

CYRIL DALMASSO

PHILIPPE BROËT

**Procédures d'estimation du False Discovery Rate basées
sur la distribution des degrés de signification**

Journal de la société française de statistique, tome 146, n° 1-2 (2005),
p. 63-75

http://www.numdam.org/item?id=JSFS_2005__146_1-2_63_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/>

PROCÉDURES D'ESTIMATION DU FALSE DISCOVERY RATE BASÉES SUR LA DISTRIBUTION DES DEGRÉS DE SIGNIFICATION

Cyril DALMASSO *, Philippe BROËT **, (1)

RÉSUMÉ

L'objectif de cet article est de présenter pour un public mixte biologiste/biostatisticien plusieurs méthodes couramment utilisées pour l'estimation du FDR (basées sur la distribution des degrés de signification) en génomique fonctionnelle.

Mots clés : Génomique fonctionnelle, taux de faux positifs, tests multiples.

ABSTRACT

The aim of this paper is to present several methods that are actually used for estimating the False Discovery Ratio (FDR) in functional genomics.

Keywords : False discovery rate, functional genomics, multiple testing.

1. Introduction

La génomique fonctionnelle, qui est l'étude de la fonction des gènes à partir de leur expression et/ou de leurs produits d'expression (les ARNm et les protéines) connaît un développement fulgurant depuis l'apparition des nouvelles biotechnologies (e.g. puces d'expression, puces d'hybridation génomique comparative, puces protéiques, ...). Les premières utilisations pratiques de ces nouvelles biotechnologies datent du milieu des années quatre-vingt dix et ont, en particulier, concerné l'étude du transcriptome à partir de puces d'expression par dépôts (lame de verre, membrane de nylon) ou synthèse in situ ([13]). Actuellement ces nouvelles technologies sont de plus en plus fréquemment intégrées dans les programmes de recherche bioclinique, où l'un des principaux objectifs est l'identification des gènes dont les modifications sont liées à un facteur biologique ou clinique particulier. En cancérologie, par exemple, de nombreux programmes de recherche portant sur des phénomènes tels que la chimiorésistance des tumeurs solides ou sur l'identification de sous-groupes pronostiques de patients utilisent ces nouvelles technologies.

* INSERM U472, 16 Avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France.

** Faculté de Médecine Paris-Sud, Le Kremlin-Bicêtre, France.

(1) Auteur correspondant : broet@vjf.inserm.fr.

La majorité des études actuellement réalisées concerne le transcriptome. Dans ce cadre, les puces d'expression permettent de mesurer simultanément l'activité transcriptionnelle de plusieurs milliers de gènes pour un échantillon biologique donné. En pratique, les résultats issus de ces nouvelles technologies posent à l'expérimentateur un problème majeur quant à la prise en compte du grand nombre de comparaisons effectuées simultanément. En effet, le contrôle du nombre de fausses découvertes (gènes déclarés comme modifiés alors qu'ils ne le sont pas) est une nécessité compte tenu du coût humain et financier engendré par les futures études confirmatoires.

Dans ce cadre de comparaisons multiples dites à «très grande échelle», certains auteurs se sont intéressés au contrôle du taux global d'erreur ou *Family Wise Error Rate* (FWER) [9], défini dans notre contexte comme la probabilité de sélectionner à tort un ou plusieurs gènes sur l'ensemble des gènes étudiés. Cependant, les procédures contrôlant ce critère sont d'autant moins puissantes que le nombre d'hypothèses testées est important, et malgré les nombreux développements méthodologiques récemment proposés ([7]), elles s'avèrent parfois inopérantes dans le cadre des études de génomique fonctionnelle. En 1995, Benjamini et Hochberg ont proposé un nouveau critère d'erreur dénommé taux de fausses découvertes ou *False Discovery Rate* (FDR) [1]. Notant V le nombre total de faux positifs et R le nombre total de positifs (gènes déclarés intéressants), le FDR est défini comme l'espérance du taux de faux positifs : $FDR = E \left[1_{\{R>0\}} \left(\frac{V}{R} \right) \right]$.

À la différence de situations décisionnelles telles que les essais thérapeutiques pour lesquels plusieurs critères de jugement sont considérés et pour lesquels le FWER semble le seul critère d'erreur envisageable, pour des études de génomique fonctionnelle, un critère tel que le FDR semble relativement bien adapté. En effet, si dans le cadre d'un essai thérapeutique d'efficacité, il est nécessaire d'être «à 95% confiant» de l'efficacité réelle globale d'un nouveau traitement pour décider de son utilisation en pratique courante, dans le cadre des études de génomique fonctionnelle, l'information apportée par l'analyse des données d'expression ne représente que l'une des composantes d'une stratégie exploratoire prenant également en compte des informations extérieures à l'étude (e.g. : fonctions biologiques, localisation,...). Dans ce contexte exploratoire, le FDR, qui conduit à des procédures beaucoup moins restrictives que le FWER, semble un critère plus opérationnel.

Les méthodes les plus couramment utilisées sont les procédures d'estimation du critère basées sur la distribution marginale des degrés de signification sans hypothèse sur leurs distributions dans le cas d'une liaison entre le gène et le critère considéré. Dans ce cadre, quatre procédures dénommées QVALUE [15], BUM [11], SPLOSH [10] et LBE [4] (implémentées dans R) sont actuellement proposées à l'investigateur. L'objectif de cet article est de présenter, illustrer et discuter ces procédures. La section suivante présente le cadre méthodologique dans lequel se situent ces procédures qui sont décrites dans la troisième section. La quatrième section présente les résultats d'une étude de simulations et la cinquième section illustre l'utilisation pratique de ces procédures. Une conclusion clôture l'article.

2. Cadre méthodologique

Considérons l'étude de la liaison entre m gènes et un critère biologique ou clinique. Il s'agit d'effectuer m tests pour les hypothèses nulles et alternatives suivantes :

H_0 : « absence de liaison entre le gène et le critère considéré »

H_1 : « liaison entre le gène et le critère considéré »

On notera que H_0 est ici considérée comme une hypothèse nulle unique, identique pour tous les gènes, alors que H_1 peut, en revanche, être multiple (ou composite) et différente selon le gène considéré.

Pour chaque gène, une statistique de test T est calculée, ainsi que le degré de signification P correspondant. Une procédure de test conduisant à rejeter ou non l'hypothèse nulle est ensuite appliquée à l'ensemble des gènes. Au terme de la procédure, un nombre inconnu V de faux positifs se trouve parmi les R gènes pour lesquels l'hypothèse nulle est rejetée.

Alors que les premières procédures relatives au FDR introduites par Benjamini et coll. ([1], [2], [12]) avaient pour objectif le contrôle de ce critère, de nombreuses procédures visant à l'estimer ont récemment été proposées. Ces dernières reposent sur la relation entre le FDR et le positive False Discovery Rate ($pFDR$), défini comme l'espérance du taux de fausses découvertes conditionnelle à l'existence d'au moins une découverte ($pFDR = E\left(\frac{V}{R} | R > 0\right)$) [14]. Ce critère, intuitivement plus pertinent que le FDR , n'avait pas été retenu par Benjamini et Hochberg dans leur article princeps (1995) car il n'est pas contrôlable. En effet, sous l'hypothèse nulle complète (*i.e.* lorsque l'hypothèse nulle H_0 est vraie pour tous les gènes), tous les positifs sont nécessairement des faux positifs, et le $pFDR$ est alors égal à 1. L'hypothèse nulle complète ne pouvant être écartée, il est impossible d'assurer que le $pFDR$ est inférieur à un seuil α fixé plus petit que 1.

Cependant, par définition, le $pFDR$ est lié au FDR par la relation $pFDR = \frac{FDR}{\Pr(R > 0)}$. Ces deux critères sont donc asymptotiquement équivalents puisque lorsque le nombre d'hypothèses testées tend vers l'infini, $\Pr(R > 0)$ tend vers 1. Dans la suite de cet article et pour simplifier l'exposé, nous utilisons indistinctement les notions de FDR et de $pFDR$.

Dans le cadre des procédures pour lesquelles aucune hypothèse distributionnelle n'est faite sous H_1 , l'estimation du FDR est généralement basée sur la relation suivante démontrée par Storey en 2001. En supposant les statistiques de test indépendantes et identiquement distribuées sous H_0 , pour une région de rejet Γ fixée, la même pour tous les gènes, le $pFDR$ peut s'écrire comme une probabilité conditionnelle :

$$pFDR(\Gamma) = \Pr(H = 0 | T \in \Gamma) = \frac{\Pr(H = 0) \Pr(T \in \Gamma | H = 0)}{\Pr(T \in \Gamma)}$$

où T est la statistique de test et H est la variable indicatrice telle que $H = 0$ si l'hypothèse nulle H_0 est vraie, $H = 1$ si l'hypothèse alternative H_1 est vraie [14].

Considérons la statistique de test $P = \Phi(T)$ où Φ est la transformation permettant d'obtenir, à partir de la statistique T , le degré de signification associé P . Sous l'hypothèse nulle, la variable aléatoire P suit une distribution uniforme sur l'intervalle $[0, 1]$, et si l'on considère une région de rejet de la forme $\Gamma = [0, \gamma]$, $\gamma \in [0, 1]$, alors $\Pr(P \in \Gamma | H = 0) = \gamma$. L'estimation du $pFDR(\Gamma)$ repose donc sur l'estimation de la quantité : $\gamma \Pr(H = 0) [\Pr(P \in \Gamma)]^{-1}$.

Dans le cas des études de génomique fonctionnelle, où le nombre de gènes étudiés est de l'ordre de plusieurs milliers, Storey a proposé d'estimer $pFDR(\Gamma)$ par l'estimation séparée de $\Pr(H = 0)$ et de $\Pr(P \in \Gamma)$. En pratique, $\Pr(P \in \Gamma)$ est estimé par la proportion de gènes pour lesquels le degré de signification observé est inférieur ou égal à γ . La principale difficulté est l'obtention d'un estimateur de $\Pr(H = 0)$.

En notant f_0 et f_1 les densités conditionnelles de P sous H_0 et sous H_1 , et $\pi_0 = \Pr(H = 0)$, la densité marginale de P , notée f , s'écrit :

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p). \quad (1)$$

Sous l'hypothèse nulle, les P - values sont distribuées selon une loi uniforme sur $[0, 1]$ d'où $f_0(p) = 1_{[0,1]}(p)$, ce qui implique : $f(p) = \pi_0 + (1 - \pi_0) f_1(p)$ où f_1 et π_0 sont inconnus. La quantité $(1 - \pi_0) f_1(p)$ étant positive, si l'on suppose que f_1 (qui peut elle-même être la densité d'un mélange) est décroissante pour $p \in [0, 1]$, alors d'après (1) $f(1) = \pi_0 + (1 - \pi_0) f_1(1)$ est le plus petit majorant de π_0 . Ainsi, un estimateur sans biais de $f(1)$ est un estimateur de π_0 dont le biais est positif. Les trois procédures QVALUE, BUM et SPLOSH sont basées sur une estimation de $f(1)$. La procédure LBE, en revanche, est basée sur une estimation de l'espérance d'une transformation des P - values.

3. Procédures Qvalue, BUM, Splosh, LBE

3.1. Qvalue

L'un des estimateurs de π_0 les plus largement utilisés est celui proposé par Storey et Tibshirani en 2003 [15]. Soit $\lambda \in [0, 1]$, π_0 peut être estimé par :

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)}.$$

En considérant la limite lorsque λ tend vers 1, cet estimateur d'un majorant de π_0 est un estimateur de $f(1)$. En effet, en notant F la fonction de répartition de P , il peut s'écrire :

$$\hat{\pi}_0(\lambda) = \frac{1 - \hat{F}(\lambda)}{1 - \lambda}.$$

Ainsi, la quantité estimée est :

$$\lim_{\lambda \rightarrow 1} \hat{\pi}_0(\lambda) = \lim_{\lambda \rightarrow 1} \frac{1 - \hat{F}(\lambda)}{1 - \lambda} = \hat{f}(1).$$

Plutôt que de choisir une valeur particulière de λ proche de 1 pour laquelle l'estimateur de π_0 aurait nécessairement une grande dispersion, Storey et Tibshirani ont proposé d'utiliser une méthode de lissage basée sur des splines cubiques afin d'estimer la fonction $\hat{\pi}_0(\lambda)$ et d'obtenir ainsi une extrapolation de $\hat{\pi}_0(1)$. Le principe de cette méthode peut être résumé comme suit : pour un ensemble de valeurs t_i ($i = 1, \dots, k$) appartenant à l'intervalle $[0, 1]$ telles que $t_1 \leq t_2 \leq \dots \leq t_k$ (l'entier k étant choisi de façon arbitraire), la fonction $\hat{\pi}_0(\lambda)$ est estimée par une fonction $\tilde{\pi}_0(\lambda)$ de classe C^2 telle que chaque restriction sur $[t_1, t_2], \dots, [t_{k-1}, t_k]$ est un polynôme d'ordre 3. L'estimation de π_0 est ensuite obtenue en prenant $\lim_{\lambda \rightarrow 1} \tilde{\pi}_0(\lambda)$.

3.2. BUM (Beta-Uniform Mixture)

L'estimateur de π_0 proposé par Pounds et Morris [11] est basé sur une modélisation de la distribution marginale des degrés de signification par le mélange d'une loi bêta $[B(a, 1)]$ et d'une loi uniforme dont la densité s'écrit :

$$f(p|a, \omega) = \omega + (1 - \omega)ap^{a-1} \text{ où } \omega \in [0, 1].$$

L'estimation des paramètres de ce modèle est basée sur la méthode du maximum de vraisemblance. On notera que dans le but d'améliorer les performances de la méthode d'optimisation numérique utilisées les auteurs ont proposé d'effectuer le changement de variable suivant : $\psi = \text{logit}(a)$ et $\phi = \text{logit}(\lambda)$. Les valeurs $\hat{\psi}$ et $\hat{\phi}$ maximisant la log-vraisemblance sont ensuite obtenues par une méthode d'optimisation numérique [6], puis, les estimations de a et λ en sont déduites $\left(\hat{a} = \frac{\exp(\hat{\psi})}{1 + \exp(\hat{\psi})} \text{ et } \hat{\lambda} = \frac{\exp(\hat{\phi})}{1 + \exp(\hat{\phi})} \right)$. Enfin, π_0 est estimé par $\hat{\pi}_0 = \hat{f}(1) = \hat{\lambda} + (1 - \hat{\lambda})\hat{a}$.

À la différence de la précédente procédure, la méthode proposée par Pounds et Morris repose sur de fortes hypothèses sur la distribution marginale des degrés de signification.

3.3. SPLOSH (SPacing LOessS Histogram)

La procédure d'estimation de π_0 proposé par Pounds et Cheng [10] repose sur une méthode de régression locale (LOESS) visant à estimer $f(p)$ après transformation des degrés de signification. Les principales étapes de la procédure peuvent être résumées comme suit : (i) Les gènes sont d'abord ordonnés suivant leurs degrés de signification ($p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$) et la fonction de répartition est estimée pour chaque valeur de $p_{(i)}$ ($i = 1, \dots, m$) par : $Pr(P \leq p_{(i)}) = a_i = \left(i - \frac{1}{2}\right) / m$. (ii) La densité en un point m_i , milieu de l'intervalle $[p_{(i)}, p_{(i+1)}]$ est ensuite estimée par l'accroissement δ_i de la fonction

de répartition empirique sur cet intervalle : $\delta_i = \frac{a_{(i+1)} - a_{(i)}}{P_{(i+1)} - P_{(i)}}$. (iii) Enfin, la densité en tout point est estimée par une méthode de régression polynomiale locale dénommée LOESS [3] appliquée aux ensembles obtenus par le changement de variable $x_i = \arcsin \left[2 \times \left(p_i - \frac{1}{2} \right) \right]$ et $y_i = \ln(\delta_i)$.

Cette dernière étape consiste à effectuer, pour toute valeur x , une régression polynomiale d'ordre 2 pour un voisinage de ce point (ensemble des couples (x_i, y_i) tels que $|x_i - x| \leq e_1$, où e_1 est arbitrairement fixé), la régression étant basée sur un critère des moindres carrés pondérés. Le changement de variable $x_i = \arcsin \left[2 \times \left(p_i - \frac{1}{2} \right) \right]$, qui conduit à éloigner les valeurs extrêmes des valeurs centrales, permet de limiter l'influence de ces valeurs centrales dans l'estimation de la densité des degrés de signification proches de 0 ou de 1. Le changement de variables $y_i = \ln(\delta_i)$ est quant à lui considéré afin d'uniformiser les écarts entre les différentes valeurs de δ_i . Après avoir effectué la transformation inverse : $\widehat{f}^*(p_{(i)}) = \exp[\widehat{y}(x_i)]$ qui permet de passer de $y = \widehat{y}(x)$ à $\delta = f(p)$, la densité $f(p)$ est estimée par le rapport entre la fonction estimée $\widehat{f}^*(p)$ et une estimation de $\int_0^1 \widehat{f}^*(p) dp$ (méthode des trapèzes), de manière à ce que $\int_0^1 \widehat{f}(p) dp \approx 1$. Finalement, π_0 est estimé par $\widehat{\pi}_0 = \min_{1 \leq i \leq m} \widehat{f}(p_i) \approx \widehat{f}(1)$.

3.4. LBE (Location Based Estimator)

Alors que les trois procédures précédentes sont basées sur une estimation de la quantité $f(1)$, la procédure LBE [4] est basée sur l'espérance d'une transformation des degrés de signification.

À partir de la relation (1), on déduit aisément la relation suivante :

$$\frac{E(P)}{E_0(P)} = \pi_0 + (1 - \pi_0) \frac{E_1(P)}{E_0(P)}$$

où $E_0(P)$ et $E_1(P)$ sont les espérances de la distribution conditionnelle de P sous l'hypothèse nulle et alternative respectivement.

Sous l'hypothèse nulle, les P -values sont uniformément distribuées, de sorte que $E_0(P) = \frac{1}{2}$. Ainsi, $2 \frac{1}{m} \sum_{i=1}^m P_i$ est un estimateur sans biais de $\frac{E(P)}{E_0(P)}$. C'est donc un estimateur de π_0 dont le biais est positif.

À partir de cet estimateur, une transformation des degrés de signification permet d'obtenir un estimateur de π_0 moins biaisé. En effet, il a été montré ([4]) que si φ est une fonction continue vérifiant les quatre conditions suivantes :

- (i) : $\lim_{x \rightarrow 1} \varphi(x) = +\infty$; (ii) : $\lim_{x \rightarrow 0} \varphi(x) < +\infty$; (iii) : φ convexe;
- (iv) : $\varphi(E_0(P)) \geq E_0(P)$

alors le biais de l'estimateur $\hat{\pi}_0 = \frac{1}{m} \sum_{i=1}^m \varphi(p_i) - E_0[\varphi(P)]$ est inférieur à celui de $2 \frac{1}{m} \sum_{i=1}^m P_i$.

Dans l'ensemble des fonctions vérifiant ces conditions, les fonctions de la forme $\varphi(P) = -\ln(1-x)^n$, $n \in \mathbb{N}$ ont été considérées. Il a été montré ([4]) que ces fonctions conduisent à une famille d'estimateurs de π_0 dont le biais est décroissant avec n . Comme sous l'hypothèse nulle, $-\ln(1-P)$ suit une loi exponentielle de paramètre 1, $E_0([-\ln(1-P)]^n) = n!$ et la famille d'estimateurs proposée est :

$$\hat{\pi}_{0(n)} = \frac{\frac{1}{m} \sum_{i=1}^m [-\ln(1-p_i)]^n}{n!}, \quad n \in \mathbb{N}.$$

Pour cette famille d'estimateurs, des résultats théoriques ont été obtenus sur la distribution asymptotique. En effet, d'après le théorème central limite, $\hat{\pi}_{0(n)}$ suit asymptotiquement une loi normale d'espérance $E(P)/E_0(P)$ et de

variance $Var(\hat{\pi}_0)$, dont un majorant s'écrit $\frac{\binom{n}{2n} - 1}{m}$.

Comme le biais est décroissant avec n , alors que la variance est croissante, la règle pratique proposée pour choisir un estimateur particulier dans la famille d'estimateurs consiste, pour un nombre spécifique m d'hypothèses testées, à choisir n en fonction d'un seuil arbitraire l pour le majorant de la variance asymptotique. De manière plus formelle,

$$n = \max \left(1, \max(n \in \mathbb{N}^* \mid \frac{\binom{n}{2n} - 1}{m} \leq l) \right).$$

4. Étude de simulations

4.1. Plan de simulation

Une étude de simulations a été menée afin de comparer les propriétés des quatre estimateurs de π_0 : QVALUE, BUM, SPLOSH et LBE.

Les données ont été générées de façon à mimer une expérience comparant deux groupes de dix échantillons soumis à deux conditions différentes. Pour chaque pseudo-gène i ($i = 1, \dots, m$ avec $m = 500$ ou 2000) et chaque échantillon k ($k = 1, \dots, 10$) sous la condition j ($j = 1, 2$), la valeur simulée $X_{i,j,k}$ correspond à une pseudo-intensité (ou rapport d'intensité) normalisée (pour une étude des méthodes de normalisation des données d'expression voir [16]).

Pour chaque échantillon k , les valeurs ont été générées indépendamment à partir de distributions normales, $X_{i,j,k} \sim N(\mu_{ij}, 1)$. Pour le premier groupe ($j = 1$), toutes les valeurs ont été simulées avec $\mu_{i1} = 0$. Pour le deuxième groupe, une proportion π_0 de valeurs a été simulée avec $\mu_{i2} = 0$ (« gènes non modifiés »). Les valeurs correspondant aux gènes modifiés ont été simulées suivant 3 configurations différentes : (a) $\mu_{i2} = 1$ pour tous les gènes modifiés ; (b) $\mu_{i2} = 2$ pour tous les gènes modifiés ; (c) $\mu_{i2} = 1$ pour la moitié des gènes modifiés, $\mu_{i2} = 2$ pour l'autre moitié. Les différentes valeurs de π_0 considérées sont : $\pi_0 = 0, 2, 0,5$ et $0,8$.

Dans chaque cas, les P -values, calculées sous l'hypothèse nulle $H_0 : \mu_{i1} = \mu_{i2}$, ont été obtenues à partir de la statistique de Student, puis nous avons estimé π_0 à partir des quatre procédures QVALUE, BUM, SPLOSH and LBE. La règle présentée précédemment pour choisir la valeur n de l'estimateur LBE conduit à choisir $n = 1$ pour $m = 500$ et $n = 2$ pour $m = 2000$.

Pour chaque configuration, 1000 matrices de données ont été simulées. Les moyennes, écarts-type et erreurs quadratiques moyennes des quatre estimateurs ont été estimées à partir des 1000 résultats ainsi obtenus.

4.2. Résultats

Le tableau 1 présente les moyennes des quatre estimateurs (pour chaque configuration). Il montre qu'alors que tous les estimateurs sont supposés avoir un biais positif, les procédures BUM et SPLOSH conduisent à sous estimer π_0 dans la majorité des cas simulés. Par exemple, pour la configuration (b) et avec $\pi_0 = 80\%$ et $m = 500$, les moyennes des estimations obtenues par les procédures BUM et SPLOSH sont $\hat{\pi}_0 = 71\%$ et $\hat{\pi}_0 = 55\%$. En revanche, la moyenne des estimations obtenues par les procédures QVALUE et LBE ne sont légèrement inférieures à π_0 que dans quelques cas. Par exemple, pour $m = 500$, $\pi_0 = 0.5$ et pour la configuration (b), l'estimation obtenue par la procédure QVALUE (0.495) est légèrement inférieure à 0.5. Notons cependant que l'espérance des estimateurs est estimée à partir de 1000 simulations, et que cette valeur (0.495) n'est pas incompatible avec une espérance théorique supérieure à 50%. Les estimations de π_0 obtenues avec la méthode LBE sont supérieures à celles fournies par QVALUE dans tous les cas (excepté un). Cependant, la différence n'est jamais plus grande que 8.7%.

Le tableau 2, qui présente les estimations des écarts-type pour chacune des méthodes, montre que l'écart-type de l'estimateur LBE est toujours inférieur à celui de l'estimateur QVALUE (la plus petite différence est de 1.8%). De plus, les écarts-type estimés pour LBE sont inférieurs au majorant calculé à partir des résultats théoriques. En effet, pour $m = 500$ et $m = 2000$, les valeurs calculées sont 0.045 et 0.05.

Le tableau 3 présente les erreurs quadratiques moyennes de chaque estimateur. En comparaison de l'estimateur QVALUE, l'estimateur LBE a la plus petite erreur quadratique moyenne dans 14 cas sur 18. Pour quelques cas, les estimateurs BUM et SPLOSH ont une erreur quadratique moyenne plus petite, cependant, ces derniers résultats sont difficiles à interpréter, compte tenu du fait que ces estimateurs tendent à sous estimer π_0 dans de nombreux cas.

FALSE DISCOVERY RATE : DISTRIBUTION DES DEGRÉS DE SIGNIFICATION

TABLEAU 1. - Moyenne des estimations (pour chaque configuration) pour les méthodes QVALUE, BUM, SPLOSH et LBE avec $n = 1$ et $n = 2$.

m	π_0	$conf.$	<i>QVALUE</i>	<i>BUM</i>	<i>SPLOSH</i>	<i>LBE</i>
500	0.2	(a)	0.251933	0.321501	0.236210	0.338584
		(b)	0.197976	0.142870	0.156950	0.203072
		(c)	0.223192	0.197882	0.230946	0.270274
	0.5	(a)	0.536112	0.440906	0.486246	0.586282
		(b)	0.495076	0.365937	0.418347	0.500949
		(c)	0.513054	0.430535	0.479036	0.543493
	0.8	(a)	0.806984	0.748141	0.671984	0.832719
		(b)	0.800555	0.705589	0.553156	0.801681
		(c)	0.808455	0.749179	0.634703	0.817921
2000	0.2	(a)	0.252816	0.320720	0.253153	0.281443
		(b)	0.198622	0.142818	0.170251	0.200213
		(c)	0.225407	0.197825	0.255394	0.241148
	0.5	(a)	0.533424	0.436461	0.524105	0.550845
		(b)	0.499708	0.366203	0.473210	0.499902
		(c)	0.515855	0.431369	0.526475	0.526171
	0.8	(a)	0.810799	0.751605	0.739347	0.818360
		(b)	0.797784	0.705598	0.588101	0.799031
		(c)	0.803206	0.750330	0.708212	0.807745

TABLEAU 2. - Écart-type des estimations (pour chaque configuration) pour les méthodes QVALUE, BUM, SPLOSH et LBE avec $n = 1$ et $n = 2$.

m	π_0	$conf.$	<i>QVALUE</i>	<i>BUM</i>	<i>SPLOSH</i>	<i>LBE</i>
500	0.2	(a)	0.067970	0.009311	0.054648	0.026380
		(b)	0.055716	0.002369	0.017699	0.019997
		(c)	0.061759	0.003785	0.034406	0.023189
	0.5	(a)	0.091064	0.020922	0.069278	0.033387
		(b)	0.091746	0.013303	0.034630	0.032180
		(c)	0.092356	0.020756	0.068751	0.034396
	0.8	(a)	0.109255	0.031400	0.108568	0.039787
		(b)	0.109257	0.013100	0.082031	0.039990
		(c)	0.113246	0.019601	0.102903	0.041502
2000	0.2	(a)	0.034272	0.004592	0.030933	0.027721
		(b)	0.028699	0.001148	0.007631	0.023457
		(c)	0.030548	0.002085	0.016979	0.024628
	0.5	(a)	0.045018	0.010728	0.031426	0.035501
		(b)	0.043649	0.006624	0.021714	0.034110
		(c)	0.045579	0.009997	0.034938	0.035719
	0.8	(a)	0.055588	0.016036	0.069977	0.044740
		(b)	0.056874	0.006834	0.042342	0.046171
		(c)	0.056032	0.009539	0.065325	0.044640

TABLEAU 3. – Erreur quadratique moyenne des estimations (pour chaque configuration) pour les méthodes QVALUE, BUM, SPLOSH et LBE avec $n = 1$ et $n = 2$.

m	π_0	<i>conf.</i>	<i>QVALUE</i>	<i>BUM</i>	<i>SPLOSH</i>	<i>LBE</i>
500	0.2	(a)	0.007313	0.014849	0.004295	0.019901
		(b)	0.003105	0.003270	0.002166	0.000409
		(c)	0.004349	0.000019	0.002140	0.005476
	0.5	(a)	0.009588	0.003929	0.004984	0.008558
		(b)	0.008434	0.018150	0.007865	0.001035
		(c)	0.008691	0.005256	0.005162	0.003074
	0.8	(a)	0.011974	0.003674	0.028164	0.002652
		(b)	0.011925	0.009085	0.067654	0.001600
		(c)	0.012883	0.002967	0.037901	0.002042
2000	0.2	(a)	0.003963	0.014594	0.003781	0.007401
		(b)	0.000825	0.003271	0.000943	0.000549
		(c)	0.001577	0.000009	0.003357	0.002299
	0.5	(a)	0.003142	0.004152	0.001567	0.003844
		(b)	0.001903	0.017945	0.001189	0.001162
		(c)	0.002327	0.004810	0.001920	0.001959
	0.8	(a)	0.003204	0.002599	0.008570	0.002337
		(b)	0.003237	0.008958	0.046692	0.002131
		(c)	0.003147	0.002558	0.012688	0.002051

5. Application

Dans cette partie, nous illustrons les différentes procédures présentées précédemment en les appliquant aux données issues d'une étude de génomique fonctionnelle réalisée par Hedenfalk *et coll.* [8] visant à mettre en évidence les gènes différemment exprimés entre deux types de tumeurs du sein héréditaires liées à une mutation du gène BRCA1 ou du gène BRCA2.

Les expressions de sept tumeurs ($n_1 = 7$) ayant une altération pour le gène BRCA1 et de huit tumeurs ($n_2 = 8$) ayant une altération pour le gène BRCA2 ont été analysées par puces d'expression par dépôts (puces à ADN) sur lames de verre. L'analyse de Hedenfalk *et coll.* [8] portait sur 3 226 clones. Nous avons exclu de notre analyse les clones pour lesquels l'un des rapports (tumeur/référence) était inférieur à 0,1 ou supérieur à 10, ces valeurs étant considérées comme aberrantes par la plupart des expérimentateurs. Notre étude a donc porté sur les 3 030 gènes restants. Les données analysées peuvent se résumer sous la forme d'une matrice de dimension $n \times p$ où $n = 3 030$ (rapports d'intensités tumeur/référence) et $p = 15$ (prélèvements). Comme dans toute analyse de données issues de puces d'expression, une étape dite de « standardisation » a été effectuée ([16]) dont l'objectif est de rendre comparable les ensembles d'intensités obtenus sur les différentes puces

d'expression. La méthode utilisée ici procède d'un simple modèle d'analyse de la variance à partir des données log-transform

En considérant séparément chaque gène, la problématique est une simple comparaison de deux moyennes et la statistique de test est une statistique de Student à 13 degrés de liberté. Les degrés de signification utilisés pour chaque procédure ont été calculés à partir de la distribution de Student. Les procédures QVALUE, BUM, SPLOSH et LBE ont ensuite été appliquées afin d'estimer π_0 .

Les estimations obtenues pour π_0 par ces quatre procédures sont, respectivement : 0.669, 0.586, 0.622 et 0.688. Les estimations du FDR en ont ensuite été déduites.

La figure 1 présente l'évolution du *FDR* en fonction du nombre de gènes sélectionnés pour les quatre procédures. Elle montre que les estimations du *FDR* obtenues par les procédures QVALUE et LBE sont très proches et que les estimations du *FDR* obtenues par les procédures BUM et SPLOSH sont les plus petites, quel que soit le nombre de gènes sélectionnés. Cependant, les résultats de l'étude de simulation ont montré que ces deux dernières procédures ne garantissent pas l'obtention d'un majorant de π_0 . Or, une sous-estimation de π_0 conduit nécessairement à une sous-estimation du *FDR*.

On notera que pour la procédure LBE, un majorant de la variance asymptotique du FDR peut se déduire du majorant de la variance asymptotique de π_0 qui est ici égale à 1.65×10^{-3} (ce qui correspond à un écart-type de 4.06 %).

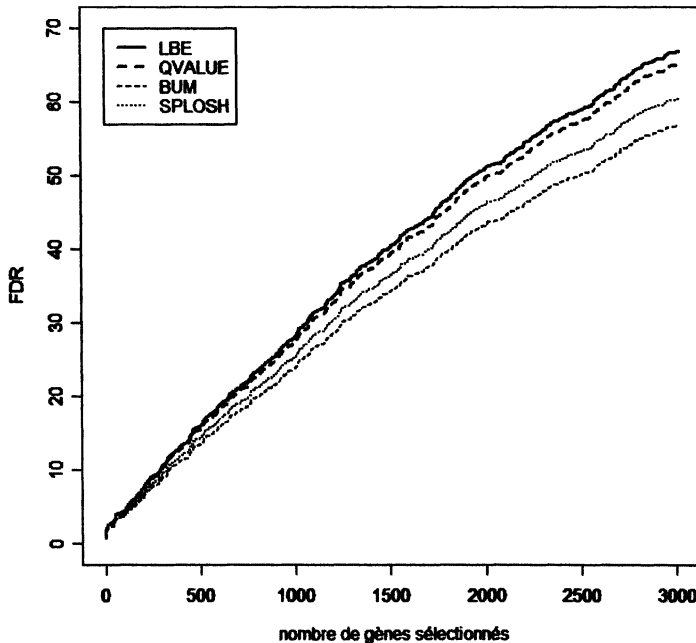


FIG 1. — Évolution du *FDR* en fonction du nombre de gènes sélectionnés pour les quatre procédures.

6. Conclusion

La problématique des comparaisons multiples est de plus en plus au centre des préoccupations des analyses de données utilisant les informations issues des biotechnologies de la génomique fonctionnelle.

Il existe actuellement de nombreuses procédures permettant d'estimer différents critères d'erreur relatifs aux faux positifs, parmi lesquels le *False Discovery Rate*, introduit par Benjamini et Hochberg. Les procédés d'estimation de ce critère basés sur la distribution des degrés de signification et ne faisant aucune hypothèse sur la distribution sous l'hypothèse alternative sont fréquemment employés. Dans ce cadre, l'estimation du *FDR* est toujours biaisée, et le biais théorique, positif, est d'autant plus important que la proportion π_0 de gènes vérifiant l'hypothèse nulle est faible et que les différences d'expression sont peu importantes. Cependant, la simplicité de ces procédures par rapport aux modélisations paramétriques ou semi-paramétriques de mélange (bayésienne ou fréquentiste) en ont fait un outil de choix en pratique courante.

Dans ce contexte et à notre connaissance, quatre méthodes sont actuellement proposées et implémentées. Nous en avons brièvement présenté les principes généraux. Les résultats de l'étude de simulations présentés dans cet article montrent que les estimateurs BUM et SPLOSH conduisent dans de nombreux cas à sous-estimer π_0 et par conséquent le FDR. Cette sous-estimation peut conduire l'investigateur à une sélection trop « optimiste » d'un ensemble de gènes en vue d'études confirmatoires. En revanche les estimateurs QVALUE et LBE ont de bonnes performances et peuvent être utilisés en pratique courante. Les simulations ont également montré les bons résultats, en terme d'erreur quadratique moyenne, de l'estimateur LBE qui, en outre, est le seul pour lequel la distribution asymptotique est connue. La connaissance d'un majorant de la variance asymptotique permet de plus d'apporter à l'investigateur une information sur la dispersion de π_0 ce qui n'est pas le cas des autres méthodes.

On notera que ces procédures sont basées sur l'hypothèse d'une distribution uniforme des degrés de signification sous H_0 , qui peut être incorrecte, en dépit de l'utilisation de méthodes de ré-échantillonnage [5]. De plus, les propriétés de ces estimateurs en cas de dépendance des degrés de signification restent à étudier.

En conclusion, les comparaisons multiples représentent actuellement un domaine de recherche très actif, comme en témoigne le nombre d'articles publiés récemment sur ce sujet, et constituent l'une des connexions importantes entre la méthodologie biostatistique et la biologie moléculaire.

Remerciements : Les auteurs remercient le Dr Thierry Moreau pour sa relecture attentive du manuscrit.

Références

- [1] BENJAMINI Y., HOCHBERG Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Stat Soc Ser. B*, **57**, 289–300.
- [2] BENJAMINI Y., YEKUTIELI D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
- [3] CLEVELAND W.S., DEVLIN S.J. (1988). Locally-weighted Regression : An Approach to Regression Analysis by Local Fitting. *J. Am. Statist. Assoc.*, **83**, 596–610.
- [4] DALMASSO C, BROËT P., MOREAU T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, (5), 660–668.
- [5] EFRON B. (2004). Large-scale simultaneous hypothesis testing : the choice of a null hypothesis. *J. Am. Statist. Assoc.*, **99**, 96–104.
- [6] GAY D.M. (1983). Algorithm 611. Subroutines for Unconstrained Minimization using a Model/Trust-Region Approach. *ACM Transactions on Mathematical Software*, **9**, 503–524.
- [7] GE Y., DUDOIT S., SPEED T.P. (2003). Resampling-based Multiple Testing for Microarray Data Analysis, *TEST*, **12**, 1, 1-66.
- [8] HEDENFALK I., DUGGAN D., CHEN Y., RADMACHER M., BITTNER M., SIMON R., MELTZER P., GUTERSON B., ESTELLER M., KALLIONIEMI O.P. *et al.* (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **22**, 539–548.
- [9] HOCHBERG Y., TAMHANE A. (1987). *Multiple comparison procedures*. Wiley & Sons, New-York.
- [10] POUNDS S., CHENG C. (2004). Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1745.
- [11] 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**, 1236–42.
- [12] REINER A., YEKUTIELI D., BENJAMINI Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.
- [13] SCHENA M. (2000). *Microarray biochip technology*. Bio Techniques Press.
- [14] STOREY J.D. (2001). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B*, **64** 479–498.
- [15] STOREY J.D., TIBSHIRANI R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.
- [16] SPEED T.P. (2003). *Statistical Analysis of Gene Expression Microarray Data*, CRC Press.