. Risque Marginal :

$$E \left[\frac{V(t)}{R(t)} \right] + E \left[\frac{T(t)}{m - R(t)} \right]$$

2. Risque Conditionnel :

$$\lambda E \left[\frac{V(t)}{R(t)} / p \right] + E \left[\frac{T(t)}{m - R(t)} / p \right]$$

où p est le vecteur des probabilités critiques et λ est un paramètre (à fixer par l'utilisateur) pondérant l'importance relative des deux types d'erreurs.

Dans le cas où $F_1 \in F_C$, ils obtiennent la loi de $V/(R = r, p)$ sous forme de fonction génératrice faisant intervenir $F_1(p)$. Puis ils calculent le risque conditionnel dans le cas particulier où F_1 est une loi normale. Ils montrent que la règle de décision de Benjamini et Hochberg n'est pas optimale si on connaît F_1 , mais elle n'en est pas très loin.

Ils étudient le comportement asymptotique des processus $V(t)$ et $S(t)$ et montrent la convergence vers des processus gaussiens quand $m \rightarrow \infty$.

Ils obtiennent des résultats sur le contrôle de $\frac{V(t)}{V(t) + S(t)}$ par trois méthodes différentes (bootstrap, asymptotique et calcul exact). La méthode asymptotique semble bien adaptée dans la mesure où m est généralement très grand. Ce résultat est important car il permet un contrôle du taux de faux positifs et pas seulement de son espérance, le FDR.

À la suite de ces travaux, d'autres auteurs ont explicitement pris en compte dans le modèle de mélange la contrainte liée aux contexte des tests multiples qui veut que $F_1 \in F_S$: Cheng *et al.* [4], Pan *et al.* [18], Liao *et al.* [15].

6. Conclusion et perspectives

Le domaine des tests multiples est très actif, dans de très nombreuses revues, allant des plus formelles aux plus appliquées. C'est un domaine de recherche

en statistique mathématique qui peut avoir beaucoup de retombées pour les utilisateurs dans de nombreux domaines allant largement au-delà des microarrays. Les résultats déjà obtenus sont à la fois très riches et très utiles. Cependant on rencontre certaines limites qui suscitent actuellement de nouvelles recherches.

Parmi les pistes actuelles on peut citer les suivantes :

Tous les modèles proposés postulent que les probabilités critiques suivent une loi uniforme sous H_0 . Dans le monde réel c'est loin d'être le cas. Ces probabilités critiques sont calculées sous un modèle statistique qui n'est pas complètement réaliste. Les hypothèses faites (normalité, homoscedasticité, absence de valeurs aberrantes...) sont souvent partiellement fausses. Les tests de permutation qui paraissent moins chargés en hypothèses préalables ne sont pas sans défaut, bien au contraire. On dispose en général de petits échantillons et l'ensemble des permutations possibles est bien trop petit pour permettre d'obtenir des probabilités critiques correctes et distribuées selon une loi uniforme. Notre expérience est qu'il vaut mieux les éviter et préférer un modèle paramétrique souple et adaptatif comme celui proposé par Delmar *et al.* [6]. Dans le cas où les probabilités critiques ne sont pas distribuées selon une loi uniforme, on peut s'attendre à ce que les modèles de mélange les plus sophistiqués intègrent dans la distribution f_1 ces défauts d'uniformité, ce qui conduit à des résultats faussés. Il faut donc privilégier les méthodes qui sont robustes à la non-uniformité des probabilités critiques sous H_0 .

On peut considérer que le problème des tests multiples est un problème de classification non supervisée. En pratique on dispose souvent de quelques gènes dont la différence d'expression est connue par d'autres études et d'autres méthodes expérimentales. On peut ainsi prendre en compte cette information pour améliorer les estimations des paramètres et des densités. On est alors dans le cadre de la classification semi-supervisée, ce qui nécessite de nouveaux développements théoriques.

Remerciements : Nous remercions un relecteur anonyme pour ses remarques pertinentes et précises qui ont beaucoup amélioré la présentation de ce texte.

Références

- [1] ALLISON D.B., GADBURY G.L., MOONSEONG Heo, FERNÁNDEZ J.R., CHEOL-KOO Lee, PROLLA T.A., WEINDRUCH R. (2002). A Mixture Model Approach for the Analysis of Microarray Gene Expression Data, *Computational Statistics & Data Analysis*, 39, 1-20.
- [2] AUBERT J., BAR-HEN A., DAUDIN J.J., ROBIN S (2004). Determination of the differentially expressed genes in microarray experiments using local FDR, *BMC Bioinformatics*, 5, 125.
- [3] BENJAMINI Y., HOCHBERG Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*, 57 (1), 289-300.
- [4] CHENG C., POUNDS S.B., BOYETT J.M., PEI D., KUO M.L., ROUSSEL M.F. (2004). Statistical Significance Threshold Criteria For Analysis of Microarray

- Gene Expression Data, *Statistical Applications in Genetics and Molecular Biology*, Vol. 3, No. 1, Article 36.
- [5] DALMASSO C., BROËT P., MOREAU T. (2005). A simple procedure for estimating the false discovery rate, *Bioinformatics*, 21, 660-668.
 - [6] DELMAR P., ROBIN S., DAUDIN J.J. (2005). VarMixt : efficient variance modelling for the differential analysis of replicated gene expression data, *Bioinformatics*, 21, 502-508.
 - [7] DUDOIT S., YANG Y.H., CALLOW M.J., SPEED T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, 12,1.
 - [8] DUDOIT S., SHAFFER J.P., BOLDRICK J.C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, 18 :1, 71-103.
 - [9] DUDOIT S., VAN DER LAAN M.J., POLLARD K.S. (2004). Multiple hypothesis. Part I. Single-Step procedures for control of general type I error rates, *Statistical Applications in Genetics and Molecular Biology*, Vol 3, No. 1, article 13.
 - [10] EFRON B., TIBSHIRANI R., STOREY J.D., TUSHER V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, 96, 1151-1160.
 - [11] EFRON B. (2004), Large scale simultaneous hypothesis testing : the choice of a null hypothesis, *Journal of the American Statistical Association*, 99, 96-104.
 - [12] GENOVESE .C, WASSERMAN L. (2002). Operating Characteristics and Extensions of the FDR Procedure, *Journal of the Royal Statistical Society Series B*, 64 (1), 499-518.
 - [13] GENOVESE C., WASSERMAN L. (2004). A stochastic process approach to false discovery control, *Ann. Statist.*, 32, 3, 1035-61.
 - [14] HEDENFALK I., DUGGAN D., CHEN Y., RADMACHER M., BITTNER M., SIMON R., MELTZER P., GUSTERSON B., ESTELLER M., KALLIONIEMI O.P. *et al.* (2001). *N. Engl.J.Med.*, 344, 539-548.
 - [15] LIAO J.G., LIN Y., SELVANAYAGAM Z.E., SHIH W.J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis, *Bioinformatics*, 20, 2694-2701.
 - [16] POUNDS S., CHENG C. (2004). Improving false discovery rate estimation, *Bioinformatics*, 20, 1737-1745.
 - [17] POUNDS S., MORRIS S.W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values, *Bioinformatics*, 19, 1236-1242.
 - [18] PAN W., LIN J., LE C. (2003). A Mixture model approach to detecting differentially expressed genes with microarray data, *Functional Integrative Genomics*, 3, 117-124.
 - [19] STOREY J.D., TAYLOR J.E., SIEGMUND D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates : A unified approach, *Journal of the Royal Statistical Society Series B*, 66, 187-205.
 - [20] STOREY J.D., TIBSHIRANI R. (2003). Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences*, 100, 16, 9440-45.