Contributions to the statistical analysis of microarray data.

Pierre Neuvial

Université Paris Diderot — Paris 7
UFR de Mathématiques

Année 2008

Thèse

Spécialité: Mathématiques Appliquées

Présentée par

Pierre NEUVIAL

Contributions à l'analyse statistique des données de puces à ADN

---

soutenue le 30 septembre 2008, devant le jury composé de:

Christophe AMBROISE	Univ. Évry	Examinateur
Emmanuel BARILLOT	Institut Curie	Co-directeur
Stéphane BOUCHERON	Univ. Paris Diderot	Directeur
Stéphane ROBIN	AgroParisTech	Rapporteur
Terry SPEED	Univ. of California, Berkeley	Rapporteur
Jean-Philippe VERT	Mines ParisTech	Examinateur
Mark van de WIEL	Vrije Univ., Amsterdam	Examinateur

INSERM, U900, Paris, F-75248
École des Mines de Paris, ParisTech, Fontainebleau, F-77300
Institut Curie, 26 rue d'Ulm, Paris cedex 05, F-75248 France

Laboratoire de Probabilités et Modèles Aléatoires
CNRS-UMR 7599, UFR de Mathématiques, case 7012
Université Paris Diderot (Paris 7)
2, place Jussieu, 75251 Paris Cedex 05

Chapter illustrations: courtesy of Prof. David Relman's lab, Stanford Univ.
`http://asiago.stanford.edu/RelmanLab/Microarray_art`

Résumé

Cette thèse traite de questions statistiques soulevées par l'analyse de données génomiques de grande dimension, dans le cadre de la recherche contre le cancer. La première partie est consacrée à l'étude des propriétés asymptotiques de procédures de tests multiples visant à contrôler l'espérance (FDR) du taux de fausses découvertes (FDP) parmi les hypothèses rejetées. On introduit un formalisme flexible qui permet de calculer la loi asymptotique du FDP et les conditions de régularité associées pour une vaste famille de procédures de tests multiples, et de comparer la puissance de ces procédures. On s'intéresse ensuite aux liens en termes de contrôle du FDR entre les bornes intrinsèques à trois problèmes de tests multiples: la détection, l'estimation, et la sélection. On relie en particulier la vitesse de convergence dans le problème d'estimation à la régularité de la loi des probabilités critiques au voisinage de 1.

La seconde partie est dédiée au développement de méthodes d'analyse des données de puces à ADN en cancérologie. On propose une méthode de pré-traitement des données de puces à ADN combinant une régression robuste et un modèle de mélange avec contrainte spatiale, qui permet d'éliminer les biais spatiaux en préservant le signal biologique. On développe ensuite une méthode d'inférence de régulations entre gènes à partir de données d'expression de gènes, qui repose sur des techniques d'apprentissage informatique et de tests multiples. Enfin, on construit un test génomique permettant de déterminer, pour une patiente traitée pour un cancer du sein, si un second cancer survenant sur le même sein est ou non une récidive du premier.

Mots-clés: Tests multiples, Méthode du delta fonctionnelle, Taux de fausses découvertes, Puces à ADN, Nombre de copies d'ADN, Normalisation, Réseaux de régulation.

Abstract

This thesis deals with statistical questions raised by the analysis of high-dimensional genomic data for cancer research. In the first part, we study asymptotic properties of multiple testing procedures that aim at controlling the False Discovery Rate (FDR), that is, the expected False Discovery Proportion (FDP) among rejected hypotheses. We develop a versatile formalism to calculate the asymptotic distribution of the FDP an the associated regularity conditions, for a wide range of multiple testing procedures, and compare their asymptotic power. We then study in terms of FDR control connections between intrinsic bounds between three multiple testing problems: detection, estimation and selection. In particular, we connect convergence rates in the estimation problem to the regularity of the p-value distribution near 1.

In the second part, we develop statistical methods to study DNA microarrays for cancer research. We propose a microarray normalization method that removes spatial biases while preserving the true biological signal; it combines robust regression with a mixture model with spatial constraints. Then we develop a method to infer gene regulations from gene expression data, which is based on learning and multiple testing theories. Finally, we build a genomic score to predict, for a patient treated for a breast tumor, whether or not a second cancer is a true recurrence of the first cancer.

Keywords: Multiple testing, Functional Delta method, False Discovery Rate, DNA microarrays, DNA copy number, Normalization, Regulation networks.

Remerciements

Lorsque j'ai commencé à travailler à l'Institut Curie en septembre 2003, je me suis très vite passionné pour le développement de nouvelles méthodes bioinformatiques et leurs applications biologiques et cliniques. Au bout d'un an environ, j'ai décidé de commencer une thèse pour concilier mon intérêt pour ces questions avec ma curiosité pour les problèmes statistiques plus théoriques qu'elles soulèvent.

C'est grâce à la bienveillance et l'implication de mes directeurs de thèse, Stéphane Boucheron et Emmanuel Barillot, que j'ai pu trouver l'équilibre qui me convenait entre théorie et application. Merci à tous deux de la confiance et de la liberté que vous m'avez accordée. Stéphane, merci d'avoir su t'investir dans le domaine des tests multiples pour m'orienter vers des questions passionnantes, tout en m'aidant à étoffer mon bagage théorique. Emmanuel, merci de m'avoir permis de travailler sur des projets aussi variés, du développement de méthodes bioinformatiques et leur implémentation à la collaboration étroite avec les biologistes et cliniciens de l'Institut Curie. J'espère continuer longtemps à travailler avec vous.

Merci à Dominique Picard de m'avoir permis de concrétiser mon projet de thèse en m'orientant vers Stéphane pour l'encadrement de ma thèse. Merci à Laure Elie pour son soutien discret et efficace, et à Michèle Wasse pour sa gentillesse et son efficacité.

Je remercie Stéphane Robin et Terry Speed de m'avoir fait l'honneur de rapporter cette thèse. Merci également à Christophe Ambroise, Jean-Philippe Vert et Mark van de Wiel d'avoir accepté de faire partie du jury; cela témoigne de l'intérêt qu'ils portent à ce travail.

Merci à l'association "Courir pour la Vie, Courir pour Curie" et au programme ANR blanc TAMIS, qui ont financé ma thèse.

⋆ ⋆
⋆

Merci à tous ceux avec qui j'ai eu la chance de travailler dans l'équipe de bioinformatique, en particulier Isabel Brito, Sabrina Carpentier, Pierre Gestraud, Philippe Hupé, Stéphane Liva, Nicolas Servant et Éric Viara. J'ai beaucoup appris à votre contact, et travailler avec vous est un réel plaisir.

Je remercie également Olivier Delattre et les biologistes de l'unité INSERM 830 qui ont réussi à me faire comprendre un peu de biologie, en particulier Isabelle Janoueix-Lerosey et Sarah Fattet; je pense aussi à Gaëlle Pierron et Élodie Manié, grâce à qui j'ai pu réaliser moi-même (ou presque) une expérience de puce à ADN (voir la preuve en annexe E).

Merci à Mohamed Elati, François Radvanyi et Céline Rouveirol de m'avoir initié avec enthousiasme aux applications de l'apprentissage informatique à l'inférence de réseaux de régulation transcriptionnelle. Je remercie enfin Marc Bollet et Nicolas Servant, pour la collaboration efficace et très agréable que nous avons menée sur un projet passionnant.

⋆ ⋆
⋆

Merci à tous mes collègues de l'U900, grâce à qui ces cinq années à l'Institut Curie ont passé si rapidement et agréablement. Je pense en particulier à Stéphane, Sabrina, Franck, Laurence, Gautier et Fantine pour leurs contributions à la bonne ambiance — au travail ou ailleurs. Merci à mes compagnons de route statisticiens ou probabilistes de Chevaleret de m'avoir accueilli dans leur bureau: Mohamed, Karim, François, Julien, Marc, merci pour votre curiosité pour mon travail et votre gentillesse.

Je remercie également ceux de mes amis qui se sont orientés vers la recherche, et dont l'exemple m'a donné confiance en mon propre projet et m'a permis de le mener à bien: Vincent, Romu, Flora, Fred, Hugo, Ismaël, Greg, Cédric, Christelle, et Pierre-Yves, qui m'a initié aux joies des applications de la statistique à la génomique et de leur enseignement.

Je remercie mes parents de m'avoir fait confiance quand j'ai choisi de m'orienter vers ce monde inconnu qu'était la recherche; merci pour votre présence, votre soutien et votre conseil.

Merci à Dominique d'avoir souvent aménagé son emploi du temps de façon à soulager le mien. Mon petit Naël, merci pour "ton rire qui lézarde les murs, qui sait surtout guérir mes blessures"... merci pour ce que tu m'apprends chaque jour.

Agathe, merci pour ton écoute, ta compréhension et ton soutien dans les moments difficiles. Merci de me montrer ce qui est important dans la vie. Par-dessus tout, merci pour ton amour qui me donne des ailes.

“— Vous avez beau dire... y’a pas seulement que de la pomme, y’a aut’chose. Ça serait pas des fois de la betterave, hein ?
— Si, y’en a aussi.”

Les Tontons flingueurs, Michel Audiard, 1963.

Résumé en français

Cette thèse aborde des questions statistiques soulevées par l'analyse de données génomiques de grande dimension, dans le cadre de la recherche contre le cancer. Les principaux objectifs de la recherche en cancérologie sont d'ordre biologique, avec la compréhension des mécanismes de développement et de progression des cancers, et clinique, avec l'amélioration du diagnostic, du pronostic, et des traitements.

Les cancers résultent d'une accumulation de désordres génétiques, que les nouvelles techniques de la biologie moléculaire comme les puces à ADN permettent d'étudier quantitativement et à grande échelle. Une des questions statistiques soulevées par l'analyse des données de puces à ADN est le contrôle du risque de première espèce dans les tests d'hypothèses multiples.

Cette question peut être illustrée par la recherche de gènes significativement associés à un type de cancer: lorsqu'on teste simultanément un grand nombre de gènes candidats, il est utile de définir une mesure de risque qui tolère un certain nombre de faux positifs (c'est-à-dire de gènes déclarés significatifs alors qu'ils ne sont pas réellement associés), pourvu que cette proportion ne soit pas trop importante. C'est le sens du contrôle du False Discovery Rate (FDR), qui correspond à l'espérance de la False Discovery Proportion (FDP), proportion de faux positifs parmi les hypothèses rejetées.

Les développements sur les tests multiples que nous présentons peuvent s'appliquer dans un contexte plus général que celui des puces à ADN; réciproquement, les applications des statistiques aux données de puces à ADN que nous avons conduites incluent des questions de tests multiples, mais englobent des problématiques plus vastes. Nous avons donc choisi de présenter ce travail en deux parties.

1. Tests multiples

1.1. Mesures de risque de première espèce. On s'intéresse à des situations dans lesquelles m hypothèses sont testés simultanément. Une procédure de tests multiples (PTM) détermine quelles hypothèses doivent être rejetées, comme illustré dans le tableau 1 (tiré de [7]). Le nombre (inconnu) d'hypothèses nulles vraies est noté m_0.

Dans ce tableau, R est le nombre (aléatoire) de rejets effectués, et S, T, U, et V sont des variables aléatoires inobservables. Si toutes les hypothèses sont testées au même niveau α, R est une fonction croissante de α. Comme dans le cas d'un test d'hypothèse unique, le choix de ce niveau réalise un compromis entre le nombre V de rejets erronés (erreurs de première espèce)

	Non significatif	Significatif	Total
Hypothèses vraies	U	V	m_0
Hypothèses fausses	T	S	$m - m_0$
Total	$m - R$	R	m

Tableau 1: *Résultat d'une procédure de tests multiples.*

et le nombre T de non-rejets erronés (erreurs de seconde espèce); nous nous intéresserons aux mesures de risque qui assurent un contrôle du risque de première espèce.

Nous étudions les mesures de risque de première espèce les plus populaires: le Family-Wise Error Rate (FWER) ou taux d'erreurs par famille, et le False Discovery Rate (FDR) ou taux de faux positifs. Le FWER d'une procédure de tests multiples est défini comme la probabilité qu'au moins une hypothèse ait été rejetée à tort:

$$\mathsf{FWER} = \mathbb{P}(V > 0)$$

avec les notations du tableau 1. Les procédures de contrôle du FWER ont initialement été développées pour tester un petit nombre d'hypothèses, voire quelques dizaines, dans des situations où aucun faux positif ne saurait être toléré. Le FDR a été introduit par Benjamini et Hochberg [7] pour permettre un contrôle moins stringent, donc potentiellement plus adapté à des études exploratoires où un petit nombre de faux positifs peut être toléré. Le FDR est l'espérance du taux de faux positifs parmi les hypothèses rejetées; il s'écrit donc

$$\mathsf{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] ,$$

où

$$\mathsf{FDP} = \frac{V}{R \vee 1}$$

est la proportion de faux positifs parmi les hypothèses rejetées. Dans le même ordre d'idées, le pFDR (positive False Discovery Rate) est défini comme l'espérance conditionnelle du FDP sachant qu'au moins une hypothèse est rejetée:

$$\mathsf{pFDR} = \mathbb{E}\left[\mathsf{FDP} | R > 0\right] .$$

On parle de *contrôle faible* du risque de première espèce pour une procédure qui contrôle ce risque dans le cas où toutes les hypothèses nulles sont vraies, et de *contrôle fort* lorsque le risque est contrôlé quelle que soit la combinaison d'hypothèses nulles considérée. Dans cette thèse on s'intéresse uniquement au contrôle fort, qui est une propriété souhaitable dans les applications génomiques: des milliers d'hypothèses sont testées simultanément, et il est donc plausible qu'au moins une hypothèse nulle est fausse.

1.2. Modèle de mélange. Pour $i \in \{1 \dots m\}$, on note $Y_i = 0$ si l'hypothèse i est tirée selon l'hypothèse nulle $\mathcal{H}_0$, et $Y_i = 1$ sinon; X_i désigne la statistique de test associée. On suppose que les variables aléatoires $(X_i, Y_i)_{1 \leq i \leq m}$ sont indépendantes et identiquement distribuées: Y_i suit une loi de Bernoulli de paramètre ε_m, où ε_m est la proportion (inconnue) de vraies alternatives. La distribution conditionnelle de X_i sachant Y_i est notée

$F_1^{(m)}$ si $Y_i = 1$ et $F_0^{(m)}$ si $Y_i = 0$. La loi marginale des X_i est donc donnée par

$$F^{(m)} = (1 - \varepsilon_m)F_0^{(m)} + \varepsilon_m F_1^{(m)} .$$

On suppose que $F_0^{(m)}$ et $F_1^{(m)}$ sont C^1. Ce modèle de mélange peut également être formulé en termes de probabilités critiques plutôt que de statistiques de test. Puisque $F_0^{(m)}$ est continue, les probabilités critiques $(P_i)_{1\leq i\leq m}$, définies par

$$P_i = 1 - F_0^{(m)}(X_i) ,$$

sont uniformément distribuées sur $[0, 1]$ sous $\mathcal{H}_0$; on note $G_0^{(m)}(x) = x$ pour $0 \leq x \leq 1$. La loi marginale des probabilités critiques a alors pour fonction de répartition $G^{(m)} = (1 - \varepsilon_m)G_0^{(m)} + \varepsilon_m G_1^{(m)}$ et pour densité $g^{(m)} = (1-\varepsilon_m) + \varepsilon_m g_1^{(m)}$, où $G_1^{(m)}$ et $g_1^{(m)}$ désignent respectivement la fonction de répartition et la densité des probabilités critiques sous l'alternative $\mathcal{H}_1$. Enfin, on note $(P_{(i)})_{1\leq i\leq m}$ le vecteur des probabilités critiques ordonnées associé à $(P_i)_{1\leq i\leq m}$.

On considère ce modèle de mélange dans deux cadres distincts. Dans le cas *creux*, on fait converger la proportion ε_m vers 0 et la distance entre $\mathcal{H}_0$ et $\mathcal{H}_1$ vers l'infini quand le nombre m d'hypothèses testées tend vers l'infini. Dans le cas *non creux*, tous les paramètres du modèle de mélange sont fixés; on note alors $\pi_0 = 1-\varepsilon_m$ la proportion d'hypothèses nulles vraies.

La procédure de Benjamini et Hochberg. La procédure BH95 [7] rejette les hypothèses dont les probabilités critiques sont inférieures à $\widehat{\tau} = \alpha\widehat{I}_m/m$, où

$$\widehat{I}_m = \max\left\{i \in \{1, \dots m\}, P_{(i)} \leq \alpha i/m\right\} .$$

Cette définition peut se réécrire de la façon suivante: si $\hat{\mathbb{G}}_m$ est la fonction de répartition empirique des probabilités critiques, alors

$$\widehat{\tau} = \sup\left\{u \in [0, 1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\} .$$

La figure 1 illustre ces deux formulations équivalentes du seuil de la procédure BH95. L'application $u \mapsto u/\alpha$ est appelée *courbe de rejet* de la procédure BH95 (ou *droite de Simes* [83]). Lorsque les probabilités critiques

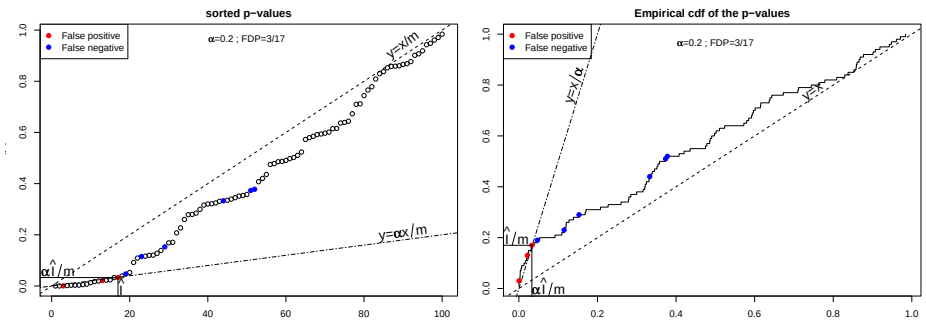


FIGURE 1. *Interprétations duales du seuil de décision de la procédure* BH95.

tirées sous l'hypothèse nulle sont indépendantes ou satisfont certaines conditions de dépendance positive, cette procédure conduit à un taux de fausses découvertes exactement égal à $\pi_0\alpha$ [8, 28, 79, 93], et ce quelle que soit la distribution jointe des probabilités critiques sous l'alternative.

Lorsque $\pi_0 < 1$, la procédure BH95 est donc conservatrice d'un facteur π_0. De nombreuses méthodes permettant d'estimer π_0 implicitement ou explicitement ont été proposées, dans l'espoir de fournir un contrôle du FDR plus exact que celui de la procédure BH95 lorsque les hypothèses testées sont indépendances: ces procédures sont dites *adaptatives*. Les procédures adaptatives à une étape utilisent des courbes de rejet autres que la droite de Simes, sans incorporer explicitement un estimateur de π_0; les procédures adaptatives à deux étapes appliquent la procédure BH95 au niveau $\alpha/\widehat{\pi_0}$, où $\widehat{\pi_0}$ est un estimateur de π_0.

Chi [16] a récemment mis en évidence l'existence d'une *valeur critique* $\alpha^\star$ qui dépend uniquement de la fonction de répartition G des probabilités critiques, telle que si $\alpha < \alpha^\star$, le nombre d'hypothèses rejetées par la procédure BH95 est borné en probabilité lorsque le nombre d'hypothèses testées augmente, alors que si $\alpha > \alpha^\star$, la proportion d'hypothèses rejetées converge en probabilité vers une valeur strictement positive.

1.3. Chapitre 2: propriétés asymptotiques du FDP. La proportion de rejets erronés (FDP) étant une grandeur aléatoire, il est utile d'étudier ses fluctuations autour de son espérance, le FDR. Le comportement asymptotique du processus $(\mathsf{FDP}_m(t))_{0<t\leq 1}$, où t est un seuil déterministe, a déjà été étudié [35, 89, 93]. Nous nous intéressons aux propriétés du *seuil aléatoire* $\widehat{\tau}$ d'une procédure de tests multiples donnée, et en particulier à la loi asymptotique de $\mathsf{FDP}_m(\widehat{\tau})$, c'est-à-dire du FDP effectivement atteint par cette procédure.

Nous considérons un vaste ensemble de procédures, dont le seuil $\widehat{\tau}$ peut s'écrire comme une fonctionnelle $\mathcal{T}$ (que nous appellerons fonction de seuil) de la fonction de répartition empirique $\hat{\mathbb{G}}_m$ des probabilités critiques. C'est notamment le cas de la procédure BH95 procédure, dont le seuil au niveau α est donné par

$$\widehat{\tau} = \sup\left\{u \in [0,1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\}.$$

Plus généralement, un grand nombre de procédures adaptatives à une ou deux étapes peuvent être décrites grâce à ce formalisme. Le seuil $\widehat{\tau}$ de ces procédures et le FDP associé s'écrivent donc comme des *fonctionnelles des trajectoires d'un processus aléatoire*: le processus empirique. Ce formalisme permet donc de découpler l'étude de la régularité de la fonctionnelle, qui ne dépend que de la procédure de tests multiples, de la régularité du processus empirique, qui ne dépend que de la distribution des probabilités critiques.

Nous avons prouvé que lorsque le nombre d'hypothèses testées tend vers l'infini, le FDP d'une procédure de tests multiples ayant pour fonction de seuil $\mathcal{T}$ converge en loi à vitesse $1/\sqrt{m}$ vers un niveau de FDR explicite et qui dépend de la procédure, dès que $\mathcal{T}$ est différentiable au sens de Hadamard

au point G, tangentiellement à l'ensemble $C[0,1]$ des fonctions continues de $D[0,1]$. Une condition suffisante pour que cette hypothèse de différentiabilité soit vérifiée est l'existence d'un unique point de croisement intérieur entre la fonction de répartition des probabilités critiques et la courbe de rejet de la procédure. L'existence s'interprète comme une généralisation naturelle de la notion de *valeur critique* introduite par Chi [16].

Ce résultat permet de caractériser le comportement asymptotique ainsi que les conditions de régularité associées pour différentes procédures, dont la procédure BH95 et plusieurs procédures adaptatives à une ou deux étapes. Comme toutes les procédures convergent à la même vitesse $(1/\sqrt{m})$, leur puissance asymptotique peut être comparée explicitement via le FDR qu'elles atteignent asymptotiquement.

Nous avons également, grâce au formalisme introduit, mis en évidence des connexions intéressantes entre les procédures adaptatives: la procédure BR08 [10] s'interprète comme un point fixe de la procédure BKY06 [9], alors que la procédure FDR08 [30] s'interprète comme un point fixe de la procédure Sto02 [93].

1.4. Chapitre 3: bornes intrinsèques et contrôle du FDR. Les questions statistiques qui se posent lorsqu'on teste un grand nombre d'hypothèses incluent non seulement la sélection des hypothèses à rejeter, que nous étudions au chapitre 2, mais également un problème de détection et un problème d'estimation:

> **Détection:** Y a-t-il des hypothèses nulles fausses ?
> **Estimation:** Combien ?
> **Sélection:** Lesquelles ?

Le problème de détection est le test de l'hypothèse nulle que la proportion d'hypothèses nulles fausses est 0 contre l'alternative qu'elle est strictement positive, tandis que le problème d'estimation consiste à estimer cette proportion. Bien qu'il s'agisse du test d'une seule hypothèse et de l'estimation d'une seule quantité, le contexte de comparaison multiples dans lequel ces problèmes sont posés motive le recours à des procédures de test et d'estimation dédiées.

Des travaux récents mentionnent l'existence de bornes intrinsèques pour ces trois problèmes. Pour le problème de sélection, le phénomène de *valeur critique* mentionné ci-dessus [16] illustre l'existence d'une borne inférieure qui peut être strictement positive, en deçà de laquelle aucune procédure de tests multiples ne peut contrôler le pFDR. Pour le problème de détection, Donoho et Jin [24] ont identifié une *frontière de détection* qui caractérise les situations dans lesquelles le test du rapport de vraisemblance détecte correctement avec probabilité 1. De façon similaire, une *frontière d'estimation* pour les modèles de mélanges Gaussiens creux caractérise les situations dans lesquelles la proportion d'hypothèses nulles vraies peut être estimées de façon consistente.

Le chapitre 3 est motivé par la comparaison, dans le contexte du contrôle du FDR, de ces bornes intrinsèques dans le cadre creux et non creux, c'est-à-dire selon que la proportion ε de vraies alternatives tend vers 0 ou non.

Dans le cadre non creux, nous avons prouvé que le phénomène de valeur critique n'intervient que lorsque la loi des probabilités critiques a des queues épaisses (par exemple pour la loi de Laplace (exponentielle bilatère) ou la loi de Student. Nous avons ensuite prouvé que dans les problèmes de localisation symétriques, pour lesquels la densité de la statistique de test sous l'alternative est une translation positive celle sous l'hypothèse nulle, le phénomène de valeur critique intervient si et seulement si $\pi_0 = 1 - \varepsilon$ n'est pas identifiable dans le problème d'estimation.

Nous avons également établi un lien entre les vitesses de convergence atteignables pour le problème d'estimation dans le cadre non creux et la régularité au voisinage de 1 de la fonction de répartition des probabilités critiques G. La faible régularité de G au voisinage de 1 attendue dans les applications usuelles implique que les vitesses de convergence des procédures de tests multiples adaptatives à deux étapes sont faibles.

Enfin, nous avons proposé une interprétation de la frontière de décision de la procédure BH95 dans le cas creux en termes de contrôle du pFDR.

2. Analyse statistique de données de puces à ADN

2.1. Utilisation des puces à ADN en cancérologie. Les cancers sont un ensemble de maladies au cours desquelles des cellules anormales prolifèrent sans contrôle, échappent à la mort cellulaire programmée, et deviennent capables d'envahir d'autres tissus. La transformation d'une cellule normale en une cellule tumorale passe par l'altération du fonctionnement de gènes qui régulent la croissance et la différenciation des cellules. Ces altérations du génome peuvent avoir lieu à différents niveaux: gains ou pertes de chromosomes entiers, mutations affectant une seule lettre de la séquence d'ADN, et peuvent avoir pour conséquence directe ou indirecte des modifications de l'expression des gènes. La nécessité de comprendre et caractériser ces altérations a stimulé le développement de nouveaux outils de biologie moléculaire tels que les puces à ADN, qui permettent notamment de mesurer simultanément le nombre de fragments d'ARN (niveau d'expression) ou d'ADN en un grand nombre de loci du génome de l'échantillon.

Les puces à ADN constituent ainsi une technologie de choix, à la fois pour la recherche biologique en cancérologie, qui a pour objectif la compréhension des mécanismes de développement et de progression des cancers, et la recherche clinique, qui a pour objectif l'amélioration du diagnostic, du pronostic, et des traitements. L'avènement de ces nouvelles technologies a nécessité le développement de méthodes statistiques adaptées à des données d'aussi grande dimension, ainsi qu'à chaque question biologique ou clinique particulière.

2.2. Méthodes statistiques pour l'analyse des données de puces à ADN. On distingue les analyses dites *de bas niveau* des analyses dites *de haut niveau*. Les premières sont nécessaires à l'étude des données mais ne permettent pas directement de répondre aux questions biologiques et cliniques d'intérêt: elles comprennent la planification expérimentale, l'analyse d'image pour exploiter les données brutes en sortie du scanner, et la normalisation, qui a pour objectif d'éliminer des données les variations sans

lien avec le signal biologique d'intérêt, et de rendre comparable entre eux les résultats quantitatifs de plusieurs expériences.

Les analyses dites *de haut niveau* doivent permettre de répondre aux questions biologiques et cliniques, qui mettent en jeu des problématiques classiques en statistique: analyses exploratoires (analyses factorielles, classification non supervisée), tests d'hypothèses, classification et régression.

Cependant les cellules sont des systèmes complexes, dont le comportement ne saurait être expliqué grâce à l'étude d'un seul niveau d'information biologique, ou en n'utilisant qu'un type de technologie à la fois. Ceci justifie les *approches interactives*, qui visent à exploiter simultanément plusieurs sources d'informations, que l'on pense complémentaires. Le développement de ce type d'approche va de pair avec l'émergence d'une nouvelle discipline, la *biologie des systèmes*, qui s'intéresse aux interactions entre les différents niveaux d'information génétique.

Dans les sections qui suivent nous donnons un aperçu des contributions de cette thèse aux méthodes d'analyse de données de puces à ADN en cancérologie: les sections 2.3 et 2.4 présentent deux méthodes génériques qui sont maintenant largement utilisées, notamment à l'Institut Curie. Les sections 2.5 et 2.6 présentent deux méthodes respectivement conçues pour répondre à une question biologique et à une question clinique.

2.3. Chapitre 5: normalisation de données de puces à ADN. Ce travail a été initié à la suite de l'analyse d'échantillons tumoraux de deux plates-formes de puces à hybridation génomique comparative (CGH), mesurant le nombre de copies d'ADN: Université de Californie, San Francisco (UCSF) [85], et Institut Curie. Nous avons montré que sur ces deux plates-formes, des biais spatiaux constituaient la principale source de variabilité non attribuable à un signal biologique. Nous avons identifié deux types de biais spatiaux: des régions entières de la puce présentant un niveau de signal moyen bien plus élevé que le reste de la puce, et des effets de gradient de signal d'une extrémité à l'autre de la puce.

Ces deux types d'effets n'étant pas corrigés de façon satisfaisante par les techniques existantes, nous avons développé une méthode de segmentation spatiale [66], qui comporte trois étapes:

(i) estimation d'une tendance spatiale sur la puce par une méthode de régression robuste (LOESS [20, 21]);
(ii) segmentation de la puce en régions dont la tendance spatiale est similaire à l'aide de la méthode NEM, une méthode de classification non supervisée qui inclut une contrainte spatiale [4, 5];
(iii) identification des régions effectivement affectées par un biais spatial localisé.

Cette méthode est très utile pour les applications car elle permet de distinguer des artefacts expérimentaux les gènes potentiellement impliqués dans la progression tumorale. Nous avons développé MANOR, un logiciel (paquet) en langage R destiné à combler l'absence de logiciels dédiés à la normalisation des puces CGH. MANOR intègre en particulier la méthode

de normalisation spatiale que nous avons développée. Ce paquet R a été intégré à l'environnement Bioconductor, un projet libre de développement de logiciels dédiés à l'analyse et l'interprétation de données génomiques [38]. Une description des fonctionnalités de MANOR et des exemples d'application sont livrés dans la vignette donnée en Annexe A.

Intégration aux plates-formes d'analyse. Les biologistes de l'unité INSERM 830 ont développé leur propre plate-forme d'expérience de puces CGH; une plate-forme d'analyse dédiée, appelée CAP pour "CGH-array Analysis Pipeline", a été implémentée par l'équipe Bioinformatique afin de stocker, analyser, et visualiser les données ainsi produites. J'ai participé à l'intégration de MANOR dans CAP. MANOR a été utilisé pour analyser plus de 6000 puces CGH via CAP, par 132 utilisateurs travaillant dans le cadre de 94 projets de recherche (données: juin 2008).

Nous avons également décidé d'implémenter CAPweb, une version de CAP qui peut être utilisée directement depuis notre site internet: `http://bioinfo.curie.fr/CAPweb`, ou installée localement pour une utilisation interne dans un centre de recherche particulier. J'ai participé à l'intégration de MANOR à CAPweb [59]. CAPweb a été utilisé pour analyser plus de 5000 puces CGH depuis notre site internet, par 21 utilisateurs dans le cadre de 468 projets de recherche. CAPweb a été installé dans 10 laboratoires de recherche publics, et une entreprise privée. Plusieurs publications rapportent déjà des résultats ayant été obtenus à l'aide de CAPweb [31, 45, 46, 103, 108].

2.4. Corrélation entre nombre de copies d'ADN et niveaux d'expression des gènes. Ce travail a été effectué en collaboration avec Pierre Gestraud.

Des études récentes ont cherché à caractériser l'effet dosage génique, c'est-à-dire l'influence globale du nombre de copies d'ADN sur le niveau d'expression du gène correspondant, à l'aide de mesures parallèles — issues d'expériences de puces à ADN — du nombre de copies d'ADN ("génome") et du niveau d'expression ("transcriptome") sur les mêmes échantillons [19].

Nous avons développé GTCA (pour Genome Transcriptome Correlation Analysis), un logiciel R qui permet de quantifier cet effet dosage génique à partir de différents types de données de puces à ADN. L'implémentation permet d'effectuer les étapes suivantes:

Pré-traitement: appariement non ambigu entre les sondes génome et les sondes transcriptome, à partir de leur position sur le génome. Les valeurs manquantes dans les données génome sont interpolées en utilisant la continuité du nombre de copies d'ADN le long du génome;

Analyse statistique: pour chaque couple formé, on calcule un coefficient de corrélation entre le nombre de copies d'ADN et le niveau d'expression, ainsi qu'une probabilité critique associée. Une étape de correction de tests multiples est ensuite effectuée, qui assure un contrôle du FWER [43] ou du FDR [7];

Visualisation et interprétation: les coefficients de corrélation et les significativités associées (après correction de tests multiples)

sont représentées le long du génome, et peuvent être exportées pour une utilisation dans des logiciels permettant l'interprétation biologique des résultats, tels que GSEA [95].

Un poster décrivant cet outil a été présenté à la conférence ISMB en 2007 (Annexe B). GTCA est intégré à VAMP [54], un logiciel de visualisation et d'analyse développé par l'équipe de bioinformatique de l'Institut Curie (Annexe D). GTCA sera bientôt soumis à Bioconductor en tant que paquet R. Il peut d'ores et déjà être utilisé via VAMP sur des jeux de données publiques contenus dans ACTuDB [44], une base de données dédiée aux données publiques de nombre de copies d'ADN: http://bioinfo.curie.fr/actudb.

2.5. Chapitre 6: apprentissage de réseaux de régulation transcriptionnelle. Ce travail a été effectué en collaboration avec Mohamed Elati et Céline Rouveirol [27].

Les facteurs de transcription sont des protéines qui activent ou répriment l'expression de leurs gènes cibles en se fixant sur des séquences d'ADN spécifiques situées en amont de la partie codante de ces gènes. La reconstruction de réseaux d'interaction transcriptionnels représente un défi important pour la compréhension du fonctionnement des cellules, et peut aussi être utile pour découvrir de nouvelles cibles thérapeutiques.

Plusieurs approches locales ont été proposées pour ce problème, qui infèrent un ensemble de régulateurs candidats pour chaque gène d'intérêt, à partir d'une mesure de corrélation ou d'information mutuelle entre le gène régulé et ses régulateurs potentiels; pour une fonction de score donnée, la recherche exhaustive des meilleurs candidats a une complexité exponentielle en le nombre de candidats, et ne peut donc être effectuée sur des données d'expression compte tenu de leur grande dimension.

Nous avons mis en place une méthode appelée LICORN pour LearnIng Cooperative Regulation Networks, qui est décrite en détail au chapitre 6. Elle met à profit le fait que plusieurs facteurs de transcription peuvent être impliqués dans la régulation du même gène pour associer à chaque gène cible un réseau de régulation génique (GRN pour *Gene Regulatory Network*), qui est un couple d'ensembles de facteurs de transcription: un *ensemble d'activateurs* et un *ensemble d'inhibiteurs*. Cette méthode fonctionne comme suit:

(i) utilisation d'une méthode d'extraction de motifs fréquents pour identifier des ensembles de co-activateurs ou de co-inhibiteurs à partir de données d'expression discrétisées;
(ii) construction et représentation structurée d'un ensemble de co-activateurs et co-inhibiteurs candidats pour chaque gène, au sein duquel une recherche exhaustive peut être faite de manière efficace;
(iii) définition d'un score permettant d'associer un *meilleur GRN* à chaque gène parmi tous les couples possibles de co-activateurs et co-inhibiteurs candidats, et sélection des gènes dont le score est significatif à l'aide d'une méthode de correction de tests multiples appropriée;
(iv) estimation des *performances de prédiction* des GRN sélectionnés à l'aide d'une méthode de validation croisée.

J'ai participé aux deux dernières étapes. Comme nous travaillions avec des données d'expression discrétisées, nous avons choisi les moindres écarts absolus (MAE) comme mesure de distance entre profils d'expression, à la fois pour le score dans l'étape (iii) et pour l'erreur de prédiction dans l'étape (iv). La significativité du *meilleur GRN* à l'étape (iii) a été calculée en comparant son score au meilleur score obtenu en permutant aléatoirement les données de la matrice d'expression originale, et nous avons utilisé comme procédure de tests multiples l'approche conservatrice de Benjamini et Yekutieli [8], afin de garantir un contrôle fort du FDR bien que les hypothèses testées n'étaient pas indépendantes. La performance de prédiction (iv) a été estimée à l'aide d'une validation croisée en 10 paquets.

LICORN est implémenté en CaML et distribué librement: `http://www.lri.fr/~elati/licorn.html`. Sur les deux jeux de données d'expression de levure de référence que nous avons testés [32, 88], LICORN obtient des erreurs de prédiction significativement plus faibles que Minreg, la méthode de référence pour l'inférence non supervisée de réseaux de régulation transcriptionnels [67]. Les résultats obtenus permettent de retrouver des interactions connues entre facteurs de transcription et gènes cibles, et suggèrent de nouveaux groupes de co-régulateurs candidats.

2.6. Chapitre 7: distinction des vraies récidives parmi des seconds cancers du sein homolatéraux. Ce travail a été effectué en collaboration avec Marc Bollet et Nicolas Servant [11].

Le traitement des cancers du sein par chirurgie conservatrice (sans ablation du sein) est plus facilement accepté que l'ablation. Bien que ces deux traitements sont équivalents en termes de survie globale [99], les patientes traitées par chirurgie conservatrice courent le risque de développer une seconde tumeur sur le même sein. Dans ce cas il est fondamental de déterminer si cette seconde tumeur est une nouvelle tumeur primaire (NP) ou une vraie récidive du premier (VR): dans le premier cas, le même traitement peut être appliqué alors que dans le second, un traitement plus agressif est nécessaire, puisque la première tumeur n'a pas été guérie.

La principale difficulté est l'absence de définition objective de "nouvelle tumeur primaire" et "vraie récidive"; la définition clinique standard repose sur différentes caractéristiques histopathologiques: localisation, type histologique, grade, récepteurs hormonaux. Plusieurs études récentes suggèrent que l'utilisation de données génomiques peut permettre d'améliorer cette définition; en particulier, les altérations de nombre de copies d'ADN peuvent être utilisées comme marqueurs du lien clonal entre les deux tumeurs. Ces études utilisent la proximité de la tumeur primaire et de la seconde tumeur sur le dendrogramme issu d'une classification ascendante hiérarchique pour inférer le statut (NP/VR) de cette dernière [98].

Nous avons cherché à proposer une définition plus pertinente de NP/VR à l'aide de données de nombres de copies d'ADN, en exploitant les deux idées suivantes:

idée biologique: il est possible que deux tumeurs sans lien clonal aient des altérations génomiques en commun, simplement du fait que ces altérations constituent un point de passage obligé dans le processus de progression tumorale. En revanche, le fait que

les tumeurs aient les mêmes points de cassure, c'est-à-dire que les régions altérées commencent ou finissent au même endroit sur le génome devrait être un indicateur plus spécifique de leur lien clonal;

idée statistique: utiliser les résultats d'une classification ascendante hiérarchique pour séparer NP et VR paraît arbitraire et peu robuste puisque la distinction NP/VR pour un couple de tumeurs donné peut être remis en cause par l'ajout ou la suppression d'un autre cas dans l'analyse. Le fait de travailler avec un *score* plutôt qu'un dendrogramme semble plus adapté, et permet en outre de fixer une tolérance en termes de faux positifs ou négatifs.

Nous avons donc construit un *score d'identité partielle* reposant sur le nombre de points de cassure communs entre les deux tumeurs, en pondérant chaque point de cassure en fonction de sa fréquence sur un jeu indépendant de cancers du sein. On estime la distribution du score sous l'hypothèse nulle d'absence d'identité partielle entre les deux tumeurs à l'aide de *paires artificielles*, construites en appariant chaque tumeur primaire à l'une des autres secondes tumeurs.

La qualité du score dépend beaucoup de la précision de localisation des points de cassure: nous avons utilisé l'algorithme ITALICS pour localiser les points de cassure, dont il a été prouvé qu'il fait mieux que ses concurrents en termes de sensibilité de détection des points de cassure, et de précision de leur localisation [74].

J'ai contribué à la construction du score et à l'élaboration de la méthode d'estimation de la distribution du score sous l'hypothèse nulle. Bien qu'il soit difficile d'évaluer la performance de ce score puisque la vraie classification TP/VR est inconnue, le score ainsi construit est plus performant que les définitions reposant sur les caractéristiques cliniques pour le pronostic, c'est-à-dire la prédiction de la survie sans métastases.

Ce score est utilisé dans une nouvelle étude biologique qui vise à identifier des gènes dont le nombre de copies d'ADN diffère entre les tumeurs primaires pour lesquelles la seconde tumeur est une nouvelle tumeur primaire, et celles pour lesquelles il s'agit d'une vraie récidive.

Contents

General introduction

This thesis is motivated by **statistical questions** raised by the analysis of high-dimensional **genomic data** for **cancer research**. This research field includes ambitious biological and medical aims as fundamental as gaining insight into biological mechanisms of cancer development and progression, and as practical as improving cancer diagnosis, prognosis, and treatment.

As cancers have been shown to result from an accumulation of genetic disorders, these biological and medical questions triggered the use of new experimental techniques, including DNA microarrays, that allow high-throughput molecular characterization at several informational levels, including DNA, RNA, and protein. The advent of DNA microarrays raised a number of statistical questions of interest, including multiple hypothesis testing.

This thesis has been done in collaboration between the Laboratoire de Probabilités et Modèles Aléatoires (LPMA, Paris VI and VII Universities and CNRS) and the Bioinformatics group of Institut Curie (now INSERM U900/Institut Curie/Mines ParisTech). Due to its situation at the intersection between the Research Center and the Hospital of Institut Curie, the Bioinformatics group has a pivotal role in **translational research**, that is, in bridging the gap between fundamental research in biology, physics and chemistry on the one hand, and applied medicine on the other hand.

A typical issue in cancer research is to find genes that are significantly associated with a cancer type. When a large number of genes are tested simultaneously, false positives (genes declared significant whereas they are not associated) may be tolerated provided that their proportion among significant genes is not too large. In such situations, it is thus particularly adapted to control the False Discovery Rate (FDR), the expected proportion of false positives among a set of rejected hypotheses.

Although this work was initiated by multiple hypothesis testing questions that arise from the analysis of DNA microarrays in the context of cancer research, multiple testing issues studied in this thesis have their own interest, independently of the application to DNA microarrays. Conversely, the applications of statistics to DNA microarray analysis we present here include, but are not restricted to, multiple hypothesis testing problems. This thesis therefore consists of two parts.

In the first part, we study statistical issues raised by multiple hypothesis testing problems, with a focus on FDR. After an introduction to multiple testing (chapter 1), we investigate the asymptotic performance of a family of FDR controlling procedures (chapter 2), and study intrinsic bounds to three

multiple testing problems (detection, estimation, and selection) through the performance of FDR controlling procedures in these contexts (chapter 3).

The second part of this thesis is dedicated to applications of statistics, and especially multiple testing procedures, to microarray data analysis. Chapter 4 motivates the use of high throughput techniques such as DNA microarrays in the context of cancer research. In chapter 5, we introduce a method for low-level analysis of a specific type of microarrays, with the aim of separating the true biological signal of interest from experimental artifacts, especially spatial biases. In chapter 6, we propose an unsupervised method to infer regulatory networks from expression data. Finally, in chapter 7 we define a genomic score that permits testing the hypothesis that a second cancer is a true recurrence from a first cancer, against the alternative that it may be considered as a new primary tumor.

Publications and documents

P. Neuvial, P. Hupé, I. Brito, S. Liva, E. Manié, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7(1):264, May 2006.

S. Liva, P. Hupé, **P. Neuvial**, I. Brito, E. Viara, P. La Rosa, and E. Barillot. CAPweb: a bioinformatics CGH array Analysis Platform. *Nucleic Acids Res*, 34(Web Server issue):477–481, Jul 2006.

P. La Rosa, E. Viara, P. Hupé, G. Pierron, S. Liva, **P. Neuvial**, I. Brito, S. Lair, N. Servant, N. Robine, E. Manié, C. Brennetot, I. Janoueix-Lerosey, V. Raynal, N. Gruel, C. Rouveirol, N. Stransky, M.-H. Stern, O. Delattre, A. Aurias, F. Radvanyi, and E. Barillot. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*, 22(17):2066–2073, Sep 2006.

M. Elati, **P. Neuvial**, M. Bolotin-Fukuhara, E. Barillot, F. Radvanyi, and C. Rouveirol. LICORN: LearIng COoperative Regulation Networks. *Bioinformatics*, 23(18):2407–2414, 2007.

M. A. Bollet, N. Servant, **P. Neuvial**, C. Decraene, I. Lebigot, J.-P. Meyniel, Y. De Rycke, A. Savignoni, G. Rigaill, P. Hupé, A. Fourquet, B. Sigal-Zafrani, E. Barillot, and J.-P. Thiery. High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers. *J Natl Cancer Inst*, 100(1):48–58, 2008.

P. Neuvial. Asymptotic properties of false discovery rate controlling procedures under independence. In revision for *Electronic Journal of Statistics*.

P. Neuvial. Intrinsic bounds and false discovery rate control in multiple testing problems. Submitted.

Part 1

Multiple testing

CHAPTER 1

Large-scale multiple testing

EXPRESSIONISM

Mary Brinig, *B19n031*, 2003

Wassily Kandinsky, *Squares with Concentric Circles*, 1913

Contents

This chapter gives an introduction to large-scale multiple testing. A short historical perspective (section 1.1) motivates the need of dedicated error rates, which are defined in section 1.2. Section 1.3 introduces FDR controlling procedures and gives an overview of their properties. In section 1.4 we recall technical tools that will be used in subsequent chapters, and section 1.5 gives an overview of the contribution of the thesis to FDR control in large-scale multiple testing.

1.1. Multiple testing situations: historical perspective

Multiple testing refers to the testing of more than one hypothesis at a time; it is a sub-field of multiple inference, which also covers multiple estimation. Although most recent multiple testing literature is concerned with *large-scale* multiple testing, that is, the simultaneous testing of thousands or more hypotheses, *small-* and *middle-scale* multiple testing have been an active field of statistics since the second half of the twentieth century.

This section gives a short historical perspective of multiple testing problems, inspired by a comprehensive review of the field [82]. Even though this thesis focuses on statistical issues raised by large-scale multiple testing problems, crucial questions such as adaptivity to the number of true null hypotheses or robustness to dependencies between hypotheses had already been mentioned and studied for small- or medium-scale multiple testing problems.

1.1.1. Small- and medium-scale multiple testing. Multiple testing questions have been mentioned as early as 1843 [22]. Taking the example of testing whether the probability of a male birth is influenced by birth order, age, profession, wealth or religion of the parents, the French mathematician Antoine-Augustin Cournot noticed that such repeated "cuts" of a reference population into two groups increases the risk that one observed difference is called significant by pure chance[1]:

> The probability that an observed deviation can not be attributed to the vagaries of chance takes on very different values depending on whether one has tried a more or less large number of splits before having hit on the observed deviation.

To illustrate this point, Table 1 displays the probability that at least one of m independent tests performed at level α is called significant, which is given by $1-(1-\alpha)^m$:

m	1	2	5	10	20	50	100
$\alpha = 0.05$	0.05	0.10	0.23	0.40	0.64	0.92	0.99
$\alpha = 0.01$	0.01	0.02	0.05	0.10	0.18	0.39	0.63

TABLE 1. *Probability* $1-(1-\alpha)^m$ *of (at least) one false rejection among* m *hypotheses tested at level* α.

[1] translated by Shaffer [82] from the following: "La probabilité qu'un écart de grandeur donnée n'est pas imputable aux anomalies du hasard, prendra des valeurs très différentes selon qu'il aura essayé un plus ou moins grand nombre de coupes avant de tomber sur l'écart observé."

Multiple testing situations with a moderate number of hypotheses arise in a number of situations, including biology and medicine: toxicity studies on animals typically have *multiple outcomes*, for example when testing carcinogenicity of a potential drug, several cancer sites are simultaneously monitored; in clinical trials, *interim tests* are often performed at several stages of the trial. In both cases, multiplicity should be taken into account.

1.1.2. Large-scale multiple testing. Since a dozen years, several domains of applied statistics have been challenged with the analysis of large data sets, generally with many more variables than observations. Any field involving high-dimensional data is concerned with multiple testing questions; here we give three specific examples in genomics, medical imaging, and astronomy, for which many tests are typically performed simultaneously.

DNA microarrays permit measuring the expression level of dozens of thousands of genes within a single biological experiment. A typical question of interest to biologists and clinicians is to find those genes whose expression differ significantly between two groups of samples (for example, two cancer subtypes). In neuroimaging, new experimental techniques such as functional Magnetic Resonance Imaging (fMRI) or Electro- or Magneto-Encephalography (EEG, MEG) permit inferring four-dimensional pictures of the brain's activity: each data point measures the activity of a voxel (volumetric pixel) integrated over a short time period, and a primary objective is to delineate areas of *significant* brain activity [37]. Finally, the field of source detection in astronomy is concerned by discoveries of stars and galaxies from observations of the cosmic microwave background (CMB) [62].

In these motivating examples, thousands to millions of hypotheses may be tested simultaneously; the need of definition of error rates that take multiplicity into account is thus even more crucial than for small- or middle-scale multiple testing.

1.2. From single testing to multiple testing

In this section we recall notions from classical testing theory, and cast them into a multiple testing framework.

1.2.1. Testing a single hypothesis. Suppose we wish to test whether a specific gene is overexpressed in breast cancers. We assume that gene expression levels have been measured in a series of healthy breast samples (group 1), and in a series of breast tumor samples (group 2). A testing procedure uses these observations to declare whether the expression level is significantly larger in group 2 than in group 1. Formally, we are interested in testing the *null hypothesis* that expression levels are the same in both groups, against the *alternative hypothesis* that they are larger in group 2 than in group 1. Assuming that gene expression levels are Gaussian distributed within each group, with the same variance, one will typically reject the null hypothesis if the observed absolute difference between mean expression levels in group 1 and 2 is large enough.

As the null hypothesis may be true or false, and the procedure may accept or reject it, such a procedure has four possible outcomes, that include

two different types of error. A type I error or *false positive* is made when the null hypothesis is rejected whereas it is true, that is, when the gene is declared overexpressed in breast cancers whereas it is not; conversely, a type II error or *false negative* is made when the null hypothesis is not rejected whereas it is false, that is, when the gene is not declared overexpressed in breast cancers whereas it is. The *level* of the test is defined as the type I error rate, and the *power* of the test is defined as one minus the type II error rate.

An ideal test should have small type I and type II errors; however, both of these risks cannot be minimized at the same time for a given set of observations: the smaller type I error rate, the larger type II error rate, and *vice versa*. One of the goals of testing theory is to motivate the choice of the threshold above which the null hypothesis will be rejected.

Neyman and Pearson have developed an optimality theory for tests of a simple null hypothesis against a simple alternative. They proved that for each target level α, there exists a test with level α exactly that has maximum power among all tests with level α, and give an explicit formulation for this test. Such a test fundamentally treats type I and type II error asymmetrically, because it maximizes power (that is, it minimizes type II error) for a given level (or type I error). Neymann and Pearson showed that this test is still optimal for testing a simple null hypothesis against a one-sided alternative (as in the above example of breast cancers and a single gene's expression) if the likelihood ratios are monotone.

1.2.2. Multiple testing: definitions and error rates.

Testing several hypotheses. We are interested in situations in which m hypotheses are simultaneously tested, and denote by m_0 the number of true nulls. A Multiple Testing Procedure (MTP) decides which of these hypotheses should be rejected. The outcome of a MTP is described by Table 2, which is taken from Benjamini and Hochberg [7]. In this table m_0

	Non significant	Significant	Total
True hypotheses	U	V	m_0
False hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

TABLE 2. *Outcome of a multiple testing procedure.*

is unknown, R is the observed (random) number of rejections, and S, T, U, and V are unobservable random variables. If each hypothesis is tested at individual level α, then R is a non-decreasing function of α. As for for single hypothesis testing, the choice of α balances the number V of false rejections with the number T of false non rejections, and we focus on risk measures that provide a control of type I error.

Error rates. As we are testing more than one hypothesis, several type I error rates may be defined. We focus on the most widely used error rates: Family-Wise Error Rate (FWER), and False Discovery Rate (FDR). The FWER of a Multiple Testing Procedure is defined as the probability of one

false rejection, that is,

$$\mathsf{FWER} = \mathbb{P}(V > 0)$$

with notation of Table 2. FWER controlling procedures have been introduced for small- or medium-scale multiple testing, for situations in which high confidence in the rejected hypotheses is needed. The FDR has been introduced by Benjamini and Hochberg [7], who argue that FWER control might be too demanding for large-scale multiple testing, and especially exploratory approaches for which a small number of false positives may be tolerated. FDR is defined as the expected *False Discovery Proportion* (FDP), the fraction of false rejections among all rejected hypotheses: letting

$$\mathsf{FDP} = \frac{V}{R \vee 1} ,$$

we have

$$\mathsf{FDR} = \mathbb{E}\left[\mathsf{FDP}\right] .$$

A related quantity is the *positive False Discovery Rate* (pFDR), which is defined as the conditional expectation of the FDP given that at least one hypothesis is rejected:

$$\mathsf{pFDR} = \mathbb{E}\left[\mathsf{FDP} | R > 0\right] .$$

Weak control of a type I error rate means control under the complete null hypothesis (all null hypotheses are true), whereas *strong control* means control of a Type I error rate under any combination of true and false hypotheses. Throughout the thesis, we will focus on *strong control*, which is desirable for large-scale multiple testing as it is likely that at least some alternative hypotheses are true.

1.3. FDR control for multiple testing procedures

1.3.1. Mixture model. For $i \in \{1 \dots m\}$, where m is the number of tests performed, we let $Y_i = 0$ if hypothesis i is drawn from the null hypothesis $\mathcal{H}_0$, and $Y_i = 1$ if it is drawn from the alternative $\mathcal{H}_1$; X_i denotes the corresponding test statistic. We assume that the random variables $(X_i, Y_i)_{1 \leq i \leq m}$ are identically independently distributed: Y_i is a Bernoulli random variable with success probability ε_m, where ε_m is the unknown proportion of true alternatives; the conditional distribution of X_i given Y_i is denoted by $F_1^{(m)}$ if $Y_i = 1$ and $F_0^{(m)}$ if $Y_i = 0$. The marginal distribution of each X_i is thus

$$F^{(m)} = (1 - \varepsilon_m) F_0^{(m)} + \varepsilon_m F_1^{(m)} .$$

$F_0^{(m)}$ and $F_1^{(m)}$ are assumed to be C^1. This model may be equivalently formulated in terms of p-values rather than test statistics. Since $F_0^{(m)}$ is continuous, the p-values $(P_i)_{1 \leq i \leq m}$, which are defined by

$$P_i = 1 - F_0^{(m)}(X_i) ,$$

are uniformly distributed on $[0, 1]$ under $\mathcal{H}_0$; we let $G_0^{(m)}(x) = x$ for $0 \leq x \leq 1$. Letting $G_1^{(m)}$ and $g_1^{(m)}$ denote the distribution function and density function of the p-values under $\mathcal{H}_1$, the marginal distribution function and density of the p-values under the mixture are given by $G^{(m)} = (1-\varepsilon_m) G_0^{(m)} + \varepsilon_m G_1^{(m)}$

and $g^{(m)} = (1-\varepsilon_m) + \varepsilon_m g_1^{(m)}$. $(P_{(i)})_{1\leq i\leq m}$ denotes the vector of ordered p-values associated with $(P_i)_{1\leq i\leq m}$.

Settings. This mixture model will be considered in two different settings. In the *sparse setting*, we let ε_m converge to 0 and the distance between $\mathcal{H}_0$ and $\mathcal{H}_1$ (typically measured by the shift parameter in a location model) grow to $+\infty$ as the number m of tested hypotheses tends to $+\infty$. In the *fixed setting*, all parameters of the mixture model are fixed. In order to alleviate notation, the superscript m will be omitted in this setting. The fraction of $1-\varepsilon$ of true null hypotheses will be denoted by π_0.

Step-up and step-down multiple testing procedures. We will consider p-value based multiple testing procedures, that is, functions $\mathcal{M} : [0,1] \to [0,1]$ such that all hypotheses i satisfying

$$P_i \leq \mathcal{M}(P_1, \dots P_m)\,.$$

are rejected. $\mathcal{M}(P_1, \dots P_m)$ is called the *threshold* of procedure $\mathcal{M}$. A more formal definition of multiple testing procedures is given in chapter 2 (Definition 2.2.1). Step-up and step-down procedures are defined as follows:

DEFINITION 1.3.1 (Step-up procedure). *Let $(\alpha_i)_{1\leq i\leq m}$ be a non-decreasing sequence of numbers of $[0,1]$. The step-up procedure associated with $(\alpha_i)_{1\leq i\leq m}$ rejects all p-values less than α_K, with*

$$K = \sup\left\{i \in \{1\dots m\}, P_{(i)} \leq \alpha_i\right\}\,.$$

DEFINITION 1.3.2 (Step-down procedure). *Let $(\alpha_i)_{1\leq i\leq m}$ be a non-decreasing sequence of numbers of $[0,1]$. The step-down procedure associated with $(\alpha_i)_{1\leq i\leq m}$ rejects all p-values less than α_K, with*

$$K = \sup\left\{j \in \{1\dots m\}, \forall i \in \{1\dots j\}, P_{(i)} \leq \alpha_i\right\}\,.$$

The step-up procedure associated with the vector $(\alpha_i)_{1\leq i\leq m}$ therefore rejects more hypotheses than the step-down procedure associated with the same vector. In this thesis we will focus on step-up procedures. Since we have assumed that p-values are independent, "FDR control" means "FDR control under independence", unless otherwise specified.

1.3.2. The BH95 procedure.

Definition and graphical interpretation. The BH95 procedure is the step-up procedure associated with the vector $(\alpha i/m)_{1\leq i\leq m}$: it rejects p-values less than $\widehat{\tau} = \alpha \widehat{I}_m/m$, with

$$\widehat{I}_m = \max\left\{i \in \{1,\dots m\}, P_{(i)} \leq \alpha i/m\right\}\,.$$

This definition can be rewritten as follows. Let $\hat{\mathbb{G}}_m$ be the empirical distribution function of the p-values, then

$$\widehat{\tau} = \sup\left\{u \in [0,1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\}\,.$$

These two equivalent formulations of the BH95 threshold are illustrated in Figure 1. Following Finner et al. [30], $u \mapsto u/\alpha$ will be called the *rejection curve* of the BH95 procedure (also known as *Simes' line* [83]).

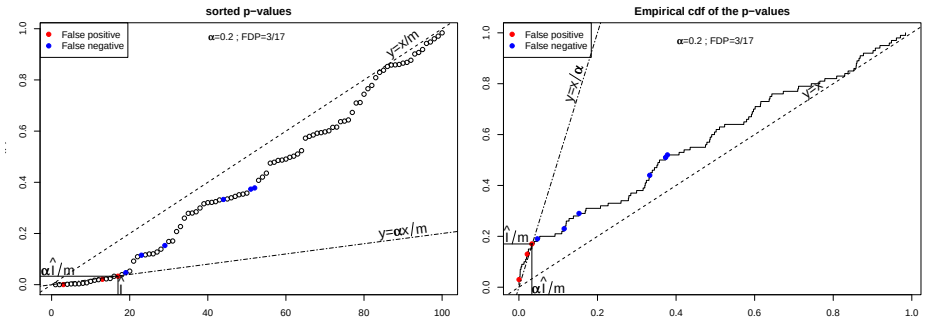


FIGURE 1. *Dual interpretations of the* BH95 *threshold.*

Properties. When the BH95 procedure is applied at level α, it yields $\mathsf{FDR} = \pi_0\alpha$ exactly when true null p-values are independent, or satisfy a specific positive dependency condition [8, 28, 79, 93]. In particular, this result holds for any joint distribution of the p-values under the alternative.

Criticality. Chi [16] recently demonstrated the existence of a *critical value* $\alpha^\star$ depending solely on the distribution function G of the p-values, such that if $\alpha < \alpha^\star$, the number of discoveries made by the BH95 procedure is stochastically bounded as the number of tested hypotheses increases, whereas if $\alpha > \alpha^\star$, the proportion of discoveries converges in probability to a positive value.

1.3.3. Estimation of π_0. When $\pi_0 < 1$, the BH95 procedure is thus conservative by a factor π_0 since it yields $\mathsf{FDR} = \pi_0\alpha$ for a target FDR level α. A number of methods have been proposed that estimate π_0, either implicitly or explicitly, in order to provide tighter (that is, more powerful) FDR control under independence: one-stage adaptive procedures use rejection curves other than Simes' line, without explicitly incorporating an estimate of π_0; two-stage adaptive procedures apply the BH95 procedure at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0.

Two-stage adaptive procedures. This class of procedures builds estimates of π_0 from the p-values or the test statistics; we refer to Broberg [12] and Langaas et al. [55] for a review.

Procedure Sto02**.** A number of such methods can be viewed as variants from an original graphical method by Schweder and Spjotvoll [80], which estimates π_0 from p-values larger than a given threshold because large p-values are more likely to come from the distribution under the null. This estimator has been popularized by Storey [89] in the context of FDR control:

$$\widehat{\pi_0}^{\mathsf{Sto02}}(\lambda) = \frac{1 - \hat{\mathbb{G}}_m(\lambda)}{1 - \lambda},$$

for $\lambda \in (0, 1)$. As $G(\lambda) = \pi_0\lambda + (1 - \pi_0)G_1(\lambda)$, we have

$$\mathbb{E}\left[\widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)\right] = \pi_0 + (1 - \pi_0)\frac{1 - G_1(\lambda)}{1 - \lambda},$$

and $\widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)$ is asymptotically anti-conservatively (that is, positively) biased. λ balances a bias-variance trade-off: when λ goes to 1, the bias of

$\widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)$ decreases but its variance increases because fewer points are used for the estimation. Several estimators based on $\widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)$ have been proposed [6, 91, 93], including

$$\widehat{\pi_0}^{\mathsf{STS04}}(\lambda) = \frac{1 + \frac{1}{m} - \hat{\mathbb{G}}_m(\lambda)}{1 - \lambda}\,.$$

The corresponding plug-in procedure, in which the BH95 procedure is applied at level $\alpha/\widehat{\pi_0}^{\mathsf{STS04}}(\lambda)$, is denoted by procedure STS04(λ); it controls FDR at level α [93].

Procedure BKY06**.** Letting $\beta = \frac{\alpha}{1+\alpha}$, procedure BKY06 applies procedure BH95 at level $\frac{\beta}{1-R(\beta)/m}$, where $R(\beta)$ is the number of hypotheses rejected by a first application of the BH95 procedure at level β. We consider a recently proposed generalization of this procedure [10], in which procedure BH95 is applied at level $\frac{1-\lambda}{1-R(\lambda)/m}\alpha = \alpha/\widehat{\pi_0}^{\mathsf{BKY06}}(\lambda)$, with

$$\widehat{\pi_0}^{\mathsf{BKY06}}(\lambda) = \frac{1 - \hat{\mathbb{G}}_m(u_\lambda)}{1 - \lambda}\,,$$

where u_λ is the threshold of the BH95 procedure applied at level λ (which satisfies $\hat{\mathbb{G}}_m(u_\lambda) = R(\lambda)/m$). The original BKY06 procedure corresponds to $\lambda = \frac{\alpha}{1+\alpha}$, and has been proved to control FDR at level α. A slightly modified version of procedure BKY06(λ) (with $1/m$ added in the numerator of the estimator of π_0) has been proved to control FDR at level α, for any $\lambda \in (0, 1)$ [10].

Other approaches. Other methods directly estimate the density g of the p-values at 1, for example by modeling g as a Beta-Uniform mixture [72], or by using histograms [12, 14]. An interesting alternative approach [23] is motivated by the fact that

$$\mathbb{E}\left[Q(P)\right] = \pi_0 \mathbb{E}_{g_0}\left[Q(P)\right] + (1 - \pi_0)\mathbb{E}_{g_1}\left[Q(P)\right]\,,$$

where $\mathbb{E}_{g_0}$ and $\mathbb{E}_{g_1}$ are expectations under the null and alternative distributions, and Q is a monotone transformation. They propose to use

$$\widehat{\pi_0}(Q) = \frac{1/m \sum_{i=1}^{m} Q(P_i)}{\mathbb{E}_0\left[Q(P)\right]}\,,$$

and exhibit conditions on Q under which $\widehat{\pi_0}(Q)$ has smaller bias than the estimator obtained with $Q = \mathrm{Id}$. In practice they suggest to use $Q : x \mapsto -\ln(1 - x)$. Unfortunately there is a gap in the proof of the main theorem in [23], as shown in section 1.6.1. Theorem 1.3.3 below provides a slightly different result; as it covers the case $Q : x \mapsto -\ln(1 - x)$, it proves that the estimator proposed in [23] indeed has smaller bias than the estimator obtained with $Q = Id$.

THEOREM 1.3.3 (Adapted from Dalmasso et al. [23]). *Let g_0 and g_1 be two probability density functions on $[0, 1]$ such that g_1/g_0 is non-increasing,*

and let Q be a real continuous function defined on $[0,1]$, such that $R : x \mapsto Q(x)/x$ is non decreasing. Then

$$\frac{\mathbb{E}_{g_1}[Q(X)]}{\mathbb{E}_{g_0}[Q(X)]} \leq \frac{\mathbb{E}_{g_1}[X]}{\mathbb{E}_{g_0}[X]}\,. \tag{1.3.4}$$

One-stage adaptive procedures. Some procedures have been proposed that do not explicitly incorporate an estimate of π_0, but are less conservative than the original BH95 procedure, because they use a non linear rejection curve instead of Simes' line: $u \mapsto u/\alpha$. The FDR08 procedure [30] is the step-up procedure associated with the rejection curve

$$f_\alpha : u \mapsto \frac{u}{\alpha + (1-\alpha)u}\,,$$

and the BR08(λ) procedure [10] is the step-up procedure associated with the rejection curve

$$b_\alpha^\lambda : u \mapsto \begin{cases} \frac{u}{\alpha(1-\lambda)+u} & \text{for } 0 \leq u \leq \lambda \\ +\infty & \text{else} \end{cases}\,,$$

for $0 < \lambda < 1$. Procedure BR08(λ) controls FDR for finitely many hypotheses, for any λ; as $f_\alpha(1) = 1$, procedure FDR08 always rejects all hypotheses and thus does not control FDR. Truncated versions of this procedure have been shown to control FDR asymptotically [30], and a step-down version using the same rejection curve has recently been proved to control FDR for finitely many hypotheses [33].

1.3.4. Unifying proofs of FDR control. Several arguments have been used to prove FDR control by adaptive procedures: continuous-time martingales for procedure STS04 [93], direct calculation for procedure BKY06 [9], and a self-consistency condition for procedure BR08 [10]. For the last two procedures, the corresponding proofs also both rely on the following counting argument:

LEMMA 1.3.5 (Benjamini et al. [9]). *If $Y \sim Bin(k-1, p)$, then we have $\mathbb{E}\left[1/(Y+1)\right] < 1/kp$.*

The arguments used to prove FDR control by procedures BKY06 and BR08 can be applied to procedure STS04 as well [9, 10], thus providing unifying proofs of FDR control. Conversely, the martingale argument permits proving that (slightly modified versions of) procedures BKY06(λ) and BR08(λ) control FDR, because both of these procedures are always more conservative than procedure STS04(λ).

To see this for procedure BKY06(λ), note that $\widehat{\pi_0}^{\mathsf{BKY06}}(\lambda) \geq \widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)$, since the threshold u_λ of procedure BH95 at level λ satisfies $u_\lambda \leq \lambda$. Therefore, adding $1/m$ to the numerator of $\widehat{\pi_0}^{\mathsf{BKY06}}(\lambda)$ yields to a more conservative procedure than STS04(λ). For procedure BR08(λ), note that[2] the threshold of this procedure may be written as

$$\widehat{\tau} = \sup\left\{ u \in [0,\lambda], \hat{\mathbb{G}}_m(u) \geq \frac{u}{\alpha}\frac{1-\hat{\mathbb{G}}_m(u)}{1-\lambda} \right\}\,.$$

[2]See chapter 2, section 2.5 for a formal proof.

Since $\frac{1-\hat{\mathbb{G}}_m(u)}{1-\lambda} \geq \widehat{\pi_0}^{\mathsf{Sto02}}(\lambda)$ for $u \leq \lambda$, using $\frac{1+1/m-\hat{\mathbb{G}}_m(u)}{1-\lambda}$ instead of $\frac{1-\hat{\mathbb{G}}_m(u)}{1-\lambda}$ in the preceding display also yields to a more conservative procedure than $\mathsf{STS04}(\lambda)$.

A continuous time martingale argument. In the remainder of this section, we discuss a continuity issue for using the martingale argument invoked by Storey et al. [93]. The idea of using martingales to prove FDR is motivated by the following Lemma for the BH95 procedure. The statements of this Lemma have been proved by Storey et al. [93]; for completeness we recall the proof in section 1.6.2.

LEMMA 1.3.6 (Motivation for a martingale argument [93]). *For $t \in (0,1]$, let V_t be the number of true null p-values smaller than t, and R_t the total number of p-values smaller than t. Denote by $\widehat{\tau}$ the threshold of the* BH95 *procedure at level α, and by $\mathcal{F} = (\mathcal{F}_t)_{0<t\leq 1}$ the natural (decreasing) filtration associated with (V_t), augmented with the p-values under the alternative distribution. $\mathcal{F}_t$ is defined for any $t \in (0,1]$ by $\mathcal{F}_t = \sigma\left\{(V_s)_{s\geq t}, (P_{(i)})_{i\sim\mathcal{H}_1}\right\}$. Then*

(i) $V_t|V_s \sim Bin\,(V_s, t/s)$;
(ii) (V_t/t) *is a $\mathcal{F}$-martingale with time running backwards;*
(iii) $\widehat{\tau}$ *is a $\mathcal{F}$-stopping time;*
(iv) $\frac{V_{\widehat{\tau}}}{R_{\widehat{\tau}}} = \frac{\alpha}{m}\frac{V_{\widehat{\tau}}}{\widehat{\tau}}$.

As a consequence of Lemma 1.3.6(iv), the FDR attained by procedure BH95 is given by $\mathsf{FDR}(\widehat{\tau}) = \mathbb{E}\left[V_{\widehat{\tau}}/R_{\widehat{\tau}}\right] = \alpha/m\mathbb{E}\left[V_{\widehat{\tau}}/\widehat{\tau}\right]$. As $\mathbb{E}\left[V_t/t\right] = \mathbb{E}\left[V_1\right] = m_0$ for any $t \in (0,1]$ (Lemma 1.3.6(ii)), one only needs an optional sampling argument to prove that $\mathsf{FDR}(\widehat{\tau}) = \pi_0\alpha$.

Existing optional sampling theorems for continuous-time martingales [70] require right-continuity of the martingale; as we are working with a reversed-time martingale, the condition we need, is *left-continuity*. However, the process (V_t/t) is cadlag and has left-discontinuities at each p-value coming from the null hypothesis. Storey's proof may nevertheless be rescued by noting that:

(i) the proof of the Optional Sampling Theorem only requires continuity of the process *at the stopping time*;
(ii) almost surely, V is continuous at $\widehat{\tau}$, as $\widehat{\tau}$ is not one of the p-values under the null hypothesis.

For (ii), note that V has less than m_0 points of left-discontinuity, which correspond to the distinct p-values drawn from the null distribution, and each of them has a null probability of being equal to $\widehat{\tau}$. For (i) we give a slight generalization of the classical Optional Sampling Theorem in section 1.6.2 (Theorem 1.6.2).

Combined with Lemma 1.3.6, Theorem 1.6.2 proves that procedure BH95 controls FDR at level exactly $\pi_0\alpha$, under any configuration of alternative hypotheses. Following the same lines, the martingale argument can be used to prove FDR control at level smaller than α by procedure $\mathsf{STS04}(\lambda)$ [93], and thus by (slight modifications of) procedures $\mathsf{BR08}(\lambda)$ and $\mathsf{BKY06}(\lambda)$, which are more conservative.

1.4. Tools for an asymptotic study

A widely used approach in asymptotic statistics (advocated by Pollard [71] for example) is to write a statistic as a functional on the sample paths of a stochastic process in order to break the analysis into two parts: the study of regularity of the functional; the study of the stochastic process as a random element of a space of functions.

This idea will be illustrated in Chapter 2, in which we establish Central Limit Theorems for the FDP achieved by a class of FDR controlling procedures. The classical tool to establish such theorems in Euclidean spaces is the Delta method [105]; we recall its formulation for real-valued random variables.

THEOREM 1.4.1 (Delta method [105]). *Let (X_m) be a sequence of real-valued random variables, $\theta \in \mathbb{R}$, and (r_m) a sequence growing to $+\infty$ as $m \to +\infty$. Assume that:*

(i) *$r_m (X_m - \theta) \rightsquigarrow X$, where X is a real-valued random variable;*
(ii) *$\phi : \mathbb{R} \to \mathbb{R}$ is differentiable at θ, with derivative $\phi'(\theta)$.*

Then we have

$$r_m (\phi(X_m) - \phi(\theta)) \rightsquigarrow \phi'(\theta) X$$

We are interested in more general situations in which X_n lives in the functional space $D[0,1]$ of cadlag functions on $[0,1]$, that is, right-continuous functions on $[0,1]$ with left limits[3], and ϕ maps $D[0,1]$ to $\mathbb{R}$. The extension of the usual definition of convergence in distribution to the non separable metric space $(D[0,1], \|\cdot\|_\infty)$ turns out to raise measurability issues that we discuss in section 1.6.3.

In the remainder of the present section, we begin by recalling a version of Donsker's Theorem that extends Assumption (i) of Theorem 1.4.1 to stochastic processes of $D[0,1]$ (section 1.4.1). Then we define Hadamard differentiability, which provides an extension for Assumption (ii) of Theorem 1.4.1 to normed spaces (section 1.4.2). Finally we show how these tools may be combined to yield a *functional Delta method* [105] (section 1.4.3).

1.4.1. Donsker's theorem. Letting $(X_m)_{m\in\mathbb{N}}$ be a sequence of independent, uniform, real-valued random variables on $[0,1]$, the uniform empirical process $U_m = (U_m(t))_{0\le t\le 1}$ is defined by

$$U_m(t) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{X_i \le t} \,,$$

for any $t \in [0,1]$. Donsker's Theorem establishes the convergence in distribution of U_m, as a process of $D[0,1]$, to the Brownian bridge, that is, a Gaussian process $\mathbb{B}$ on $[0,1]$, with covariance function $(s,t) \mapsto s \wedge t(1 - s \vee t)$, and such that $\mathbb{B}(0) = 0$, $\mathbb{B}(1) = 1$. It is also known as Empirical Central Limit Theorem. The earliest results about convergence in distribution of the empirical process go back to Donsker [26]; the version we recall, which is valid in $D[0,1]$ equipped with the uniform metric, is taken from Pollard [71].

[3] *cadlag* stands for the French "continues à droite, avec limites à gauche".

THEOREM 1.4.2 (Empirical Central Limit Theorem [71]). *The uniform empirical process on* $[0,1]$ *converges in distribution to the Brownian bridge.*

By the Continuous Mapping Theorem, this result may be extended to the empirical process associated with a sequence of independent random variables with any distribution function F:

COROLLARY 1.4.3. *Let* $F \in D[0,1]$ *be any distribution function, denote by* $\hat{\mathbb{F}}_m$ *the associated empirical distribution function. Then, as* $m \to +\infty$,

$$\sqrt{m}\left(\hat{\mathbb{F}}_m - F\right) \rightsquigarrow \mathbb{B} \circ F\,,$$

where $\mathbb{B}\circ F(t) = B(F(t))$ *is a* F-dilated Brownian bridge*, that is, a Gaussian process on* $[0,1]$ *such that* $\mathbb{B} \circ F(0) = \mathbb{B} \circ F(1) = 0$*, with covariance function*

$$(s,t) \mapsto F(s) \wedge F(t)(1 - F(s) \vee F(t))\,.$$

1.4.2. Hadamard differentiability. Although several notions of differentiability for maps between normed spaces $\mathbb{D}$ and $\mathbb{E}$ have been defined, not all of them are appropriate for translating convergence in distribution in $\mathbb{D}$ into convergence in distribution in $\mathbb{E}$: Gâteaux differentiability is too weak for this purpose, while Fréchet differentiability in unnecessarily strong (and thus might not hold for a particular application). It turns out that the appropriate choice is Hadamard differentiability, which is "intermediate"between Gâteaux and Fréchet differentiability, in the sense that Fréchet differentiability implies Hadamard differentiability, which in turn implies Gâteaux differentiability.

DEFINITION 1.4.4 (Hadamard differentiability). *Let* $\mathbb{D}$ *and* $\mathbb{E}$ *be two normed spaces, and* $\phi : \mathbb{D} \to \mathbb{E}$ *be defined on a subset* $\mathbb{D}_\phi$ *of* $\mathbb{D}$*. The function* ϕ *is Hadamard differentiable at* $\theta \in \mathbb{D}$ *if and only if there is a continuous linear map* $\dot{\phi}_\theta : \mathbb{D} \to \mathbb{E}$*, such that for any family* $(h_t)_{t>0}$ *of* D_ϕ *with limit* h *as* $t \to 0$*,*

$$\left\| \frac{\phi(\theta + t h_t)\phi(\theta)}{t} - \dot{\phi}_\theta(h) \right\|_{\mathbb{E}} \to 0\,. \tag{1.4.5}$$

The function ϕ *is Hadamard differentiable at* θ tangentially *to a set* $\mathbb{D}_0 \subset \mathbb{D}$ *if display (1.4.5) is only required to hold for* (h_t) *with limits* $h \in \mathbb{D}_0$*; the derivative needs then be defined on* $\mathbb{D}_0$ *only.*

1.4.3. Functional delta method.

THEOREM 1.4.6 (Functional delta method [105]). *Let* $\mathbb{D}$ *and* $\mathbb{E}$ *be two normed spaces, and* $\phi : \mathbb{D} \to \mathbb{E}$*, and* $\mathbb{D}_0$ *be a separable subset of* $\mathbb{D}$*. Let* (X_m) *be a sequence of* $\mathbb{D}$*-valued processes,* $\theta \in \mathbb{D}$*, and* (r_m) *a sequence growing to* $+\infty$ *as* $m \to +\infty$*. Assume that:*

(i) $r_m\,(X_m - \theta) \rightsquigarrow X$*, where* X *takes its values in* $\mathbb{D}_0$*;*
(ii) ϕ *is Hadamard differentiable at* $\theta \in \mathbb{D}$ tangentially *to* $\mathbb{D}_0$*.*

Then we have

$$r_m\,(\phi(X_m) - \phi(\theta)) \rightsquigarrow \dot{\phi}_\theta(X)$$

In chapter 2 we use Theorem 1.4.6 with $\mathbb{D} = D[0,1]$ and $\mathbb{D}_0 = C[0,1]$, the set of continuous functions on $[0,1]$, which is a separable subset of $D[0,1]$ for the uniform metric.

1.5. Contributions

1.5.1. Chapter 2: Asymptotic FDP.

Motivation. As the proportion of erroneous rejections (FDP) is a stochastic quantity, its fluctuations around its mean value (FDR) are worth being investigated. The asymptotic behavior of process $(\mathsf{FDP}_m(t))_{0<t\leq 1}$, where t is a deterministic threshold, has already been studied [35, 89, 93]. In this chapter we are interested in the properties of the *random threshold* $\widehat{\tau}$ of a given multiple testing procedure, and especially in the asymptotic distribution of $\mathsf{FDP}_m(\widehat{\tau})$, that is, of the FDP actually reached by the procedure.

We consider procedures whose threshold $\widehat{\tau}$ may be written as a functional $\mathcal{T}$ of the empirical distribution function $\hat{\mathbb{G}}_m$ of the p-values: $\mathcal{T}$ will be called a *threshold function*. This is the case for the BH95 procedure, whose threshold at level α is given by

$$\widehat{\tau} = \sup\left\{u \in [0,1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\} .$$

More generally, many one-stage adaptive and two-stage adaptive procedures may be written using the formalism of threshold functions; thus the threshold $\widehat{\tau}$ of these procedures and its associated FDP may be written as *stochastic processes of a random threshold*. The tools described in section 1.4 help us studying the asymptotic behavior of these quantities.

Contributions. We prove that FDP of a multiple testing procedure with threshold function $\mathcal{T}$ converges in distribution at rate $1/\sqrt{m}$ to a conservative, procedure-specific FDR level under the assumption that $\mathcal{T}$ is Hadamard-differentiable at G tangentially to the set $C[0,1]$ of continuous functions of $D[0,1]$. This general regularity assumption is implied by the existence and uniqueness of an interior right-crossing point between the distribution function of the p-values and the rejection curve of the procedure; the existence condition for a given procedure may be interpreted as a natural generalization of the notion of *criticality* discussed in section 1.3.2 for the BH95 procedure.

We derive the asymptotic behavior and the associated regularity conditions for a number of FDR controlling procedures, including one-stage adaptive procedures (BR08 and FDR08), and two-stage adaptive (or plug-in) procedures (Sto02 or STS04, and BKY06). As all procedures converge at the same rate $1/\sqrt{m}$, their asymptotic power may be explicitly compared through their attained asymptotic FDR.

We demonstrate the existence of interesting connections between one-stage and two-stage adaptive procedures under investigation: with a striking symmetry, procedure BR08 may be interpreted as a fixed point of the iteration of procedure BKY06, and procedure FDR08 as a fixed point of the iteration of procedure Sto02.

1.5.2. Chapter 3: Intrinsic bounds and FDR control.

Three multiple testing problems. Statistical questions that arise when testing a large number of hypotheses include not only the selection of false null hypotheses, which has been discussed in this chapter, but also a detection and an estimation problem. We are therefore concerned with the following questions:

detection: Are there any false null hypotheses ?
estimation: How many null hypotheses are false ?
selection: Which null hypotheses are false ?

The detection problem is a test of the null hypothesis that the proportion ε_m is 0 against the alternative hypothesis that it is positive, whereas the estimation problem is to estimate ε_m. Although they are *single* testing and estimation problems, the multiple comparison context in which they are cast motivates the need for appropriate testing and estimation procedures.

Intrinsic bounds on multiple testing problems. Recent work demonstrates the existence of intrinsic bounds to these problems. For the selection problem, the criticality phenomenon mentioned in section 1.3.2 illustrates the existence of a possibly positive lower bound below which no multiple testing procedure can control the pFDR. For the detection problem, a *detection boundary* has been identified, which characterizes situations in which the Likelihood Ratio Test asymptotically almost surely correctly detects [24]. Likewise, an *estimation boundary* for sparse Gaussian mixtures characterizes situations in which ε_m can be consistently estimated [13].

Contributions. Chapter 3 is motivated by the comparison of these intrinsic bounds in the sparse and non sparse settings, in the context of FDR control. In the non sparse setting, we demonstrate that the criticality phenomenon only occurs for heavy-tailed distributions such as the Laplace (bilateral exponential) or Student distributions, and we prove that for symmetric location problems in which the test statistics under the alternative is a positive shift of the test statistics under the null, criticality for the selection problem occurs if and only if $\pi_0 = 1 - \varepsilon$ is not identifiable in the estimation problem. We also connect attainable convergence rates for the estimation problem in the non sparse setting to the regularity of the distribution function of the p-values in a neighborhood of 1, and argue that this regularity is typically poor, which results in slow rates of convergence for plug-in procedures defined in section 1.3.3. Finally, we discuss the performances of the BH95 procedure in the sparse setting, and propose an interpretation of the detection boundary of this procedure in terms of pFDR control.

1.6. Proofs

1.6.1. On using bias-reducing transformations to estimate π_0. Recall that the p-value P is distributed as $g_0 = \mathcal{U}[0,1]$ under the null hypothesis, and as g_1 under the alternative. In the proof of the Theorem of [23] it is claimed (end of page 667) that

$$\frac{\mathbb{E}_{g_1}[P] - \mathbb{E}_{g_0}[P]}{\mathbb{E}_{g_0}[Q(P)]} \leq \frac{\mathbb{E}_{g_1}[P] - \mathbb{E}_{g_0}[P]}{Q(\mathbb{E}_{g_0}[P])}\,. \tag{1.6.1}$$

As Q is convex, Jensen's inequality ensures that $\mathbb{E}_{g_0}[Q(P)] \geq Q(\mathbb{E}_{g_0}[P])$. By Assumption (iv) of the Theorem, $Q(\mathbb{E}_{g_0}[P]) \geq \mathbb{E}_{g_0}[P] = 1/2$. Thus, $\mathbb{E}_{g_0}[Q(P)]$ and $Q(\mathbb{E}_{g_0}[P])$ are positive, and inequality (1.6.1) holds if and only if $\mathbb{E}_{g_1}[P] \geq \mathbb{E}_{g_0}[P]$, which typically does not hold because p-values under the alternative are *smaller* in expectation than under the null. In this section we prove Theorem 1.3.3 (page 14), which provides a slightly different version of the result from [23].

PROOF OF THEOREM 1.3.3. Let us note that inequality 1.3.4 is equivalent to

$$\frac{\mathbb{E}_{g_1}[Q(P)]}{\mathbb{E}_{g_1}[P]} \leq \frac{\mathbb{E}_{g_0}[Q(P)]}{\mathbb{E}_{g_0}[P]},$$

which can be rewritten as

$$\mathbb{E}_{h_1}[R(P)] \leq \mathbb{E}_{h_0}[R(P)]\,,$$

where $h_0 : x \mapsto \frac{xg_0(x)}{\mathbb{E}_{g_0}[P]}$, $h_1 : x \mapsto \frac{xg_1(x)}{\mathbb{E}_{g_1}[P]}$ are two probability density functions on $[0,1]$. As R is non decreasing and $\frac{h_1}{h_0} = \frac{\mathbb{E}_{g_0}[P]}{\mathbb{E}_{g_1}[P]}\frac{g_1}{g_0}$ is non-increasing, we have

$$\begin{aligned}
\mathbb{E}_{h_1}[R(P)] &= \mathbb{E}_{h_0}\left[\frac{h_1(P)}{h_0(P)}R(P)\right] \\
&\quad \text{by Chebychev association inequality} \\
&\leq \mathbb{E}_{h_0}\left[\frac{h_1(P)}{h_0(P)}\right]\mathbb{E}_{h_0}\left[R(P)\right] \\
&= \mathbb{E}_{h_0}\left[R(P)\right]
\end{aligned}$$

□

1.6.2. A continuous time optional sampling theorem.

PROOF OF LEMMA 1.3.6. As p-values under the null distribution are independently, uniformly distributed, item (i) results from a simple counting argument, and (ii) follows from (i). For (iii), recalling that

$$\widehat{\tau} = \sup\left\{u \in [0,1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\},$$

we have $\widehat{\tau} \geq t$ if and only if $\hat{\mathbb{G}}_m(t) < t/\alpha$. As $\hat{\mathbb{G}}_m(t)$ only depends on p-values under the alternative, and on p-values under the null which are smaller than t, we have $\{\widehat{\tau} \geq t\} \in \mathcal{F}_t$, and (iii) is proved. By the definition of the threshold $\widehat{\tau}$ of the BH95 procedure, we have $\hat{\mathbb{G}}_m(\widehat{\tau}) = \widehat{\tau}/\alpha$, which proves (iv) because $\hat{\mathbb{G}}_m(\widehat{\tau}) = R_{\widehat{\tau}}/m$. □

THEOREM 1.6.2 (Optional Sampling Theorem (adapted from [70])). *Let $\{(X_t, \mathcal{F}_t) : 0 \leq t \leq 1\}$ be a martingale, and let $0 \leq \sigma \leq \tau \leq 1$ be stopping times for the filtration, such that almost surely, X_t is right-continuous at σ and τ. Then, almost surely,*

$$\mathbb{E}\left[X_\tau | \mathcal{F}_\sigma\right] = X_\sigma\,.$$

PROOF OF THEOREM 1.6.2. The proof follows the lines of [70, Theorem 6, Appendix E], as right-continuity in only needed at σ and τ. For each $n \in \mathbb{N}$, let $\tau_n = \frac{\lceil 2^n\tau\rceil}{2^n}$ and $\sigma_n = \frac{\lceil 2^n\sigma\rceil}{2^n}$ be τ and σ rounded up to the next integer multiple of 2^{-n}. As we rounded *up*, each τ_n and each σ_n are stopping times for the filtration $\{\mathcal{F}_{i/2^n} : 0 \leq i \leq 2^n\}$. The discrete version of the Optional Sampling Theorem ensures that

$$\mathbb{E}\left[X_{\tau_n} | \mathcal{F}_{\sigma_n}\right] = X_{\sigma_n}\,.$$

As X_t is right-continuous *at τ and σ*, we have $X_{\sigma_n} \to X_\sigma$ and $X_{\tau_n} \to X_\tau$ along each sample path. To conclude it is sufficient to prove that X_{σ_n} and

X_{τ_n} are uniformly integrable; we refer to [70] for the end of the proof, as it does not rely on right-continuity. □

1.6.3. Convergence in distribution in $D[0,1], \|\cdot\|$.

Random elements in $D[0,1]$. Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $(\mathbb{D}, d)$ be a metric space. Denote by $\mathcal{B}(\mathbb{D})$ the Borel σ-field $\mathcal{B}_d$ generated by the closed sets of $\mathbb{D}$ under the metric d.

When $\mathbb{D}$ is the euclidean space $\mathbb{R}^n$, a random variable is defined as a measurable map from $(\Omega, \mathcal{A}, P)$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. If $(\mathbb{D}, d)$ is a more general metric space such as $D[0,1]$, it seems natural to define a random element X of $\mathbb{D}$ as a measurable map from $(\Omega, \mathcal{A}, P)$ to $(\mathbb{D}, \mathcal{B}_b(\mathbb{D}))$, that is, a map such that $X^{-1}(D) \in \mathcal{A}$ for any $D \in \mathcal{B}_d(\mathbb{D})$. This property is known as *Borel-measurability*. When $\mathbb{D} = D[0,1]$ is equipped with the Borel σ-field $\mathcal{B}_\infty$ generated by the closed sets under the uniform metric, it turns out that *the empirical processes need not be random elements of $D[0,1]$ in this sense [71, Chapter 4]*. The reason for this is that the space $D[0,1]$, equipped with the uniform metric, is nonseparable; hence the Borel sigma field is so large that $X^{-1}(D)$ needs not belong to $\mathcal{A}$ for any $D \in \mathbb{B}_\infty$, and Borel-measurability fails to hold.

Three ways to circumvent this problem have been successively proposed. The first one is to modify the metric so as to make $D[0,1]$ separable; the most popular example is Skorokhod's J_1 metric [84]. This solution permits working with the usual notion of measurability, at the price of a greater topological complexity. The other two solutions keep working with the uniform metric, at the price of modifying the notion of measurability itself, either by using a smaller σ-field than $\mathcal{B}_\infty$ [71], or by defining a generalized expectation even for maps that need not be random elements [106]. In the remainder of this chapter we give a quick look at these two solutions.

Redefining random elements. Random elements may be defined using other σ-fields than the Borel σ-field:

Definition 1.6.3 (Random elements in metric spaces). *A random element of $(\mathbb{D}, d)$ is a measurable map from $(\Omega, \mathcal{A}, P)$ to $(\mathbb{D}, d, \mathcal{D})$, where $\mathcal{D}$ is* any *σ-field over $(\mathbb{D}, d)$.*

The choice of the σ-field $\mathcal{D}$ in this definition is of importance: on the one hand, one should guard against too large a σ-field, which would make the corresponding definition of a random element too restrictive. Conversely, too small a σ-field would make the associated weak convergence theory trivial.

In chapter 2 we are working on empirical processes associated with a continuous distribution function. The set $C[0,1]$ of continuous functions on $[0,1]$ is a separable subset of $D[0,1]$ for the uniform metric; in such situations where processes concentrate on a separable subset of the original metric space, Pollard [71] advocates the use of the *ball σ-field*, that is, the σ-field $\mathbb{P}$ generated by closed balls.

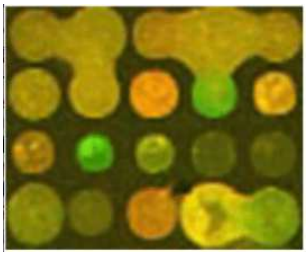
With this choice $(D[0,1], \|\cdot\|, \mathbb{P})$, Pollard [71] is able to prove a *Continuous mapping theorem* which serves as a basis for the proof of Donsker's Theorem (Theorem 1.4.2), and the corresponding functional Delta method (Theorem 1.4.6).

Outer Expectations. Instead of working with smaller σ-fields, it is possible to stick to the Borel σ-field, and relax the measurability requirement in the definition of random elements. In this case, the definition of expectation has to be generalized to non-measurable maps. *Outer expectations* provide such a generalization [105, Chapter 18], and permit proving a continuous mapping theorem from which a version of Donsker's Theorem and the functional Delta method may be derived.

CHAPTER 2

Asymptotic Properties of FDR controlling procedures

MINIMALISM



Mary Brinig, *B24n013*, 2003

Tony Smith, *Untitled*, 1961

Contents

2.1. Introduction

The BH95 procedure defined in Chapter 1 controls FDR when the true null hypotheses are independent, or display certain forms of positive dependence [7, 8]. Since applying the BH95 procedure at level α actually yields FDR $\pi_0\alpha$, where π_0 is the unknown proportion of true null hypotheses, considerable efforts have been devoted to the design of procedures that increase the number of rejections while keeping FDR $< \alpha$. These procedures are called *two-stage adaptive* when they explicitly incorporate an estimator of π_0, and *one-stage adaptive* when π_0 is estimated implicitly.

The FDR controlling properties of such procedures have been carefully studied for a finite number of hypotheses hypotheses [7, 9, 10, 28, 29, 79, 89], or asymptotically [16, 28, 30, 34, 35, 89, 90, 93]. As the proportion of erroneous rejections (FDP) is a stochastic quantity, its fluctuations around its mean value are worth investigating. Several procedures have been proposed for controlling the upper quantiles of the FDP [35, 36, 56, 68, 75–77, 104]. The asymptotic behavior of process $(\mathsf{FDP}_m(t))_{0<t\leq 1}$, where t is a deterministic threshold, has also been studied [35, 89, 93]. We focus in this chapter on the properties of the *random threshold* $\widehat{\tau}$ associated with a given multiple testing procedure, particularly in the asymptotic distribution of $\mathsf{FDP}_m(\widehat{\tau})$, the FDP actually reached by the procedure.

This chapter is organized as follows. In section 2.2 we propose a general framework for asymptotic analysis of the FDP of multiple testing procedures. In section 2.3 we derive the asymptotic distribution of the FDP of a multiple testing procedure with generic threshold function $\mathcal{T}$ and characterize the asymptotic equivalence of multiple testing procedures. These results are explicitly connected to the regularity of the map $\mathcal{T}$, which is then discussed. In section 2.4 we derive the asymptotic behavior of several existing procedures. In section 2.5 we point out interesting connections between one-stage adaptive and two-stage adaptive procedures. The main results are summarized and discussed in section 2.6, and proofs of the main results are gathered in section 2.7.

2.2. Background and notation

In this chapter, we consider the "fixed" version of the mixture model presented in Chapter 1, where the parameters do not depend on the number of tested hypotheses. The above-defined quantities are now recalled and described more formally.

2.2.1. Background. We consider a sequence $(P_i)_{i\in\mathbb{N}}$ of p-values associated with a collection of binary tests of a null hypothesis $\mathcal{H}_0^i$ against an alternative hypothesis $\mathcal{H}_1^i$.

Definition 2.2.1 (Multiple Testing Procedure (MTP)). *A multiple testing procedure $\mathcal{M}$ is a sequence of functions $\mathcal{M}_m : [0,1]^m \to [0,1]$ such that for any m-dimensional vector of p-values $(P_1, \dots P_m)$, all hypotheses i satisfying*

$$P_i \leq \mathcal{M}_m(P_1, \dots P_m)\,.$$

are rejected. Slightly abusing notation, we shall write $\mathcal{M}(P_1, \dots P_m)$ for $\mathcal{M}_m(P_1, \dots P_m)$.

Denoting by V_m and R_m the number of illegitimate rejections and the total number of rejections among the m tested hypotheses for a multiple testing procedure $\mathcal{M}$, the associated False Discovery Proportion and False Discovery Rate are $\mathsf{FDP}_m(\mathcal{M}) = \frac{V_m}{R_m \vee 1}$, and $\mathsf{FDR}_m(\mathcal{M}) = \mathbb{E}\left[\frac{V_m}{R_m \vee 1}\right]$.

Mixture model. Denoting by π_0 the proportion of true null hypotheses, we assume that p-values are uniformly distributed on $[0,1]$ under $\mathcal{H}_0$, and distributed according to G_1 under $\mathcal{H}_1$, where G_1 is a concave, C^1 distribution function, with density g_1. We also assume that all p-values are independent, so that

$$(P_i)_{1\leq i\leq m} \overset{iid}{\sim} G\,,$$

where $G(x) = \pi_0 x + (1-\pi_0)G_1(x)$. The corresponding density function is given by $g = \pi_0 + (1-\pi_0)g_1$. Using this notation, the $\mathsf{BH95}$ procedure at level α is defined as

$$\mathcal{M}^\alpha(P_1, \dots P_m) = \sup\left\{u \in [0,1], \hat{\mathbb{G}}_m(u) \geq u/\alpha\right\}\,,$$

where $\hat{\mathbb{G}}_m$ is the empirical distribution function of the p-values, and $u \mapsto u/\alpha$ is called the *rejection curve* of the $\mathsf{BH95}$ procedure (also known as *Simes' line* [83]).

Threshold functions. This interpretation of the $\mathsf{BH95}$ procedure in terms of the empirical distribution function suggests to define *threshold functions* as follows. Let $D[0,1]$ denote the set of cadlag functions defined on $[0,1]$.

Definition 2.2.2 (Threshold function). *A multiple testing procedure $\mathcal{M}$ has threshold function $\mathcal{T} : D[0,1] \to [0,1]$ if and only if*

$$\forall m \in \mathbb{N}, \mathcal{M}(P_1, \dots P_m) = \mathcal{T}(\hat{\mathbb{G}}_m)\,.$$

Note that $\mathcal{T}$ does not depend on m in Definition 2.2.2. From now on $\mathcal{T}(G)$ will be denoted by $\tau^\star$.

2.2.2. FDP as a stochastic process of a random threshold. As suggested in a previous study [35], the False Discovery Proportion can be viewed as a stochastic process. Let $\hat{\mathbb{G}}_{0,m}$ and $\hat{\mathbb{G}}_{1,m}$ denote the (unobservable) empirical distribution function of the p-values under the null and alternative hypotheses, and $\hat{\mathbb{G}}_m = \pi_0\hat{\mathbb{G}}_{0,m} + (1-\pi_0)\hat{\mathbb{G}}_{1,m}$. Then, for any $t \in [0,1]$, we have $R_m(t) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{P_i\leq t} = \hat{\mathbb{G}}_m(t)$, and $V_m(t) = \frac{1}{m}\sum_{\{i/\mathcal{H}_0^i \text{ true}\}} \mathbf{1}_{P_i\leq t} = \pi_0\hat{\mathbb{G}}_{0,m}(t)$, so that

$$\mathsf{FDP}_m(t) = \frac{\pi_0\hat{\mathbb{G}}_{0,m}(t)}{\hat{\mathbb{G}}_m(t) \vee \frac{1}{m}}$$

is the False Discovery Proportion achieved at the deterministic threshold t. The asymptotic properties of the stochastic process $(\mathsf{FDP}_m(t))_{0\leq t\leq 1}$ were analyzed by Genovese and Wasserman [35]. They noticed that $\mathsf{FDR}_m(t) = \mathbb{E}\left[\mathsf{FDP}_m(t)\right]$, so the achieved FDR at t, may be written as

$$\mathsf{FDR}_m(t) = \mathsf{p}(t)\left(1 - (1-G(t))^m\right),$$

where $\mathsf{p}(t) = \frac{\pi_0 t}{G(t)}$ is the *positive False Discovery Rate* (pFDR) at t, as defined by [89]. They proved that the FDP_m process converges to pFDR at a rate $\frac{1}{\sqrt{m}}$, and built *confidence envelopes* for the FDP process using this result.

We make use of this stochastic process approach here to study the behavior of the FDP *actually achieved by a given multiple testing procedure* $\mathcal{T}$, that is, the random variable $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m))$. We investigated the asymptotic behavior of this variable and, in particular, its fluctuations around the asymptotic FDR achieved by procedure $\mathcal{T}$, by writing $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m))$ as a function of the empirical distribution functions under the null and alternative hypotheses. Letting

$$\mathcal{V} : (F_0, F_1) \mapsto \pi_0 F_0(\mathcal{T}(\pi_0 F_0 + (1-\pi_0)F_1))$$

and

$$\mathcal{R} : F \mapsto F(\mathcal{T}(F)),$$

the FDP achieved by procedure $\mathcal{T}$ may be written as

$$\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) = \frac{\mathcal{V}(\hat{\mathbb{G}}_{0,m}, \hat{\mathbb{G}}_{1,m})}{\mathcal{R}(\pi_0\hat{\mathbb{G}}_{0,m} + (1-\pi_0)\hat{\mathbb{G}}_{1,m}) \vee \frac{1}{m}}$$

since $\hat{\mathbb{G}}_m = \pi_0\hat{\mathbb{G}}_{0,m} + (1-\pi_0)\hat{\mathbb{G}}_{1,m}$. Using the functional Delta method [105], this formalism makes it possible to break down the analysis of $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m))$ into the regularity properties of the map $\mathcal{T}$, which depend solely on the procedure, and the asymptotic behavior of the empirical distribution functions of the p-values, which can be derived from Donsker's invariance principle [26] because p-values are assumed to be independent.

REMARK 2.2.3. Although we focus on FDP, the formalism we propose here may be used to derive the asymptotic distribution of any risk measure based on the number of true/false positive/negatives, under the same regularity conditions. In particular, the results obtained here can also be applied to the False Non-discovery Proportion (FNP) [34]:

$$\mathsf{FNP}_m(t) = \frac{(1 - R_m(t)/m) - (\pi_0 - V_m(t)/m)}{1-\pi_0}$$

2.2.3. Multiple testing procedures studied. The threshold function of the BH95 procedure is defined by

$$\mathcal{T}(F) = \sup\{u \in [0,1], F(u) \geq u/\alpha\}\,.$$

As the BH95 procedure keeps the false discovery rate at a level of (exactly) $\pi_0\alpha$ when p-values are independent [8, 28, 79, 93], it is conservative by a factor π_0. Other multiple testing procedures have been proposed that estimate π_0, either implicitly or explicitly, to provide tighter (i.e. more powerful) FDR control under independence:

One-stage adaptive procedures (BR08 [10], FDR08 [30]): use rejection curves other than Simes' line, without explicitly incorporating an estimate of π_0.

Two-stage adaptive procedures (BKY06 [9], STS04 [93], Sto02 [89]): apply the BH95 procedure at a level of $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0.

We therefore consider threshold functions of the form

$$\mathcal{T}(F) \quad = \quad \mathcal{U}\left(F, \mathcal{A}(F)\right), \tag{2.2.4}$$

with

$$\mathcal{U}(F,\alpha) = \sup\{u \in [0,1], F(u) \geq r_\alpha(u)\}, \tag{2.2.5}$$

where $r_\alpha : [0,1] \to \mathbb{R}_+$ will be called a *rejection curve* (after [30]), and $\mathcal{A} : D[0,1] \to [0,1]$ will be called a *level function*. r_α will be denoted by $r(\alpha,\cdot)$ whenever the dependence on α is of importance. $\mathcal{A}$ and r_α are two degrees of freedom that can be used to describe generalizations of the BH95 procedure, corresponding to the case in which the level function is constant (equal to α), *and* the rejection curve is Simes' line. We consider increasing rejection curves satisfying $r_\alpha(0) = 0$, so that $\mathcal{U}(F,\alpha) \geq 0$ for any $F \in D[0,1]$ and $\alpha \in [0,1]$.

2.2.4. Overview of main results of this chapter. Theorem 2.3.2 shows that the FDP of a multiple testing procedure with threshold function $\mathcal{T}$ converges in distribution at rate $1/\sqrt{m}$ to a conservative, procedure-specific FDR level. This theorem holds under a general regularity condition on the map $\mathcal{T}$, which is implied by the existence and uniqueness of an interior right-crossing point between the distribution function of the p-values and the rejection curve of the procedure; the existence condition for a given procedure may be interpreted as a natural generalization of the notion of *criticality*, which has recently been introduced for the BH95 procedure [16].

Although the BH95 procedure is known to control FDR at a level of *exactly* $\pi_0\alpha$ [8, 28], other procedures have been proved to yield only an FDR *not larger than* α, either for a finite number of hypotheses (procedures STS04, BKY06 and BR08) or asymptotically (Sto02 and FDR08). In section 2.4 we derive the asymptotic behavior of each procedure of interest, and the associated regularity conditions. As all procedures converge at the same rate $1/\sqrt{m}$, their asymptotic power may be explicitly compared through their attained asymptotic FDR.

In section 2.5 we demonstrate the existence of interesting connections between the one-stage and two-stage adaptive procedures under investigation: with a striking symmetry, procedure BR08 may be interpreted as a fixed point of the iteration of procedure BKY06, and procedure FDR08 as a fixed point of the iteration of procedure Sto02.

2.3. Asymptotic properties of threshold procedures

This section provides general results about multiple testing procedures with threshold functions satisfying the following regularity condition:

Condition C.1 (Hadamard-differentiability). *The threshold function $\mathcal{T}$ satisfies $\mathcal{T}(G) > 0$, and is Hadamard-differentiable at G, tangentially to $C[0,1]$, where $C[0,1]$ is the set of continuous functions on $[0,1]$ The threshold function derivative is denoted by $\dot{\mathcal{T}}_G$.*

We begin by deriving the asymptotic distribution of the FDP of any multiple testing procedure satisfying Condition C.1 (section 2.3.1). We then

define and characterize asymptotic equivalence between multiple testing procedures in terms of Condition C.1 (section 2.3.2). Finally we interpret this Condition in terms of crossing points between the distribution function G of the p-values and the rejection curve (section 2.3.3).

2.3.1. Asymptotic False Discovery Proportion. Condition C.1 makes it possible to use the functional Delta method [105] to derive the asymptotic distribution of the False Discovery Proportion $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m))$ actually achieved by procedure $\mathcal{T}$ from the convergence in distribution of the centered empirical processes associated with $\hat{\mathbb{G}}_{0,m}$ and $\hat{\mathbb{G}}_{1,m}$, which is a consequence of Donsker's theorem [105]:

THEOREM 2.3.1 (Donsker). *If the p-values are independent, then*

(i) $\sqrt{m}\left(\begin{pmatrix}\hat{\mathbb{G}}_{0,m}\\ \hat{\mathbb{G}}_{1,m}\end{pmatrix} - \begin{pmatrix}G_0\\ G_1\end{pmatrix}\right) \rightsquigarrow \begin{pmatrix}\mathbb{Z}_0\\ \mathbb{Z}_1\end{pmatrix}$ *on* $[0,1]$, *where* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$, *where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$.

(ii) $\sqrt{m}\left(\hat{\mathbb{G}}_m - G\right) \rightsquigarrow \mathbb{Z}$ *on* $[0,1]$, *where* $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ *is a stochastic process with continuous sample paths and independent, Gaussian increments, with covariance function given by*

$$\mathbb{E}\left[\mathbb{Z}(s)\mathbb{Z}(t)\right] \;=\; \pi_0^2\gamma_0(s,t) + (1-\pi_0)^2\,\gamma_0(G_1(s),G_1(t)),$$

where γ_0 *is the covariance function of* $\mathbb{B}$, *that is,* $\gamma_0 : (s,t) \mapsto s \wedge t(1 - s \vee t)$.

THEOREM 2.3.2 (Asymptotic distribution of FDP_m for procedure $\mathcal{T}$). *Let* $\mathcal{T}$ *be a threshold function,* $\tau^\star = \mathcal{T}(G)$, *and* $\mathsf{p}(t) = \frac{\pi_0 t}{G(t)}$ *the positive False Discovery Rate at threshold* t. *Under Condition C.1,*

(i)
$$\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star)\right) \rightsquigarrow \dot{\mathcal{T}}_G(\mathbb{Z})\,,$$

(ii)
$$\lim_{m\to\infty} \mathsf{FDR}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) = \mathsf{p}(\tau^\star)\,,$$

(iii)
$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}(\tau^\star)\right) \rightsquigarrow X\,,$$

with

$$X = \mathsf{p}(\tau^\star)(1-\mathsf{p}(\tau^\star))\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\right) + \dot{\mathsf{p}}(\tau^\star)\dot{\mathcal{T}}_G(\mathbb{Z})\,,$$

where $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ *and* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$, *where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$.

According to (*ii*), the asymptotic FDR achieved by procedure $\mathcal{T}$ is the pFDR at the asymptotic threshold $\tau^\star = \mathcal{T}(G)$. This is true because $\tau^\star$ is positive (by Condition C.1). In particular, Theorem 2.3.2 provides a necessary and sufficient condition under which a multiple testing procedure with Hadamard differentiable threshold function asymptotically controls FDR:

COROLLARY 2.3.3. *A threshold function $\mathcal{T}$ satisfying Condition C.1 asymptotically controls* FDR *if and only if its* pFDR *at $\tau^\star = \mathcal{T}(G)$ (i.e. its asymptotic* FDR*) is below α, that is, if and only if*

$$\frac{\pi_0 \tau^\star}{G(\tau^\star)} \leq \alpha \,.$$

REMARK 2.3.4 (Form of $\dot{\mathcal{T}}_G$). The expression of $\dot{\mathcal{T}}_G$ for threshold functions is given by Corollary 2.7.12, which shows that for one-stage adaptive procedures (where the level function $\mathcal{A}$ is constant), $\dot{\mathcal{T}}_G$ is proportional to the inverse of the difference between the slopes of r_α and G at $\tau^\star$. For two-stage plug-in procedures, which typically estimate π_0 using $G(u_0)$ for some u_0 (e.g. $u_0 = \lambda$ for procedure Sto02), $\dot{\mathcal{T}}_G$ involves an additional term that depends on $G(u_0)$, and the asymptotic distribution of the FDP depends on $\mathbb{Z}(u_0)$, where $\mathbb{Z}$ is defined in Theorem 2.3.1.

2.3.2. Asymptotically equivalent procedures. Some multiple testing procedures cannot be written in terms of threshold functions, because they do not depend exclusively on $\hat{\mathbb{G}}_m$, but instead also directly depend on the number m of observations. When such procedures are only slight perturbations of actual threshold procedures, they share the same asymptotic distribution, as explained below.

DEFINITION 2.3.5 (Asymptotic equivalence of multiple testing procedures). *Let $\mathcal{T}$ be a threshold function for which Condition C.1 holds for $\mathcal{T}$. A multiple testing procedure $\mathcal{M}$ is asymptotically equivalent to $\mathcal{T}$ as $m \to +\infty$ if and only if*

$$\sqrt{m}\left(\mathsf{FDP}_m\left(\mathcal{M}(P_1, \dots P_m)\right) - \mathsf{FDP}_m\left(\mathcal{T}(\hat{\mathbb{G}}_m)\right)\right) \xrightarrow{P} 0 \,.$$

PROPOSITION 2.3.6 (Asymptotic equivalence of thresholding procedures). *Let $\mathcal{T}$ be a threshold function, and $\varepsilon = (\varepsilon_m)_{m\in\mathbb{N}}$ a positive sequence. For $m \in \mathbb{N}$, let $\mathcal{T}_m : D[0,1] \to [0,1]$ such that*

$$\forall F \in D[0,1], \mathcal{T}\left(F - \varepsilon_m\right) \leq \mathcal{T}_m(F) \leq \mathcal{T}(F) \,. \tag{2.3.7}$$

If Condition C.1 holds for $\mathcal{T}$, and if $\varepsilon_m = \mathrm{o}\left(\frac{1}{\sqrt{m}}\right)$, $\mathcal{T}_m$ is asymptotically equivalent to $\mathcal{T}$ as $m \to +\infty$.

Several applications of Proposition 2.3.6 are given in section 2.4. For example, the asymptotic behavior of procedure $\mathcal{T}_m = \mathsf{STS04}(\lambda)$ can be derived from that of procedure $\mathcal{T} = \mathsf{Sto02}(\lambda)$, for which Theorem 2.3.2 may be used because $\mathsf{Sto02}(\lambda)$ is an actual threshold function.

2.3.3. Regularity conditions. For the threshold functions under investigation, $\mathcal{T}(G)$ is defined as the last point for which $G \geq r(\mathcal{A}(G), \cdot)$. Therefore, the existence of a *unique interior right crossing point* between G and $r(\mathcal{A}(G), \cdot)$ ensures that Theorem 2.3.2 and Proposition 2.3.6 are applicable, i.e. that $\mathcal{T}(G) > 0$, and that $\mathcal{T}$ is Hadamard differentiable at G (Condition C.1). For two-stage adaptive (plug-in) procedures, for which the level function $\mathcal{A}$ is not constant, additional technical assumptions concerning the regularity of $\mathcal{A}$ require checking (see Corollary 2.7.12) to ensure that Condition C.1 holds.

DEFINITION 2.3.8 (Right crossing point). *Let r_α be a rejection curve, and $\mathcal{A}$ a level function. Denote by $\mathcal{T} : F \mapsto \mathcal{U}(F, \mathcal{A}(F))$ the associated threshold function, where $\mathcal{U}(F,\alpha) = \sup\{u \in [0,1], F(u) \geq r_\alpha(u)\}$. A right crossing point for the multiple comparison problem defined by $\mathcal{T}$ (or, in short, a right crossing point for $\mathcal{T}$), is a point $t \in [0,1]$ such that $G(t) = r_\alpha(t)$, and $g(t) < \frac{\partial r}{\partial u}(\mathcal{A}(G), t)$. If t belongs to the open interval $(0,1)$ it is called an interior right crossing point for $\mathcal{T}$.*

Condition $g(t) < \frac{\partial r}{\partial u}(\mathcal{A}(G), t)$ in Definition 2.3.8 ensures that G and $r_{\mathcal{A}(G)} = r(\mathcal{A}(G), \cdot)$ actually *cross* at t, *i.e.* that $G \geq r_{\mathcal{A}(G)}$ in a left-neighborhood of t, and that $G \leq r_{\mathcal{A}(G)}$ in a right-neighborhood of t.

Studies of the asymptotic distribution of the abovementioned FDR controlling procedures require investigation, in each case, of the conditions guaranteeing the existence of a unique interior right crossing point. To this end, we broke this condition down as follows:

CONDITION C.2 (Existence). *$\mathcal{T}$ has an interior right crossing point.*

CONDITION C.3 (Uniqueness). *$\mathcal{T}$ has at most one interior right crossing point.*

Condition C.3 always holds for procedures based on Simes' line (BH95, Sto02, and BKY06) because their rejection curve is linear, and G is concave. Condition C.2 typically holds in situations in which the slope of G at the origin is large enough. In the case of the BH95 procedure, Chi recently showed the existence of a *critical value* $\alpha^\star$ depending solely on the distribution function G of the p-values, such that if $\alpha < \alpha^\star$, the number of discoveries made by the BH95 procedure is stochastically bounded as the number of tested hypotheses increases, whereas if $\alpha > \alpha^\star$, the proportion of discoveries converges in probability to a positive value $\tau^\star = \mathcal{T}(G)$ [16].

In section 2.4, we provide a detailed analysis of a number of FDR controlling procedures, and present, for each, a *critical value* for the target FDR level characterising situations in which condition C.2 is guaranteed for the procedure.

2.4. Results for procedures of interest

We apply the results of the preceding section to a series of procedures with proven (asymptotic) FDR control. Starting from the original BH95 procedure and its Oracle version (section 2.4.1), we then turn to adaptive procedures, which implicitly or explicitly incorporate an estimate of the proportion π_0 of true null hypotheses: *one-stage adaptive procedures* are studied in section 2.4.2, and *two-stage adaptive procedures* (also called plug-in procedures) are studied in section 2.4.3.

2.4.1. BH95 procedure. We will first recall the definition of the BH95 procedure in our framework.

DEFINITION 2.4.1 (Procedure BH95[7]). *The BH95 procedure is the multiple testing procedure with threshold function*

$$\mathcal{T}^{\mathsf{BH95}}(F) = \sup\{u \in [0,1], F(u) \geq u/\alpha\}\,.$$

As the rejection curve of procedure BH95 is linear, and G is concave, the uniqueness Condition C.3 always holds, and the existence Condition C.2 can be reduced to $\alpha > \alpha^\star$, where $\alpha^\star = \inf_{u\to 0} u/G(u) = \lim_{u\to 0} 1/g(u)$ corresponds to the *critical value of the* BH95 *procedure* [16]:

CONDITION C.4 (Condition C.2 for the BH95 procedure). *The target* FDR *level α is greater than the critical value $\alpha^\star$ of the* BH95 *procedure.*

The criticality phenomenon is illustrated in Figure 1 for Laplace (double exponential) test statistics. The Weak Law of Large Numbers phenomenon

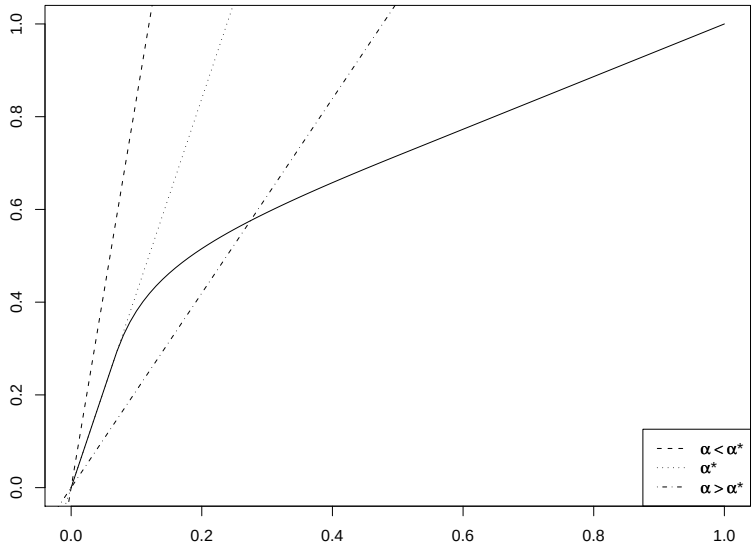


FIGURE 1. *Critical value of the* BH95 *procedure for Laplace test statistics with location parameter $\theta = 2$, and $\pi_0 = 0.5$. Solid line: distribution function G; straight lines: Simes' rejection curves for several values of α. There is an interior right crossing point between the distribution function of the p values and the Simes' line if and only if $\alpha > \alpha^\star = 1/g(0)$.*

analyzed by [16], which occurs when $\alpha > \alpha^\star$, was noted by [34]. We now derive the corresponding central limit theorem under the same hypothesis, and the asymptotic distribution of the FDP actually achieved by the BH95 procedure.

THEOREM 2.4.2 (Asymptotic properties of the BH95 procedure). *Let $\tau^\star = \mathcal{T}^{\mathsf{BH95}}(G)$. Under Condition C.4,*

(i)
$$\sqrt{m}\left(\mathcal{T}^{\mathsf{BH95}}(\hat{\mathbb{G}}_m) - \tau^\star\right) \rightsquigarrow \frac{\mathbb{Z}(\tau^\star)}{1/\alpha - g(\tau^\star)}$$
with $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ and $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are independent Gaussian processes such that $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ and $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, where $\mathbb{B}$ is a standard Brownian bridge on $[0,1]$.

(ii)

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{BH95}}(\hat{\mathbb{G}}_m)) - \pi_0\alpha\right) \rightsquigarrow \mathcal{N}\left(0, (\pi_0\alpha)^2\frac{1-\tau^\star}{\tau^\star}\right)$$

Applying the BH95 procedure at level α/π_0 leads to an Oracle procedure (as π_0 is not known) that is more powerful as it controls FDR at level exactly α. This procedure has threshold function

$$\mathcal{T}^{\mathsf{BH95o}}(F) = \sup\{u \in [0,1], F(u) \geq \pi_0 u/\alpha\},$$

and its critical value is therefore $\pi_0\alpha^\star$, which translates into the following regularity condition:

Condition C.5 (Condition C.2 for the BH95 Oracle procedure). *The target* FDR *level α is greater than $\pi_0\alpha^\star$, where $\alpha^\star$ is the critical value of the* BH95 *procedure.*

The corresponding asymptotic properties can be derived from Theorem 2.4.2:

Corollary 2.4.3 (Asymptotic properties of the BH95 Oracle procedure). *Let $\tau^\star = \mathcal{T}^{\mathsf{BH95o}}(G)$. Under Condition C.5,*

(i)

$$\sqrt{m}\left(\mathcal{T}^{\mathsf{BH95o}}(\hat{\mathbb{G}}_m) - \tau^\star\right) \rightsquigarrow \frac{\mathbb{Z}(\tau^\star)}{\pi_0/\alpha - g(\tau^\star)},$$

where $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\mathbb{Z}_1$ and $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are independent Gaussian processes such that $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ and $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, where $\mathbb{B}$ is a standard Brownian bridge on $[0,1]$.

(ii)

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{BH95o}}(\hat{\mathbb{G}}_m)) - \alpha\right) \rightsquigarrow \mathcal{N}\left(0, \alpha^2\frac{1-\tau^\star}{\tau^\star}\right).$$

2.4.2. One-stage adaptive procedures. The first class of adaptive procedures studied here are *one-stage* adaptive procedures, because they estimate π_0 implicitly, rather than through a level function $\mathcal{A}$.

Definition 2.4.4 (Adaptive procedure). *Let $r_\alpha : [0,1] \to [0,1]$. The adaptive procedure associated with r_α is the multiple testing procedure defined by the threshold function*

$$\mathcal{T}(F) = \sup\{u \in [0,1], F(u) \geq r_\alpha(u)\}.$$

The rejection curve of adaptive procedures is not linear, so the conditions under which Condition C.1 is fulfilled are more subtle than for the BH95 procedure (section 2.4.1) or for two-stage adaptive procedures (section 2.4.3).

Procedure FDR08(λ)**.** The rejection curve of the FDR08 procedure [30] is defined for $u \in [0,1]$ by $f_\alpha(u) = \frac{u}{\alpha+(1-\alpha)u}$. As $f_\alpha(1) = 1$, the corresponding threshold function is always equal to 1. This procedure therefore systematically rejects *all* hypotheses, and does not control FDR either for finite sample size or asymptotically. Several ways of overcoming this problem have been proposed [30], including *truncating* the rejection curve, yielding the following procedure:

DEFINITION 2.4.5 (Procedure $\mathsf{FDR08}(\lambda)$). *Let $\lambda \in [0,1)$. The rejection curve of the $\mathsf{FDR08}(\lambda)$ procedure is defined by $f_\alpha^\lambda(u) = f_\alpha(u)$ for $u \leq \lambda$, and $+\infty$ otherwise. The threshold function of the $\mathsf{FDR08}(\lambda)$ procedure is therefore given by*

$$\mathcal{T}^{\mathsf{FDR08}}(F) = \sup\left\{u \in [0,\lambda], F(u) \geq \frac{u}{\alpha + (1-\alpha)u}\right\}.$$

We introduce the following regularity condition:

CONDITION C.6. *$\lambda \geq \kappa$, where $\kappa = \frac{\alpha(1-\pi_0)}{(1-\alpha)\pi_0}$.*

Note that $\left(\kappa, \frac{1-\pi_0}{1-\alpha}\right)$ is the crossing point between the rejection curve f_α and the distribution function $\mathrm{DU}(\pi_0)$ in the extremal Dirac-Uniform configuration where all p-values drawn from $\mathcal{H}_1$ are equal to 0. As $G \leq \mathrm{DU}(\pi_0)$, condition C.6 ensures that any interior right crossing point between G and f_α occurs before λ. In practice, κ is unknown because it depends on π_0. However, an upper bound for κ can be deduced from a lower bound for π_0; for example, in microarray data analysis, it can often be assumed that $\pi_0 > \frac{1}{2}$: in this case, κ is smaller than $\frac{\alpha}{1-\alpha}$.

By definition 2.4.5, the rejection curve f_α^λ of any procedure $\mathsf{FDR08}(\lambda)$ satisfying Condition C.6 is equal to f_α on $[0,\kappa]$, corresponding to the admissible region for interior right crossing points. The following Proposition is a straightforward consequence of this observation:

PROPOSITION 2.4.6. *All $\mathsf{FDR08}(\lambda)$ procedures satisfying Condition C.6 are asymptotically equivalent in the sense of Definition 2.3.5.*

As the corresponding asymptotic distribution does not depend on λ, we will refer to it simply as the "asymptotic distribution of the $\mathsf{FDR08}$ procedure". In order to characterize this distribution we introduce a further technical condition to ensure that $\kappa < 1$. Combined with Condition C.4, it also ensures that existence Condition C.2 holds for procedure $\mathsf{FDR08}(\lambda)$, because the slope of f_α^λ at the origin is $1/\alpha$.

CONDITION C.7. *$\alpha < \pi_0$.*

Condition C.7 is a mild assumption in practice, because π_0 is typically expected to be greater than $1/2$, in microarray data analysis, for example. When $\alpha \geq \pi_0$, there is no need for sophisticated FDR controlling procedures because rejecting all hypotheses yields $\mathsf{FDP} = \pi_0$ and thus $\mathsf{FDR} \leq \alpha$.

THEOREM 2.4.7 (Asymptotic behavior of procedure $\mathsf{FDR08}$). *Let $\lambda \in [0,1)$ such that Condition C.6 is fulfilled, and $\tau^\star = \sup\left\{u \in [0,\kappa], G(u) \geq \frac{u}{\alpha+(1-\alpha)u}\right\}$. Under uniqueness Condition C.3, and existence Conditions C.4 and C.7, we have*

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{FDR08}(\lambda)}(\hat{\mathbb{G}}_m)) - \alpha\frac{\pi_0}{\overline{\pi_0}(\tau^\star)}\right) \rightsquigarrow X^{\mathsf{FDR08}},$$

with $\overline{\pi_0}(\tau^\star) = \frac{1-G(\tau^\star)}{1-\tau^\star}$, and

$$X^{\mathsf{FDR08}} = \mathsf{p}^\star(1-\mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)},$$

where $\mathsf{p}^\star = \alpha\pi_0/\overline{\pi_0}(\tau^\star)$ *is the* pFDR *achieved by procedure* FDR08*,*

$$\zeta(\tau^\star) = -\frac{(1-\overline{\pi_0}(\tau^\star))\overline{\pi_0}(\tau^\star)/\alpha}{\overline{\pi_0}(\tau^\star)^2/\alpha - g(\tau^\star)},$$

and $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$*, where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$*.*

As $\overline{\pi_0}(\tau^\star) = \frac{1-\tau^\star}{1-G(\tau^\star)} \in [\pi_0, 1]$, we have $\pi_0\alpha \le \mathsf{p}^\star \le \alpha$, so that procedure FDR08 is asymptotically more powerful than procedure BH95, and less powerful than procedure BH95o.

Procedure BR08(λ)**.**

Definition 2.4.8 (Procedure BR08(λ) [10]). *Let* $\lambda \in [0,1)$*. The rejection curve of the* BR08(λ) *procedure is defined by* $b_\alpha^\lambda(u) = \frac{u}{\alpha(1-\lambda)+u}$ *for* $u \le \lambda$*, and* $+\infty$ *otherwise. The threshold function of the* BR08(λ) *procedure is therefore given by*

$$\mathcal{T}^{\mathsf{BR08}(\lambda)}(F) = \sup\left\{u \in [0,\lambda], F(u) \ge \frac{u}{\alpha(1-\lambda)+u}\right\}.$$

Procedure BR08(λ) is actually defined by the rejection curve $\left(1+\frac{1}{m}\right) b_\alpha^\lambda$ [10]. However these procedures are asymptotically equivalent according to Proposition 2.3.6; we will therefore use Definition 2.4.8.

As for the FDR08 procedure, the rejection curve of the BR08(λ) procedure is not linear and we therefore need to make two assumptions to ensure that existence Condition C.2 holds: Condition C.8 ensures that there is no criticality phenomenon, that is, that the slope of the distribution function G is great enough at the origin, and Condition C.9 ensures that a right crossing point occurs before λ, because the BR08(λ) procedure is truncated at λ:

Condition C.8. *The target* FDR *level* α *satisfies* $\alpha(1-\lambda) > \alpha^\star$*, where* $\alpha^\star$ *is the critical value of the* BH95 *procedure.*

Condition C.9. *The distribution function* G *satisfies*

$$G(\lambda) \le \frac{\lambda}{\alpha}\frac{1-G(\lambda)}{1-\lambda}.$$

Remark 2.4.9. Condition C.9 may be written as $G(\lambda) \le b_\alpha^\lambda$, or as $G(\lambda) \le f_\alpha^\lambda$, because the rejection curves of procedures BR08(λ) and FDR08 intersect at λ.

Theorem 2.4.10 (Asymptotic distribution of procedure BR08(λ)). *Let* $\lambda \in [0,1)$ *and* $\tau^\star = \mathcal{T}^{\mathsf{BR08}(\lambda)}(G)$*. Under uniqueness Conditions C.3 and existence Conditions C.8 and C.9, we have*

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{BR08}(\lambda)}(\hat{\mathbb{G}}_m)) - \alpha\pi_0\frac{1-\lambda}{1-G(\tau^\star)}\right) \rightsquigarrow X^{\mathsf{BR08}(\lambda)},$$

with

$$X^{\mathsf{BR08}(\lambda)} = \mathsf{p}^\star(1-\mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)},$$

where $\mathsf{p}^\star = \alpha\pi_0 \frac{1-\lambda}{1-G(\tau^\star)}$ *is the* pFDR *achieved by procedure* BR08(λ),

$$\zeta(\tau^\star) = -\frac{G(\tau^\star)^2/\tau^\star}{G(\tau^\star)(1-G(\tau^\star))/\tau^\star - g(\tau^\star)}\,,$$

and $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$*, where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$.

Theorem 2.4.10 implies that procedure BR08(λ) controls FDR asymptotically at level α: as $\tau^\star \le \lambda$, we have $\mathsf{p}^\star \le \alpha\pi_0 \frac{1-\lambda}{1-G(\lambda)}$, which is smaller than α because $\overline{\pi_0}(\lambda) = \frac{1-G(\lambda)}{1-\lambda}$ is an upper bound for π_0.

However, as $b_\alpha^\lambda(u) \ge u/\alpha$ if and only if $u \ge \lambda\alpha$, procedure BR08(λ) need not be more powerful than procedure BH95, and we have the following characterization:

$$\mathsf{BR08}(\lambda) \gg \mathsf{BH95} \iff \tau^\star_{\mathsf{BH95}} \ge \alpha\lambda\,,$$

where $\gg$ means "is more powerful than", and $\tau^\star_{\mathsf{BH95}}$ is the asymptotic threshold of procedure BH95. An explicit characterization of situations in which BR08(λ) $\gg$ BH95 for Gaussian test statistics is given in [10].

2.4.3. Two-stage adaptive (plug-in) procedures. In this section we study two-stage adaptive or *plug-in* procedures, in which a conservative step-up procedure is applied to a data-dependent level. In particular, we consider the case of *Simes' line-based* plug-in procedures, in which procedure BH95 is applied at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is estimated from the data:

Definition 2.4.11 (Simes' line-based plug-in procedure). *Let* $\mathcal{A} : D[0,1] \to \mathbb{R}_+^*$. *The Simes' line-based plug-in procedure associated with* $\mathcal{A}$ *is the multiple testing procedure defined by the threshold function*

$$\mathcal{T}(F) = \sup\left\{u \in [0,1], F(u) \ge \frac{u}{\mathcal{A}(F)}\right\}.$$

Such procedures will simply be called plug-in procedures *hereafter.*

As r_α is linear, and G is concave, uniqueness Condition C.3 always holds for plug-in procedures, and existence Condition C.2 is the same as for procedure BH95, except that α is replaced by the value of the level function $\mathcal{A}$ at G:

Condition C.10 (Condition C.2 for plug-in procedures). *The level function* $\mathcal{A}(G)$ *associated with the target* FDR *level* α *is greater than the critical value of the* BH95 *procedure.*

Care is required when deriving the asymptotic distribution of the FDP for plug-in procedures, because the Hadamard derivative of $\mathcal{T}$, $\dot{\mathcal{T}}_G(H)$, typically involves the value of H at $\tau^\star$ and at a point $u(\lambda)$ used for the estimation of π_0: $u(\lambda) = \lambda$ for procedure Sto02, and $u(\lambda) = \mathcal{U}(G,\lambda)$ for procedure BKY06(λ). The asymptotic variance of the False Discovery Proportion therefore involves the covariance between $\mathbb{Z}(\tau^\star)$ and $\mathbb{Z}(u(\lambda))$.

Theorem 2.4.12 (Asymptotic FDP for procedures based on Simes' line). *Let* $\mathcal{T} : F \mapsto \mathcal{U}(F, \mathcal{A}(F))$ *a threshold function based on Simes' line. If the*

level function $\mathcal{A}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, and satisfies existence Condition C.10, then

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \pi_0\mathcal{A}(G)\right) \rightsquigarrow \pi_0\mathcal{A}(G)\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \frac{\dot{\mathcal{A}}_G(\mathbb{Z})}{\mathcal{A}(G)}\right),$$

with $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ and $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are independent Gaussian processes such that $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ and $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, where $\mathbb{B}$ is a standard Brownian bridge on $[0,1]$.

We consider the two types of plug-in procedures most widely used and theoretically justified: Sto02-like procedures (Sto02 [89], STS04 [93]), in which π_0 is estimated by $\frac{1-\hat{\mathbb{G}}_m(\lambda)}{1-\lambda}$ or a slight variant, and the BKY06 procedure [9], in which an upper bound for π_0 is derived from a first application of the classical BH95 procedure.

Procedure Sto02.

Definition 2.4.13 (Procedure Sto02 [89]). *Procedure Sto02 is the multiple testing procedure with threshold function*

$$\mathcal{T}^{\mathsf{Sto02}(\lambda)}(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\alpha}\frac{1-F(\lambda)}{1-\lambda}\right\}.$$

The level function of this procedure is therefore

$$\mathcal{A}(F) = \frac{\alpha}{\overline{\pi_0}^F(\lambda)},$$

with

$$\overline{\pi_0}^F(\lambda) = \frac{1-F(\lambda)}{1-\lambda}.$$

$\overline{\pi_0}^G(\lambda)$ will simply be denoted by $\overline{\pi_0}(\lambda)$.

Condition C.11 (Condition C.2 for procedure Sto02(λ)). *The target FDR level α is greater than $\overline{\pi_0}(\lambda)\,\alpha^\star$, where $\alpha^\star$ is the critical value of the BH95 procedure.*

This procedure is known to provide asymptotic control of FDR at level α [89], but does not necessarily control FDR at level α for finite sample size. This led to the definition of a modification of the Sto02 procedure that does control FDR even for finite sample size [93]:

Definition 2.4.14 (Procedure STS04(λ) [93]). *Procedure STS04(λ) rejects p-values smaller than*

$$\mathcal{T}_m^{\mathsf{STS04}(\lambda)}(\hat{\mathbb{G}}_m) = \sup\left\{u \in [0,\lambda], \hat{\mathbb{G}}_m(u) \geq \frac{u}{\alpha}\frac{1+\frac{1}{m}-\hat{\mathbb{G}}_m(\lambda)}{1-\lambda}\right\}.$$

According to Proposition 2.3.6, procedures STS04(λ) and Sto02(λ) are asymptotically equivalent provided that Conditions C.9 and C.11 hold (see Proposition 2.7.16 page 60 for a formal proof).

THEOREM 2.4.15 (Asymptotic properties of the Sto02/STS04 procedure). *Let* $\lambda \in (0,1)$, *and* $\tau^\star = \mathcal{T}^{\mathsf{Sto02}(\lambda)}(G)$. *Under existence Condition C.11, we have*

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{Sto02}}(\hat{\mathbb{G}}_m)) - \frac{\pi_0}{\overline{\pi_0}(\lambda)}\alpha\right) \rightsquigarrow X^{\mathsf{Sto02}},$$

with

$$X^{\mathsf{Sto02}} = \frac{\pi_0\alpha}{\overline{\pi_0}(\lambda)}\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \frac{\mathbb{Z}(\lambda)}{1-G(\lambda)}\right),$$

where $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\mathbb{Z}_1$ *and* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, *where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$. X^{Sto02} *is therefore a centered Gaussian random variable, with variance*

$$\left(\frac{\pi_0\alpha}{\overline{\pi_0}(\lambda)}\right)^2\left\{\frac{1-\tau^\star}{\tau^\star} + \frac{\operatorname{Var}\mathbb{Z}(\lambda)}{(1-G(\lambda))^2} + \pi_0\frac{\tau^\star_\wedge\lambda\,(1-\tau^\star\vee\lambda)}{\tau^\star(1-G(\lambda))}\right\},$$

where

$$\operatorname{Var}\mathbb{Z}(\lambda) = \pi_0^2\lambda(1-\lambda) + (1-\pi_0)^2\,G_1(\lambda)(1-G_1(\lambda))\,.$$

COROLLARY 2.4.16. *If* $\tau^\star \le \lambda$,

$$\operatorname{Var} X^{\mathsf{Sto02}}(\lambda) = \left(\frac{\pi_0\alpha}{\overline{\pi_0}(\lambda)}\right)^2\left\{\frac{1-\tau^\star}{\tau^\star} + \frac{\operatorname{Var}\mathbb{Z}(\lambda)}{(1-G(\lambda))^2} + \frac{\pi_0}{\overline{\pi_0}(\lambda)}\right\}.$$

As $\overline{\pi_0}(\lambda) = \pi_0 + (1-\pi_0)\frac{1-G_1(\lambda)}{1-\lambda}$, we have, for any $\lambda \le \lambda'$, $\pi_0 \le \overline{\pi_0}(\lambda') \le \overline{\pi_0}(\lambda) \le 1$, $\mathsf{BH95o} \gg \mathsf{Sto02}(\lambda') \gg \mathsf{Sto02}(\lambda) \gg \mathsf{BH95}$.

Procedure BKY06**.** Letting $\beta = \frac{\alpha}{1+\alpha}$, procedure BKY06 involves applying procedure BH95 at level $\frac{\beta}{1-R(\beta)/m}$, where $R(\beta)$ is the number of hypotheses rejected by a first application of the BH95 procedure at level β. We shall consider a recently proposed generalization of this procedure [10], in which procedure BH95 is applied at level $\frac{1-\lambda}{1-R(\lambda)/m}\alpha$. The original BKY06 procedure corresponds to $\lambda = \frac{\alpha}{1+\alpha}$.

DEFINITION 2.4.17 (Procedure $\mathsf{BKY06}(\lambda)$[9]). *Let* $\lambda \in [0,1)$, *and*

$$\mathcal{A}(F) = \alpha\frac{1-\lambda}{1-F\left(\mathcal{U}(F,\lambda)\right)},$$

where

$$\mathcal{U}(F,\lambda) = \sup\left\{u \in [0,1], F(u) \ge \frac{u}{\lambda}\right\}.$$

The threshold function of procedure $\mathsf{BKY06}(\lambda)$*is defined for any* $F \in D[0,1]$ *by* $\mathcal{T}^{\mathsf{BKY06}(\lambda)}(F) = \mathcal{U}(F,\mathcal{A}(F))$, *that is,*

$$\mathcal{T}^{\mathsf{BKY06}(\lambda)}(F) = \sup\left\{u \in [0,1], F(u) \ge \frac{u}{\alpha}\;\frac{1-F(\mathcal{U}(F,\lambda))}{1-\lambda}\right\}.$$

REMARK 2.4.18. As the proportion $R(\lambda)/m$ of hypotheses rejected by procedure BH95 at level λ equals $\hat{\mathbb{G}}_m(\mathcal{U}(\hat{\mathbb{G}}_m,\lambda))$ (see Proposition 2.7.8 page 54), $\mathcal{A}(\hat{\mathbb{G}}_m)$ may be written as $\alpha\frac{1-\lambda}{1-R(\lambda)/m}$.

REMARK 2.4.19. The exact definition of procedure BKY06(λ) adds $1/m$ to the denominator of the level function:

$$\mathcal{A}(F) = \alpha \frac{1-\lambda}{1+\frac{1}{m}-F\left(\mathcal{U}(F,\lambda)\right)},$$

which permits proving that this procedure controls FDR for finite sample size [10]. According to Proposition 2.3.6, these two procedures are asymptotically equivalent, so we will use Definition 2.4.17.

As procedure BKY06(λ) is based on two successive applications of procedure BH95, at level λ and $\alpha(1-\lambda)$, Condition C.2 holds if and only if Condition C.8 holds and $\lambda > \alpha^\star$.

CONDITION C.12. *The parameter λ satisfies $\lambda > \alpha^\star$, where $\alpha^\star$ is the critical value of the* BH95 *procedure.*

THEOREM 2.4.20 (Asymptotic properties of the BKY06(λ) procedure). *Let $\alpha \in [0,1]$, and $\lambda \in [0,1)$. Let $u(\lambda) = \mathcal{U}(G,\lambda)$ be the asymptotic threshold of the* BH95 *procedure applied at level λ, and $\tau^\star = \mathcal{T}^{\mathsf{BKY06}(\lambda)}(G)$. Under existence Conditions C.8 and C.12,*

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{BKY06}(\lambda)}(\hat{\mathbb{G}}_m)) - \frac{\pi_0\alpha(1-\lambda)}{1-G(u(\lambda))}\right) \rightsquigarrow X^{\mathsf{BKY06}(\lambda)},$$

with

$$X^{\mathsf{BKY06}(\lambda)} = \frac{\pi_0\alpha(1-\lambda)}{1-G(u(\lambda))}\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \frac{1}{1-\alpha(1-\lambda)g(u(\lambda))}\frac{\mathbb{Z}(u(\lambda))}{1-G(u(\lambda))}\right),$$

where $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\mathbb{Z}_1$ and $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are independent Gaussian processes such that $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ and $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, where $\mathbb{B}$ is a standard Brownian bridge on $[0,1]$.

As $u(\lambda)$ is the asymptotic threshold of the BH95 procedure applied at level λ, we have $G(u(\lambda)) = u(\lambda)/\lambda$, $u(\lambda) \le \lambda$. Therefore, $\frac{1-\lambda}{1-G(u(\lambda))} \le \frac{1-\lambda}{1-G(\lambda)}$, and the asymptotic level of procedure BKY06(λ) is less than α because $\frac{1-G(\lambda)}{1-\lambda} \ge \pi_0$.

However, as for procedure BR08(λ), procedure BKY06(λ) need not be more powerful than BH95: a comparison of the asymptotic FDR for these two procedures shows that situations in which BKY06(λ) $\gg$ BH95 are characterized by $G(u_\lambda) \ge \lambda$, that is, $G^2(u_\lambda) \ge u_\lambda$ because $G(u_\lambda) = u_\lambda/\lambda$. Hence, BH95 $\gg$ BKY06(λ) corresponds to situations in which G is too close to the Uniform distribution. For example, if $G(x) \le \sqrt{x}$ for all $x \in [0,1]$, then for any $\lambda \in [0,1]$, BH95 $\gg$ BKY06(λ).

2.5. Connection between one- and two-stage adaptive procedures

We have introduced two types of FDR controlling procedures generalizing the BH95 procedure: two-stage adaptive (plug-in) procedures explicitly incorporate an estimate of π_0 into the standard BH95 procedure, whereas one-stage adaptive procedures do not explicitly use such an estimate, but still yield tighter FDR control than the BH95 procedure.

We will now investigate connections between one-stage and two-stage adaptive procedures, which naturally appear when using the formalism of

threshold functions: with a striking symmetry, the threshold of procedure BR08(λ) may be interpreted as a fixed point of an iterated BKY06(λ) procedure, whereas the threshold of procedure FDR08 may be interpreted as a fixed point of an iterated Sto02(λ) procedure. We provide heuristic reasons for these connections in section 2.5.1; in section 2.5.2 we present general results for the connection between one-stage and two-stage adaptive procedures, and derive consequences for the connection between procedures Sto02(λ) and FDR08 on the one hand, and between procedures BKY06(λ) and BR08(λ) on the other hand.

2.5.1. Heuristics.

Procedures BKY06(λ) and BR08(λ). The BKY06 procedure was designed to derive an approximate upper bound for π_0 from a first application of procedure BH95, and to use this upper bound in a second application of the BH95 procedure, leading to less conservative FDR control. For $\lambda \in [0,1)$, the threshold function of the BKY06(λ) procedure is defined by

$$\mathcal{T}^{\mathsf{BKY06}(\lambda)}(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\alpha}\frac{1-F(\mathcal{U}(F,\lambda))}{1-\lambda}\right\},$$

where $\mathcal{U}(F,\lambda) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\lambda}\right\}$. It therefore seems natural to iterate this process, using the number of rejections at the second application to find a less conservative upper bound for π_0, and to use this new upper bound in a third application of the BH95 procedure, and so on. Based on this idea, Benjamini *et al.* suggested defining a *multi-stage procedure* for the particular situation in which $\lambda = \frac{\alpha}{1+\alpha}$ [9]. In our framework, this iterative process suggests the introduction of a *fixed-point procedure* defined for any $F \in D[0,1]$ by:

$$\mathcal{T}_\infty^{\mathsf{BKY06}(\lambda)}(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\alpha}\frac{1-F(u)}{1-\lambda}\right\}.$$

The term *fixed-point procedure* refers to the following property of the corresponding asymptotic threshold $\tau_\infty^\star = \mathcal{T}_\infty^{\mathsf{BKY06}(\lambda)}(G)$. Let us suppose that $\tau_\infty^\star$ is the threshold obtained at a given stage of the abovementioned iteration process. As $G(\tau_\infty^\star) = \tau_\infty^\star(1-G(\tau_\infty^\star))/\alpha(1-\lambda)$, $\tau_\infty^\star$ is also the asymptotic threshold at the next stage, and is thus a fixed point of the iteration process. It turns out that *this fixed-point procedure is the* BR08(λ) *procedure* investigated in section 2.4.2: $F(u) \geq \frac{u}{\alpha(1-\lambda)}(1-F(u))$ may be written as $F(u) \geq \frac{u}{\alpha(1-\lambda)+u}$, and the right-hand side is the rejection curve b_α^λ of the BR08(λ) procedure.

Procedures Sto02(λ) and FDR08(λ). The same idea may be adapted to procedure Sto02(λ), which is defined for $0 \leq \lambda < 1$ by the threshold function

$$\mathcal{T}^{\mathsf{Sto02}(\lambda)}(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\alpha}\frac{1-F(\lambda)}{1-\lambda}\right\}.$$

If $\widehat{\tau}_\lambda = \mathcal{T}^{\mathsf{Sto02}(\lambda)}(\hat{\mathbb{G}}_m)$ denotes the empirical threshold of procedure Sto02(λ), one may use $\widehat{\tau}_\lambda$ to estimate π_0, that is, calculate the threshold

given by procedure $\mathsf{Sto02}(\widehat{\tau}_\lambda)$, and so on. This suggests that an associated *fixed-point procedure* could be defined as

$$\mathcal{T}_\infty^{\mathsf{Sto02}}(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\alpha}\frac{1-F(u)}{1-u}\right\}.$$

Again, the term *fixed-point procedure* refers to the fact that if $\tau_\infty^\star = \mathcal{T}(G)$ is used as a new λ to estimate π_0 in procedure $\mathsf{Sto02}(\lambda)$, then the asymptotic threshold of procedure $\mathsf{Sto02}(\lambda)$ is also $\tau_\infty^\star$, which is therefore a fixed point of the iteration process. It turns out that *this fixed-point procedure is the* FDR08 *procedure* investigated in section 2.4.2: $F(u) \geq \frac{u}{\alpha}\frac{1-F(u)}{1-u}$ may be written as $F(u) \geq \frac{u}{\alpha+(1-\alpha)u}$, and the right-hand side is the rejection curve f_α of the FDR08 procedure.

2.5.2. Formal connections. We present a general result concerning connections between one-stage and two-stage adaptive procedures, providing a formal justification for the connections mentioned in section 2.5.1, and accounting for their symmetry. This result is based on the following assumption concerning the threshold function of the one-stage adaptive procedure:

Condition C.13. *There is a curve $c_\alpha : D[0,1] \times [0,1]$ such that the threshold function $\mathcal{T}$ may be written as*

$$\mathcal{T}(F) = \sup\{u \in [0,\lambda], F(u) \geq c_\alpha(F,u)\},$$

where $u \mapsto c_\alpha(G,u)/u$ is non increasing on $[0,\lambda]$.

Remark 2.5.1. In Condition C.13, $c_\alpha(F,\cdot)$ is *not* the rejection curve of procedure $\mathcal{T}$, because it depends on F. For example, for procedure FDR08, we will use

$$c_\alpha(F,u) = \frac{u}{\alpha}\,\frac{1-F(u)}{1-u}.$$

Theorem 2.5.2 shows that we can associate with a one-stage adaptive procedure fulfilling Condition C.13 a two-stage adaptive procedure with linear rejection curve, and level function given by

$$\mathcal{A}(F) = \frac{t}{c_\alpha(F,t)},$$

for fixed $t \in (0,1)$. The asymptotic threshold of the one-stage procedure may then be interpreted as the fixed point of iterations of the two-stage procedure.

Theorem 2.5.2 (Connection between one-stage and two-stage adaptive procedures). *Let $\lambda \in (0,1)$. Let us consider a multiple testing procedure with a threshold function $\mathcal{T}$ that may be written as*

$$\mathcal{T}(F) = \sup\{u \in [0,\lambda], F(u) \geq c_\alpha(F,u)\}$$

for any $F \in D[0,1]$. Let $\mathcal{T}_t$ be the threshold function defined by

$$\mathcal{T}_t(F) = \sup\left\{u \in [0,1], F(u) \geq \frac{c_\alpha(F,t)}{t}u\right\},$$

for any $t \in (0,1)$ and any $F \in D[0,1]$. Let us assume that existence Condition C.2 and uniqueness Condition C.3 hold for procedure $\mathcal{T}$, and that, for

any $t \in (0,1)$, existence Condition C.2 holds for procedure $\mathcal{T}_t$. Let $\tau^\star = \mathcal{T}(G)$ and $\tau(t) = \mathcal{T}_t(G)$ be the asymptotic thresholds of procedures $\mathcal{T}$ and $\mathcal{T}_t$, respectively. If c_α satisfies Condition C.13, we have

(i) *for any $t \in (0, \lambda]$,*

$$\begin{cases} t \leq \tau^\star \Rightarrow \tau(t) \in [t, \tau^\star] \\ t \geq \tau^\star \Rightarrow \tau(t) \in [\tau^\star, t] \end{cases} .$$

(ii) *Let $t \in (0, \lambda]$. Define the sequence $(t_n) \in [0,1]^{\mathbb{N}}$ by $t_0 = t$, and $t_{i+1} = \tau(t_i)$ for $i \in \mathbb{N}$. Then*

$$\lim_{n \to \infty} t_n = \tau^\star .$$

COROLLARY 2.5.3 (Asymptotic power comparison). *With the same notation and under the same conditions, the following assertions are equivalent:*

(i) *Procedure $\mathcal{T}_t$ is asymptotically more powerful than procedure $\mathcal{T}$*
(ii) $\tau(t) > \tau^\star$
(iii) $t > \tau^\star$
(iv) $t > \tau(t)$

In the remainder of this section, we use Theorem 2.5.2 to characterize the connection between the abovementioned procedures.

Procedures Sto02(λ) **and** FDR08(λ)**.** Theorem 2.5.4 gives the convergence of the process consisting of the recursive use of the asymptotic threshold of procedure Sto02(λ) as a new λ. It holds under the same regularity conditions as those required to obtain the asymptotic distribution of procedure FDR08.

THEOREM 2.5.4 (Connection between procedures Sto02(λ) and FDR08). *Let $\kappa = \frac{\alpha(1-\pi_0)}{\pi_0(1-\alpha)}$, and*

$$\tau^\star = \sup \left\{ u \in [0, \kappa], G(u) \geq \frac{u}{\alpha} \frac{1 - G(u)}{1 - u} \right\}$$

be the asymptotic threshold of the FDR08 *procedure. For $u \in [0,1]$, let*

$$\tau(u) = \sup \left\{ u \in [0,1], G(u) \geq \frac{u}{\alpha} \frac{1 - G(\lambda)}{1 - \lambda} \right\}$$

be the asymptotic threshold of procedure Sto02(u)*. For any $t \in (0,1)$, define the sequence $(t_n) \in [0,1]^{\mathbb{N}}$ by $t_0 = t$, and $t_{i+1} = \tau(t_i)$ for $i \in \mathbb{N}$. Let us assume that uniqueness Condition C.3 holds for procedure* FDR08*, and that the target* FDR *level α satisfies existence Conditions C.4 and C.7. Then,*

$$\lim_{n \to \infty} t_n = \tau^\star .$$

COROLLARY 2.5.5 (Asymptotic power comparison — Sto02(λ) vs FDR08). *With the same notation and under the same conditions, procedure* Sto02(λ) *is asymptotically more powerful than procedure* FDR08 *if and only if $\lambda > \tau(\lambda)$.*

When using procedure $\mathsf{Sto02}(\lambda)$ in practice, we would not want any of the rejected hypotheses to be incorporated into the estimation of π_0. Thus, the empirical rejection threshold $\mathcal{T}^{\mathsf{Sto02}(\lambda)}(\hat{\mathbb{G}}_m)$ should be less than λ. In such situations, as $\mathcal{T}^{\mathsf{Sto02}(\lambda)}(\hat{\mathbb{G}}_m)$ converges at rate $1/\sqrt{m}$ to $\tau(\lambda) = \mathcal{T}^{\mathsf{Sto02}(\lambda)}(G)$, procedure $\mathsf{Sto02}(\lambda)$ is probably more powerful than procedure $\mathsf{FDR08}$ according to Corollary 2.5.5.

Procedures $\mathsf{BKY06}(\lambda)$ and $\mathsf{BR08}(\lambda)$. Theorem 2.5.6 characterizes the connection between procedure $\mathsf{BKY06}(\lambda)$ and procedure $\mathsf{BR08}(\lambda)$. It holds under the same regularity conditions as those required to obtain the asymptotic distribution of procedure $\mathsf{BR08}(\lambda)$.

Let $\tau^\star = \mathcal{T}^{\mathsf{BR08}(\lambda)}(G)$ be the asymptotic threshold of the $\mathsf{BR08}(\lambda)$ procedure. Under uniqueness Condition C.3 and existence Conditions C.8, C.9 and C.12, (t_n) is non decreasing, and converges to $\tau^\star$.

THEOREM 2.5.6 (Connection between procedures $\mathsf{BKY06}(\lambda)$ and $\mathsf{BR08}(\lambda)$). *Let $\lambda \in (0,1)$. For $F \in D[0,1]$ and $\beta \in [0,1]$, let*

$$\mathcal{U}(F,\beta) = \sup\left\{u \in [0,1], F(u) \geq \frac{u}{\beta}\right\}.$$

For any $u \in [0,1)$, let $\tau(u) = \mathcal{U}\left(G, \frac{\alpha(1-\lambda)}{1-G(u)}\right)$. With this notation, $\mathcal{T}(G) = \tau(u(\lambda))$ is the asymptotic threshold of the $\mathsf{BKY06}(\lambda)$ procedure, where $u(\lambda) = \mathcal{U}(G,\lambda)$. Let

$$\tau^\star = \sup\left\{u \in [0,\lambda], G(u) \geq \frac{u}{\alpha(1-\lambda)+u}\right\}$$

be the asymptotic threshold of the $\mathsf{BR08}(\lambda)$ procedure. Define the sequence $(t_n) \in [0,1]^{\mathbb{N}}$ by $t_0 = u(\lambda)$, and $t_{i+1} = \tau(t_i)$ for $i \in \mathbb{N}$. Let us assume that uniqueness Condition C.3 holds for procedure $\mathsf{BR08}(\lambda)$, and that the target FDR level α satisfies existence Conditions C.8 and C.9. Then

$$\lim_{n\to\infty} t_n = \tau^\star.$$

COROLLARY 2.5.7 (Asymptotic power comparison — $\mathsf{BKY06}(\lambda)$ vs $\mathsf{BR08}(\lambda)$). *With the same notation and under the same conditions, procedure $\mathsf{BR08}(\lambda)$ is asymptotically more powerful than procedure $\mathsf{BKY06}(\lambda)$ if and only if the asymptotic threshold $\tau^\star$ of procedure $\mathsf{BR08}(\lambda)$ satisfies $\tau^\star \geq \lambda - \alpha(1-\lambda)$.*

For example, setting λ to a value less than $\frac{\alpha}{1+\alpha}$, corresponding to the original $\mathsf{BKY06}$ procedure [9], ensures that the associated $\mathsf{BR08}(\lambda)$ procedure is asymptotically more powerful than the associated $\mathsf{BKY06}(\lambda)$ procedure.

2.6. Concluding remarks

We have demonstrated the power and flexibility of the formalism of threshold functions, making it possible to derive the asymptotic properties of well known FDR controlling procedures with their associated regularity conditions, and to identify and characterize novel connections between one-stage and two-stage adaptive procedures. These results are summarized in

Table 1. We should recall that the threshold function associated with the level function $\mathcal{A}$ and rejection curve $r_\alpha = r(\alpha, \cdot)$ is defined by

$$\mathcal{T}(F) = \sup\{u \in [0,1], F(u) \geq r(\mathcal{A}(F), u)\} .$$

By definition, the level function $\mathcal{A}$ equals α for one-stage procedures, and the rejection curve of Simes' line-based procedures is $r_\alpha : u \mapsto u/\alpha$.

TABLE 1. Comparison of FDR controlling procedures, characterized by their level function $\mathcal{A}$ and their rejection curve r_α. Conditions for the existence and uniqueness of an interior right crossing point are recalled, together with the corresponding pFDR relative to that of the BH95 procedure: $\pi_0\alpha$.

Name	BH95 [7]	FDR08 [30]	BR08(λ) [10]	Sto02(λ) [89]	BKY06(λ) [9]
$\mathcal{A}(F)/\alpha$	1	1	1	$\frac{1-\lambda}{1-F(\lambda)}$	$\frac{1-\lambda}{1-\hat{\mathbb{G}}_m(u_\lambda)}$ $^{(a)}$
$r_\alpha(u)$	u/α	$\frac{u}{\alpha+(1-\alpha)u}$ $^{(b)}$	$\frac{u}{\alpha(1-\lambda)+u}$ $^{(c)}$	u/α	u/α
Existence	C.4	C.4 & C.7 $^{(d)}$	C.8 & C.9 $^{(d)}$	C.11	C.12
Uniqueness	—	C.3	C.3	—	—
pFDR/$\pi_0\alpha$	1	$\frac{1-\tau^\star_{\mathsf{FDR08}}}{1-G(\tau^\star_{\mathsf{FDR08}})}$	$\frac{1-\lambda}{1-G(\tau^\star_{\mathsf{BR08}})}$	$\frac{1-\lambda}{1-G(\lambda)}$	$\frac{1-\lambda}{1-G(u_\lambda)}$ $^{(a)}$

(a) : u_λ is the asymptotic threshold of the BH95 procedure at target level λ; (b) : truncated at $\frac{\alpha(1-\pi_0)}{\pi_0(1-\alpha)}$; (c) : truncated at λ; (d) : Sufficient (not necessary) conditions.

Regularity conditions. For one-stage adaptive procedures FDR08 and BR08(λ), the uniqueness Condition C.3 has to be assumed (cf. Table 1): as the rejection curve is not linear, the interior right crossing point is not necessarily unique; in practice the uniqueness condition holds except in pathological situations. For Simes' line-based procedures BH95, Sto02(λ) and BKY06(λ), existence Condition C.2 holds provided that the slope of the distribution function exceeds a certain threshold at the origin (that is, that there is no criticality phenomenon). For one-stage adaptive procedures, it is also required that the rejection curve r_α ends below the distribution function G, which corresponds to Condition C.7 for procedure FDR08, and Condition C.9 for procedure BR08(λ).

The criticality phenomenon studied by Chi [16] is intrinsic to the multiple testing problem, and not specific to a given procedure, as the minimum attainable pFDR level $\beta^\star = \inf_{t>0} \mathsf{pFDR}(t)$ depends solely on the parameters of the mixture model [16]. When $\beta^\star = 0$, say for the Gaussian location problem, *there is no criticality phenomenon for any procedure*: $\alpha^\star = 0$, and all existence Conditions concerning the behavior of the distribution function G close to 0 are fulfilled for any procedure, and for any target FDR level α. When $\beta^\star > 0$, say for the Laplace location problem (Figure 1, page 33), *there is a criticality phenomenon for every procedure*; however the critical value, that is, the minimum target FDR level for which existence Condition C.2 holds, may depend on the procedure, as illustrated by the existence conditions in Table 1.

Power comparisons. All procedures are asymptotically conservative, and therefore yield asymptotic FDR below the target level. Procedures

FDR08 and Sto02 (and thus STS04) are always more powerful than procedure BH95, but this is not the always the case for procedures BR08(λ)(section 2.4.2) and BKY06(λ) (section 2.4.3).

For one-stage adaptive procedures, for any $\lambda \in (0,1)$ such that the regularity conditions for procedures FDR08 and BR08(λ) hold, FDR08 is asymptotically more powerful than BR08(λ). Indeed, Condition C.9 ensures that the asymptotic thresholds of both procedures are less than λ. As the rejection curve f_α of procedure FDR08 is smaller than the rejection curve b_α^λ of BR08 on $[0,\lambda]$, the asymptotic threshold of procedure FDR08 is greater than that of procedure BR08(λ). However, it should be noted that procedure BR08(λ) does control FDR for a finite number of tested hypotheses, whereas procedure FDR08 does not.

For two-stage adaptive procedures, for any $\lambda \in (0,1)$ such that the regularity conditions for procedures Sto02(λ) and BKY06(λ) hold, Sto02(λ) (and thus STS04) is asymptotically more powerful than BKY06(λ), as demonstrated by the corresponding asymptotic FDR levels in Table 1: as $u_\lambda \leq \lambda$, we have $\frac{1-\lambda}{1-G(u_\lambda)} \leq \frac{1-\lambda}{1-G(\lambda)}$. This suggests that procedure STS04(λ) is preferable to procedure BKY06(λ) in practice. This recommendation should be balanced against the choice of λ and the desired robustness to dependence between null hypotheses. Based on a simulation study, procedure Sto02(α) was recently reported to be much more robust to positive dependence between null hypotheses than procedure Sto02(1/2) [10], which is still a standard choice in practical implementations, such as the SAM (Significance Analysis of Microarrays) software [92].

Towards optimality. This comparison raises the question of whether the formalism of threshold functions can be used to derive procedures more powerful than those studied here. One possible approach consists of trying to improve the estimation of π_0 to build a procedure closer to the Oracle BH95 procedure, as discussed in [35]. However, consistent estimators of π_0 have slower convergence rates than $1/\sqrt{m}$, resulting in slower convergence rates than $1/\sqrt{m}$ for the associated FDP. This may be illustrated by the influence of λ on procedure Sto02(λ): the larger λ, the smaller the bias $\mathbb{E}\left[\widehat{\pi_0}(\lambda)\right] - \pi_0$, and the larger the variance of $\widehat{\pi_0}(\lambda)$. The question of how to choose λ as a function of the number of hypotheses tested and the assumed regularity of G is addressed in another work [65].

Another possibility would be to consider procedures more general than those used in this paper: the BH95o procedure has been shown to give the lowest false non discovery rates (FNR) of the threshold procedures controlling FDR at level α [34]. The question of optimality in a broader family of testing procedures has recently been raised [96]: Z score-based threshold procedures may outperform p value-based threshold procedures, as Z score-based threshold procedures make it possible to choose different significance thresholds for positive and negative significance cutoffs. This suggests to extend our framework to Z score-based procedures.

Confidence intervals. An interesting practical application of this work concerns the derivation of asymptotic confidence intervals for the FDP of a given procedure. Our results give explicit asymptotic distributions for the

attained FDP, but this issue is not straightforward because these distributions depend on unknown quantities, including the proportion π_0, the asymptotically attained FDR $\tau^\star$, or the distribution function G and its associated density g. These quantities should, in turn, be estimated. Bootstrapping techniques could be used for this purpose; we leave this question for further research.

Extension to other dependence settings. We have derived the asymptotic properties of several multiple testing procedures and the associated regularity conditions in the situation in which p-values are independent. However, our formalism makes it possible to deal with any dependence situation for which the vector $(\hat{\mathbb{G}}_{0,m}, \hat{\mathbb{G}}_{1,m})$ of empirical distribution functions of the p-values under the null and alternative hypotheses satisfies Donsker's invariance principle. For example, the form of the asymptotic distributions of the threshold $\mathcal{T}(\hat{\mathbb{G}}_m)$ and the associated FDP would remain the same in the conditional dependence model recently proposed by Wu [109].

2.7. Proof of main results

2.7.1. Asymptotic FDP: general threshold functions. In this section, we provide proofs for the results of section 2.3.

Proof of Theorem 2.3.2. The following lemma will be used in several subsequent proofs.

LEMMA 2.7.1. *Let $H \in C[0,1]$, and H_t be a family of functions of $D[0,1]$ that converges to H on $(D[0,1], \|.\|_\infty)$ as $t \to 0$. For any sequence $(u_t)_{t>0}$ of $[0,1]$ that converges to $u \in [0,1]$ as $t \to 0$, we have*

$$\begin{aligned} \lim_{t\to 0} H_t(u_t) &= H(u) \\ \lim_{t\to 0} H_t(u_t^-) &= H(u), \end{aligned}$$

where $f(x_0^-)$ denotes $\lim_{x\to x_0, x\le x_0} f(x)$.

PROOF OF LEMMA 2.7.1. We have

$$\begin{aligned} |H_t(u_t) - H(u)| &\le |H_t(u_t) - H(u_t)| + |H(u_t) - H(u)| \\ &\le \|H_t - H\|_\infty + |H(u_t) - H(u)| \end{aligned}$$

and

$$\begin{aligned} |H_t(u_t^-) - H(u)| &\le |H_t(u_t^-) - H(u_t^-)| + |H(u_t^-) - H(u)| \\ &\le \|H_t - H\|_\infty + |H(u_t) - H(u)| \end{aligned}$$

as H is continuous. The first term goes to 0 as $t \to 0$ by the convergence of H_t to H on $D[0,1]$, and the second term also tends to 0 by the continuity of H, because $\lim_{t\to 0} u_t = u$. □

PROPOSITION 2.7.2 (Hadamard differentiability of $\mathcal{V}$ and $\mathcal{R}$). *Under Condition C.1,*

(i) *$\mathcal{V}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, with derivative*

$$\dot{\mathcal{V}}_{(G_0,G_1)} : (H_0, H_1) \mapsto \pi_0 \dot{\mathcal{T}}_G \left(\pi_0 H_0 + (1-\pi_0) H_1\right) + \pi_0 H_0(\mathcal{T}(G))$$

(ii) $\mathcal{R}$ *is Hadamard-differentiable at* G*, tangentially to* $C[0,1]$*, with derivative*

$$\dot{\mathcal{R}}_G : H \mapsto H(\tau^\star) + g(\tau^\star)\dot{\mathcal{T}}_G(H)$$

PROOF OF PROPOSITION 2.7.2. (i) Let $(H_0, H_1) \in C[0,1]^2$, and $(H_{0,t}, H_{1,t})_{t>0}$ be a family of functions of $D[0,1]^2$ that converges to $((H_0, H_1), \|.\|_\infty)$ as $t \to 0$. Let $H = \pi_0 H_0 + (1-\pi_0)H_1$, and $H_t = \pi_0 H_{0,t} + (1-\pi_0)H_{1,t}$. We have

$$\mathcal{V}(G_0 + tH_{0,t}, G_1 + tH_{1,t}) - \mathcal{V}(G_0, G_1) = \pi_0(\tau_t^\star - \tau^\star) + \pi_0 t H_{0,t}(\tau_t^\star)$$

where $\tau^\star = \mathcal{T}(G)$ and $\tau_t^\star$ denotes $\mathcal{T}(G + tH_t)$. By the Hadamard differentiability of $\mathcal{T}$ at G tangentially to $C[0,1]$, we have, as $H = \pi_0 H_0 + (1-\pi_0)H_1$ is continuous at $\tau^\star$,

$$\tau_t^\star - \tau^\star = t\left(\dot{\mathcal{T}}_G(H) + o(1)\right)$$

In order to conclude, we notice that

$$\lim_{t\to 0} H_{0,t}(\tau_t^\star) \to H_0(\tau^\star)$$

according to Lemma 2.7.1, which concludes the proof.

(ii) Let $H \in C[0,1]$, and H_t be a family of functions of $D[0,1]$ that converges to H on $(D[0,1], \|.\|_\infty)$ as $t \to 0$. We have

$$\begin{aligned}\mathcal{R}(G + tH_t) &= (G + tH_t)\mathcal{T}(G + tH_t) \\ &= G(\mathcal{T}(G + tH_t)) + tH_t(\mathcal{T}(G + tH_t))\end{aligned}$$

By the Hadamard differentiability of $\mathcal{T}$ at G tangentially to $C[0,1]$, we have

$$\mathcal{T}(G + tH_t) = \mathcal{T}(G) + t\left(\dot{\mathcal{T}}_G(H) + o(1)\right)$$

so that applying Taylor's formula to G at $\mathcal{T}(G)$ yields

$$G(\mathcal{T}(G + tH_t)) = G(\mathcal{T}(G)) + t\left(\dot{\mathcal{T}}_G(H) + o(1)\right) g(\mathcal{T}(G)) + o(t)\,.$$

For the second term, Lemma 2.7.1 ensures that

$$\lim_{t\to 0} H_t(\mathcal{T}(G + tH_t)) = H(\mathcal{T}(G))$$

because $\mathcal{T}(G + tH_t)$ converges to $\mathcal{T}(G)$ and H_t converges to H on $(D[0,1], \|.\|_\infty)$. Finally, we have

$$\lim_{t\to 0} \frac{\mathcal{R}(G + tH_t) - \mathcal{R}(G)}{t} = H(\tau^\star) + g(\tau^\star)\dot{\mathcal{T}}_G(H)$$

because $\tau^\star = \mathcal{T}(G)$, which concludes the proof.

□

THEOREM 2.7.3 (Asymptotic distribution of $(\hat{\tau}, \hat{\nu}, \hat{\rho})$). *Under Condition C.1,*

$$\sqrt{m}\left(\begin{pmatrix}\widehat{\tau}\\ \widehat{\nu}\\ \widehat{\rho}\end{pmatrix} - \begin{pmatrix}\tau^\star\\ \pi_0\tau^\star\\ r_\alpha(\tau^\star)\end{pmatrix}\right) \rightsquigarrow X$$

with

$$X = \begin{pmatrix} 1 \\ \pi_0 \\ g(\tau^\star) \end{pmatrix} \dot{\mathcal{T}}_G(\mathbb{Z}) + \pi_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \mathbb{Z}_0(\tau^\star) + (1-\pi_0) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \mathbb{Z}_1(\tau^\star)\,,$$

and $\mathbb{Z} = \pi_0 \mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ *and* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$*, where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$*.*

Proof of Theorem 2.7.3. We note that

$$\begin{pmatrix} \widehat{\tau} \\ \widehat{\nu} \\ \widehat{\rho} \end{pmatrix} = \Psi(\hat{\mathbb{G}}_{0,m}, \hat{\mathbb{G}}_{1,m})$$

where $\Psi : D[0,1]^2 \to \mathbb{R}^3$ is the map defined by

$$\Psi(F_0, F_1) = \begin{pmatrix} \mathcal{T}(\pi_0 F_0 + (1-\pi_0)F_1) \\ \mathcal{V}(F_0, F_1) \\ \mathcal{R}(\pi_0 F_0 + (1-\pi_0)F_1) \end{pmatrix}.$$

We have

$$\Psi(G_0, G_1) = \begin{pmatrix} \tau^\star \\ \pi_0 \tau^\star \\ G(\tau^\star) \end{pmatrix}.$$

By the Hadamard differentiability of $\mathcal{T}$ at $G = \pi_0 G_0 + (1-\pi_0)G_1$ and that of $\mathcal{V}$ at (G_0, G_1), Ψ is Hadamard-differentiable at (G_0, G_1) tangentially to $C[0,1]^2$, with derivative

$$\dot{\Psi}_{G_0,G_1}(H_0, H_1) = \begin{pmatrix} \dot{\mathcal{T}}_G(H) \\ \dot{\mathcal{V}}_{G_0,G_1}(H_0, H_1) \\ \dot{\mathcal{R}}_G(H) \end{pmatrix}$$

where H denotes $\pi_0 H_0 + (1-\pi_0)H_1$. Therefore Theorem 2.3.1 yields

$$\sqrt{m}(\Psi(\hat{\mathbb{G}}_{0,m}, \hat{\mathbb{G}}_{1,m}) - \Psi(G_0, G_1)) \rightsquigarrow \dot{\Psi}_{G_0,G_1}(\mathbb{Z}_0, \mathbb{Z}_1),$$

According to Proposition 2.7.2, we have

$$\begin{aligned} \dot{\mathcal{V}}_{(G_0,G_1)}(\mathbb{Z}_0, \mathbb{Z}_1) &= \pi_0 \dot{\mathcal{T}}_G(\mathbb{Z}) + \pi_0 \mathbb{Z}_0(\tau^\star) \\ \dot{\mathcal{R}}_G(\mathbb{Z}) &= g(\tau^\star)\dot{\mathcal{T}}_G(\mathbb{Z}) + \mathbb{Z}(\tau^\star) \end{aligned}$$

with $\mathbb{Z} = \pi_0 \mathbb{Z} + (1-\pi_0)\mathbb{Z}_1$, so that

$$\begin{aligned} X &= \dot{\Psi}_{G_0,G_1}(\mathbb{Z}_0, \mathbb{Z}_1) \\ &= \begin{pmatrix} 1 \\ \pi_0 \\ g(\tau^\star) \end{pmatrix} \dot{\mathcal{T}}_G(\mathbb{Z}) + \pi_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \mathbb{Z}_0(\tau^\star) + (1-\pi_0) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \mathbb{Z}_1(\tau^\star) \end{aligned}$$

, which concludes the proof. □

Proof of Theorem 2.3.2. (i) is a direct consequence of Theorem 2.7.3:

$$\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star)\right) \rightsquigarrow \dot{\mathcal{T}}_G(\mathbb{Z})\,,$$

For (ii) and (iii), we note that as $\tau^\star > 0$ (by Condition C.1), $\widehat{\tau} = \mathcal{T}(\hat{\mathbb{G}}_m)$ is bounded away from 0 if m is sufficiently large, with a probability 1. More precisely, there exist $t_0 \in (0,1)$ and $m_0 \in \mathbb{N}$ such that

$$\mathbb{P}\left(\forall m \geq m_0, \widehat{\tau} > t_0\right) = 1\,.$$

Therefore, as $\hat{\mathbb{G}}_m$ is non decreasing, and as $\hat{\mathbb{G}}_m(t_0) \to G(t_0) > 0$, the proportion $\widehat{\rho} = \hat{\mathbb{G}}_m(\widehat{\tau})$ of rejections by procedure $\mathcal{T}$ is bounded away from 0 with probability 1. Thus, $\mathbb{P}\left(R_m(\widehat{\tau}) > 0\right) = 1$, where $R_m(\widehat{\tau}) = m\hat{\mathbb{G}}_m(\widehat{\tau})$ is the number of rejections at threshold $\widehat{\tau}$.

As a first consequence, as $\mathsf{FDR}(t) = \mathsf{p}(t)\mathbb{P}\left(R(t) > 0\right)$, we have $\mathsf{FDR}_m(\widehat{\tau}) = \mathsf{p}(\widehat{\tau})$ for a sufficiently large m, which proves (ii) $\mathsf{p}(\widehat{\tau})$ converges almost surely to $\mathsf{p}(\tau^\star)$.

As a second consequence, letting $\gamma : \mathbb{R}_+ \times \mathbb{R}_+^* \to \mathbb{R}$ be defined by $\gamma(x,y) = \frac{x}{y}$, we have $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) = \gamma\left(\widehat{\nu}, \widehat{\rho}\right)\mathbf{1}_{\widehat{\rho}>0} = \gamma\left(\widehat{\nu}, \widehat{\rho}\right)$ for a sufficiently large m, with probability one. γ is differentiable on $\mathbb{R}_+ \times \mathbb{R}_+^*$, with derivative

$$\dot{\gamma}_{x,y} = \left(\frac{1}{y}, -\frac{x}{y^2}\right)\,.$$

In particular, $\dot{\gamma}_{\pi_0\tau^\star, G(\tau^\star)}(h,k) = \frac{1}{G(\tau^\star)}\left(h - \frac{\pi_0\tau^\star}{G(\tau^\star)}k\right)$. We can therefore derive the asymptotic distribution of FDP_m from Theorem 2.7.3 combined with the delta method [105]. According to Theorem 2.7.3 we have

$$\sqrt{m}\left(\begin{pmatrix}\widehat{\nu}\\ \widehat{\rho}\end{pmatrix} - \begin{pmatrix}\pi_0\tau^\star\\ G(\tau^\star)\end{pmatrix}\right) \rightsquigarrow \begin{pmatrix}\pi_0\mathbb{Z}_0(\tau^\star) + \pi_0\dot{\mathcal{I}}_G(\mathbb{Z})\\ \mathbb{Z}(\tau^\star) + g(\tau^\star)\dot{\mathcal{I}}_G(\mathbb{Z})\end{pmatrix}\,.$$

Hence, as $\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) = \gamma\left(\widehat{\nu}, \widehat{\rho}\right)$ (almost surely), and $\gamma(\pi_0\tau^\star, \tau^\star/\alpha) = \pi_0\alpha$, the delta method [105] yields

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \frac{\pi_0\tau^\star}{G(\tau^\star)}\right) \rightsquigarrow X,$$

where

$$\begin{aligned} X &= \frac{1}{G(\tau^\star)}\left(\pi_0(\mathbb{Z}_0(\tau^\star) + \dot{\mathcal{I}}_G(\mathbb{Z})) - \frac{\pi_0\tau^\star}{G(\tau^\star)}(\mathbb{Z}(\tau^\star) + g(\tau^\star)\dot{\mathcal{I}}_G(\mathbb{Z}))\right)\\ &= \frac{\pi_0\tau^\star}{G(\tau^\star)}\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)}\right) + \frac{\pi_0}{G(\tau^\star)}\left(1 - \frac{\tau^\star g(\tau^\star)}{G(\tau^\star)}\right)\dot{\mathcal{I}}_G(\mathbb{Z}) \end{aligned}$$

Since $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ and $G = \pi_0 G_0 + (1-\pi_0)G_1$, and $\mathsf{p}^\star = \frac{\pi_0 G_0(\tau^\star)}{G(\tau^\star)}$, we have

$$\frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)} = \mathsf{p}^\star\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + (1-\mathsf{p}^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\,,$$

so that

$$\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)} = (1-\mathsf{p}^\star)\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\right)\,,$$

which concludes the proof because $\mathsf{p}(t) = \frac{\pi_0 t}{G(t)}$ and $\dot{\mathsf{p}}(t) = \frac{\pi_0}{G(t)}\left(1 - \frac{tg(t)}{G(t)}\right)$. □

Proof of Proposition 2.3.6. Lemma 2.7.4 states the asymptotic equivalence between a multiple testing procedure defined as a threshold function and a slight modification of this procedure.

LEMMA 2.7.4. *Let $\mathcal{T}$ be a threshold function, $\varepsilon = (\varepsilon_m)_{m\in\mathbb{N}}$ and $\mathcal{T}_m^{\varepsilon,H} : D[0,1] \to [0,1]$, such that*

$$\forall F \in D[0,1], \mathcal{T}_m^{\varepsilon,H}(F) = \mathcal{T}\left(F + \varepsilon_m H\right) .$$

Let $\mathcal{M}$ be the multiple testing procedure naturally associated with the sequence of thresholds $\mathcal{T}_m^H(\hat{\mathbb{G}}_m)$. If Condition C.1 holds for $\mathcal{T}$, and if $\varepsilon_m = \mathrm{o}\left(\frac{1}{\sqrt{m}}\right)$, then $\mathcal{M}$ is asymptotically equivalent to $\mathcal{T}$ as $m \to +\infty$.

PROOF OF LEMMA 2.7.4. The proof is based on the idea that, as $\varepsilon_m = \mathrm{o}\left(\frac{1}{\sqrt{m}}\right)$, and $\hat{\mathbb{G}}_m$ converges at rate $\frac{1}{\sqrt{m}}$ to G, a modification of $\mathcal{T}$ of the order of ε_m does not change the asymptotic distribution of the associated FDP, because $\mathcal{T}$ is Hadamard-differentiable. For the sake of simplicity in notation, we prove only that

$$\sqrt{m}\left(\mathcal{T}^{\varepsilon,H}(\hat{\mathbb{G}}_m) - \mathcal{T}(\hat{\mathbb{G}}_m)\right) \xrightarrow{P} 0 .$$

Indeed, as the associated FDP is a Hadamard-differentiable function of the empirical distribution functions under the null and alternative hypotheses $\hat{\mathbb{G}}_{0,m}$ and $\hat{\mathbb{G}}_{1,m}$, the arguments developed below can be transposed (but with much more cumbersome notation) to prove that

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\varepsilon,H}(\hat{\mathbb{G}}_m)) - \mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m))\right) \xrightarrow{P} 0 .$$

Let $\mathbb{Z}_m = \sqrt{m}\left(\hat{\mathbb{G}}_m - G\right)$. According to Donsker's invariance principle (Theorem 2.3.1), $\mathbb{Z}_m$ converges in distribution on $[0,1]$ to a Gaussian process with continuous sample paths. For $Z \in D[0,1]$, let

$$\phi_m(Z) = \sqrt{m}\left(\mathcal{T}^{\varepsilon,H}(G + \frac{1}{\sqrt{m}}Z) - \mathcal{T}(G + \frac{1}{\sqrt{m}}Z)\right) .$$

We have

$$\phi_m(Z) = \sqrt{m}\left(\mathcal{T}^{\varepsilon,H}(G + \frac{1}{\sqrt{m}}Z) - \mathcal{T}(G)\right) - \sqrt{m}\left(\mathcal{T}(G + \frac{1}{\sqrt{m}}Z) - \mathcal{T}(G)\right) .$$

According to Condition C.1, $\mathcal{T}$ is Hadamard-differentiable at G tangentially to $C[0,1]$. Therefore, for any sequence Z_m of $D[0,1]$ that converges to $Z \in C[0,1]$, $\sqrt{m}\left(\mathcal{T}(G + \frac{1}{\sqrt{m}}Z_m) - \mathcal{T}(G)\right)$ converges to $\dot{\mathcal{T}}_G(Z)$. As $\varepsilon_m = \mathrm{o}\left(\frac{1}{\sqrt{m}}\right)$, $\sqrt{m}\left(\mathcal{T}^{\varepsilon,H}(G + \frac{1}{\sqrt{m}}Z_m) - \mathcal{T}(G)\right)$ also converges to $\dot{\mathcal{T}}_G(Z)$.

Thus, $\phi_m(Z_m)$ converges to 0 for any sequence Z_m of $D[0,1]$ that converges to $Z \in C[0,1]$. Therefore, according to the Extended Continuous Mapping Theorem [105, Theorem 18.11], $\phi_m(\mathbb{Z}_m)$ converges in distribution (hence also in probability) to 0. □

PROOF OF PROPOSITION 2.3.6. For $m \in \mathbb{N}$, let $\mathcal{T}_m^{\varepsilon} : D[0,1] \to [0,1]$ be defined by

$$\forall F \in D[0,1], \mathcal{T}_m^{\varepsilon}(F) = \mathcal{T}\left(F - \varepsilon_m\right) .$$

Let $\tau^\star = \mathcal{T}(G)$, $\widehat{\tau} = \mathcal{T}\left(\hat{\mathbb{G}}_m\right)$, $\widehat{\tau}_m = \mathcal{T}_m\left(\hat{\mathbb{G}}_m\right)$ and $\widehat{\tau}_m^\varepsilon = \mathcal{T}_m^\varepsilon\left(\hat{\mathbb{G}}_m\right)$. Write

$$\mathsf{FDP}_m(t) = \frac{\pi_0 \hat{\mathbb{G}}_{0,m}(t)}{\hat{\mathbb{G}}_m(t)},$$

where $\hat{\mathbb{G}}_{0,m}$ and $\hat{\mathbb{G}}_m$ are non decreasing functions, so that

$$\frac{\pi_0 \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}_m^\varepsilon\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}\right)} \leq \mathsf{FDP}_m(\widehat{\tau}_m) \leq \frac{\pi_0 \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}_m^\varepsilon\right)}$$

because $\widehat{\tau}_m^\varepsilon \leq \widehat{\tau}_m \leq \widehat{\tau}$. Therefore,

$$\begin{aligned}
&\frac{\pi_0\left(\hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}_m^\varepsilon\right) - \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}\right)\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}\right)} \leq \mathsf{FDP}_m(\widehat{\tau}_m) - \mathsf{FDP}_m(\widehat{\tau}) \\
\leq &\frac{\pi_0\left(\hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}\right) - \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}_m^\varepsilon\right)\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}_m^\varepsilon\right)} - \left(\mathsf{FDP}_m(\widehat{\tau}) - \mathsf{FDP}_m(\widehat{\tau}_m^\varepsilon)\right)
\end{aligned}$$

As $\mathcal{T}_m^\varepsilon \leq \mathcal{T}_m \leq \mathcal{T}$ for any $m \in \mathbb{N}$, Lemma 2.7.4 ensures that $\sqrt{m}\left(\widehat{\tau}_m - \widehat{\tau}\right)$ and $\sqrt{m}\left(\widehat{\tau}_m^\varepsilon - \widehat{\tau}\right)$ converge to 0 in probability. Therefore, as $\hat{\mathbb{G}}_m\left(\widehat{\tau}\right)$ and $\hat{\mathbb{G}}_m\left(\widehat{\tau}_m^\varepsilon\right)$ converge in probability to $G\left(\mathcal{T}(G)\right)$ as $m \to +\infty$, we have

$$\frac{\pi_0\sqrt{m}\left(\hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}_m^\varepsilon\right) - \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}\right)\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}\right)} \xrightarrow{P} 0$$

and

$$\frac{\pi_0\sqrt{m}\left(\hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}\right) - \hat{\mathbb{G}}_{0,m}\left(\widehat{\tau}_m^\varepsilon\right)\right)}{\hat{\mathbb{G}}_m\left(\widehat{\tau}_m^\varepsilon\right)} \xrightarrow{P} 0,$$

which concludes the proof because $\sqrt{m}\left(\mathsf{FDP}_m(\widehat{\tau}) - \mathsf{FDP}_m(\widehat{\tau}_m^\varepsilon)\right)$ also converges in probability to 0 (according to Lemma 2.7.4). □

2.7.2. Asymptotic FDP: specific threshold functions. We now apply the results of section 2.7.1 to threshold functions of the form

$$\mathcal{T}(F) = \mathcal{U}\left(F, \mathcal{A}(F)\right),$$

with

$$\mathcal{U}(F,\alpha) = \sup\{u \in [0,1], F(u) \geq r_\alpha(u)\}.$$

In this section, we will use the notation $r : (\alpha, u) \mapsto r_\alpha(u)$ whenever the dependence of r_α in α is of importance. We begin by giving sufficient conditions for the regularity of $\mathcal{U}$ and $\mathcal{A}$ under which Condition C.1 holds (section 2.7.2, Proposition 2.7.5). Then we provide sufficient conditions for $\mathcal{U}$ to be regular enough to be consistent with hypotheses (i) to (iii) of Proposition 2.7.5 (section 2.7.2). Finally we derive the form of the asymptotic distribution of the corresponding False Discovery Proportion (section 2.7.2).

Hadamard differentiability of $\mathcal{T}$.

PROPOSITION 2.7.5 (Hadamard differentiability of $\mathcal{T}$). *Let $C[0,1]$ be the set of continuous functions of $D[0,1]$. Suppose that*

(i) *$\mathcal{U}$ is Hadamard-differentiable with respect to its first variable at (G,α), tangentially to $C[0,1]$, for any α in a neighborhood of $\mathcal{A}(G)$; its derivative will be denoted by $\nabla_F \mathcal{U}_{G,\alpha}$;*

(ii) *$\nabla_F \mathcal{U}_{G,\cdot}$ is continuous at $\mathcal{A}(G)$;*

(iii) *$\mathcal{U}$ is differentiable with respect to its second variable; its derivative will be denoted by $\nabla_\alpha \mathcal{U}_{(G,\mathcal{A}(G))}$;*

(iv) *$\mathcal{A}$ is Hadamard-differentiable at G tangentially to $C[0,1]$; its derivative will be denoted by $\dot{\mathcal{A}}_G$.*

Then, $\mathcal{T}$ is Hadamard-differentiable at G tangentially to $C[0,1]$, with derivative $\dot{\mathcal{T}}_G$, defined for any $H \in D[0,1]$ by

$$\dot{\mathcal{T}}_G(H) = \nabla_F \mathcal{U}_{(G,\mathcal{A}(G))}(H) + \dot{\mathcal{A}}_G(H) \nabla_\alpha \mathcal{U}_{(G,\mathcal{A}(G))}\,.$$

PROOF OF PROPOSITION 2.7.5. Let $H \in C[0,1]$, and H_t be a family of functions of $D[0,1]$ that converges to H on $(D[0,1], \|.\|_\infty)$ as $t \to 0$.

As $\mathcal{A}$ is continuous at G (by (iv)), $\mathcal{A}(G + tH_t)$ lies in a neighborhood of $\mathcal{A}(G)$ for small $t > 0$, and $\mathcal{U}$ is Hadamard-differentiable with respect to its first variable at $(G, \mathcal{A}(G + tH_t))$ by (i). We therefore have

$$\begin{aligned}
\mathcal{T}(G + tH_t) &= \mathcal{U}(G + tH_t, \mathcal{A}(G + tH_t)) \\
&= \mathcal{U}(G, \mathcal{A}(G + tH_t)) + t \nabla_F \mathcal{U}_{G,\mathcal{A}(G+tH_t)}(H)(1 + o(1)) \\
&= \mathcal{U}(G, \mathcal{A}(G + tH_t)) + t \nabla_F \mathcal{U}_{G,\mathcal{A}(G)}(H)(1 + o(1))
\end{aligned}$$

by the continuity of $\nabla_F \mathcal{U}_{G,\cdot}$ at $\mathcal{A}(G)$ (ii). Then, combining (iii) and (iv) yields

$$\begin{aligned}
\mathcal{U}(G, \mathcal{A}(G + tH_t)) &= \mathcal{U}\left(G, \mathcal{A}(G) + t\dot{\mathcal{A}}_G(H)(1 + o(1))\right) \\
&= \mathcal{U}(G, \mathcal{A}(G)) + t\dot{\mathcal{A}}_G(H) \nabla_\alpha \mathcal{U}_{(G,\mathcal{A}(G))}(1 + o(1)) \\
&= \mathcal{T}(G) + t\dot{\mathcal{A}}_G(H) \nabla_\alpha \mathcal{U}_{(G,\mathcal{A}(G))}(1 + o(1))
\end{aligned}$$

so that

$$\lim_{t \to 0} \frac{\mathcal{T}(G + tH_t) - \mathcal{T}(G)}{t} = \nabla_F \mathcal{U}_{(G,\mathcal{A}(G))}(H) \dot{\mathcal{A}}_G(H) \nabla_\alpha \mathcal{U}_{(G,\mathcal{A}(G))},$$

which concludes the proof. □

Regularity of $\mathcal{U}$. The crucial point for proving the desired regularity of $\mathcal{U}$ is its Hadamard differentiability with respect to its first variable at (G,α), tangentially to $C[0,1]$, for α in a neighborhood of $\mathcal{A}(G)$. Lemma 2.7.6 is a straightforward analytical translation of Conditions C.2 and C.3.

LEMMA 2.7.6. *Under Conditions C.2 and C.3, the unique interior right crossing point $\tau^\star$ between r_α and G is positive. If r is C^1 on $(0,1] \times [0,1]$, there exists a neighborhood $V = A \times U$ of $(\mathcal{A}(G), \tau^\star)$ such that for any $(\alpha, x) \in V$, $\psi_{G,\alpha} : u \mapsto r_\alpha(u) - G(u)$ is locally invertible around $\mathcal{U}(G,\alpha)$, with $\dot{\psi}_{G,\alpha}(x) > 0$.*

We begin by proving the continuity of $\mathcal{U}$ at (G,α) for α in a neighborhood of $\mathcal{A}(G)$. We then (Proposition 2.7.11) provide the sufficient conditions for conditions (i), (ii), and (iii) of Proposition 2.7.5 to hold.

LEMMA 2.7.7. *For any $F \in D[0,1]$, and $\alpha \in [0,1]$ such that r_α is continuous, one of the following two assertions holds:*

(i) $F(\mathcal{U}(F,\alpha)) = r_\alpha(\mathcal{U}(F,\alpha))$
(ii) $F(\mathcal{U}(F,\alpha)) \leq r_\alpha(\mathcal{U}(F,\alpha)) \leq F(\mathcal{U}(F,\alpha)^-)$

PROOF OF LEMMA 2.7.7. According to the definition of $\mathcal{U}(F,\alpha)$, $F(u) \leq r_\alpha(u)$ for any $u > \mathcal{U}(F,\alpha)$. As F is right-continuous and r_α is continuous, we have $F(\mathcal{U}(F,\alpha)) \leq r_\alpha(\mathcal{U}(F,\alpha))$. Therefore, either (i) holds, or $F(\mathcal{U}(F,\alpha)) < r_\alpha(\mathcal{U}(F,\alpha))$. In the second case, according to the definition of $\mathcal{U}(F,\alpha)$, there is a non decreasing sequence (u_n) that converges to $\mathcal{U}(F,\alpha)$ such that $F(u_n) \geq r_\alpha(u_n)$. As r_α is continuous and F is left-continuous, we have $F(\mathcal{U}(F,\alpha)^-) \geq r_\alpha(\mathcal{U}(F,\alpha))$, which proves (ii). □

PROPOSITION 2.7.8. *Let $F \in D[0,1]$ be non decreasing, and $\alpha = \mathcal{A}(F)$. If r_α is continuous, then*

$$F(\mathcal{U}(F,\alpha)) = r_\alpha(\mathcal{U}(F,\alpha))\,.$$

PROOF OF PROPOSITION 2.7.8. Let us consider the two assertions of Lemma 2.7.7: as F is non decreasing, (ii) can be reduced to (i). □

PROPOSITION 2.7.9. *Let r be continuous on $(0,1]\times[0,1]$. Let $F \in D[0,1]$ be non decreasing, and $\alpha \in (0,1]$. Let F_t be a sequence of functions of $D[0,1]$ such that $(F_t)_{t>0}$ converges to F on $(D[0,1], \|.\|_\infty)$ as $t \to 0$, and $\alpha_t \to \alpha$ as $t \to 0$. Denote by $\psi_{F,\alpha}$ the function defined on $[0,1]$ by*

$$\forall u \in [0,1], \psi_{F,\alpha}(u) = r_\alpha(u) - F(u)\,.$$

Then,

$$\lim_{t\to 0} \psi_{F,\alpha}\left(\mathcal{U}(F_t,\alpha_t)\right) = \psi_{F,\alpha}\left(\mathcal{U}(F,\alpha)\right)\,.$$

PROOF OF PROPOSITION 2.7.9. For each fixed $t \in [0,1]$, one of the following two assertions holds according to Lemma 2.7.7:

(i) $F_t(\mathcal{U}(F_t,\alpha_t)) = r_{\alpha_t}(\mathcal{U}(F_t,\alpha_t))$
(ii) $F_t(\mathcal{U}(F_t,\alpha_t)) \leq r_{\alpha_t}(\mathcal{U}(F_t,\alpha_t)) \leq F_t(\mathcal{U}(F_t,\alpha_t)^-)$.

If (ii) holds, then, as F is non decreasing we have

$$(F-F_t)(\mathcal{U}(F_t,\alpha_t)^-) \leq F(\mathcal{U}(F_t,\alpha_t)) - r_{\alpha_t}(\mathcal{U}(F_t,\alpha_t)) \leq (F-F_t)(\mathcal{U}(F_t,\alpha_t))\,.$$

If (i) holds, then $F(\mathcal{U}(F_t,\alpha_t)) - r_{\alpha_t}(\mathcal{U}(F_t,\alpha_t)) = F(\mathcal{U}(F_t,\alpha_t)) - F_t(\mathcal{U}(F_t,\alpha_t))$. In either case, we have $|F(\mathcal{U}(F_t,\alpha_t)) - r_{\alpha_t}(\mathcal{U}(F_t,\alpha_t))| \leq \|F - F_t\|$, which tends to 0 as $t \to 0$. As r is continuous on $(0,1]\times[0,1]$, r_{α_t} converges uniformly to r_α on the compact $[\alpha/2,1]$, and we have

$$\lim_{t\to 0} F(\mathcal{U}(F_t,\alpha_t)) - r_\alpha(\mathcal{U}(F_t,\alpha_t)) = 0\,,$$

which concludes the proof as $\psi_{F,\alpha}\left(\mathcal{U}(F,\alpha)\right) = 0$. □

COROLLARY 2.7.10 (Continuity of $\mathcal{U}$). *Let r be C^1 on $(0,1]\times[0,1]$. According to Conditions C.2 and C.3, there is a neighborhood A of $\mathcal{A}(G)$ such that $\mathcal{U}$ is continuous at (G,α) for any $\alpha \in A$.*

PROPOSITION 2.7.11 (Differentiability of $\mathcal{U}$). *Let us assume that r is C^1 on $(0,1]\times[0,1]$. Under Conditions C.2 and C.3,*

(i) *$\mathcal{U}$ is Hadamard-differentiable with respect to its first variable at (G,α), tangentially to $C[0,1]$ for any α in a neighborhood A of $\mathcal{A}(G)$, with derivative $\nabla_F\mathcal{U}_{G,\alpha}$ defined for any $H\in C[0,1]$ by*

$$\nabla_F\mathcal{U}_{G,\alpha}(H)=\frac{H(\mathcal{U}(G,\alpha))}{\frac{\partial r}{\partial u}(\alpha,\mathcal{U}(G,\alpha))-g(\mathcal{U}(G,\alpha))}$$

(ii) *$\nabla_F\mathcal{U}_{G,\cdot}$ is continuous at $\mathcal{A}(G)$ on A.*

(iii) *$\mathcal{U}$ is differentiable with respect to its second variable, with derivative*

$$\nabla_\alpha\mathcal{U}_{G,\mathcal{A}(G)}=-\frac{\frac{\partial r}{\partial\alpha}(\mathcal{A}(G),\tau^\star))}{\frac{\partial r}{\partial u}(\mathcal{A}(G),\tau^\star))-g(\tau^\star))}\,,$$

where $\tau^\star=\mathcal{U}(G,\mathcal{A}(G))$.

COROLLARY 2.7.12. *If we also assume that $\mathcal{A}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, then $\mathcal{T}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, with derivative defined for any $H\in C[0,1]$ by*

$$\dot{\mathcal{T}}_G(H)=\frac{H(\tau^\star)-\frac{\partial r}{\partial\alpha}(\mathcal{A}(G),\tau^\star))\dot{\mathcal{A}}_G(H)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t)-g(t)}$$

PROOF OF PROPOSITION 2.7.11. Let $\tau^\star=\mathcal{U}(G,\mathcal{A}(G))$. Throughout the proof, $V=A\times U$ denotes the neighborhood of $(\mathcal{A}(G),\tau^\star)$ defined in Lemma 2.7.6.

(i) Let $\alpha\in A$. Let $H\in C[0,1]$, and H_t be a family of functions of $D[0,1]$ that converges to H on $(D[0,1],\|.\|_\infty)$ as $t\to 0$. Let $v=\mathcal{U}(G,\alpha)$ and $v_t=\mathcal{U}(G_t,\alpha)$, with $G_t=G+tH_t$. By the continuity of U (Corollary 2.7.10), $v_t\to v$ as $t\to 0$. Therefore, applying Taylor's formula to $\psi_{G,\alpha}:u\mapsto r_\alpha(u)-G(u)$ yields

$$\psi_{G,\alpha}(v_t)-\psi_{G,\alpha}(v)\underset{t\to 0}{=}(v_t-v)\dot{\psi_{G,\alpha}}(v)\,(1+o(1))\,.$$

As $\alpha\in A$, we have $\dot{\psi_{G,\alpha}}(v)=\frac{\partial r}{\partial u}(\alpha,v)-g(v)>0$. Therefore, as $\psi_{G,\alpha}(v)=0$ according to Proposition 2.7.8,

$$v_t-v\underset{t\to 0}{=}\frac{\psi_{G,\alpha}(v_t)}{\frac{\partial r}{\partial u}(\alpha,v)-g(v)}\,(1+o(1))$$

so that it is sufficient to prove that $\lim_{t\to0}\psi_{G,\alpha}(v_t)/t=H(v)$. The behavior of $\psi_{G,\alpha}(v_t)=r_\alpha(v_t)-G(v_t)$ can be derived using the same argument as in the proof of proposition 2.7.9; Lemma 2.7.7, we have either $G_t(v_t)=r_\alpha(v_t)$ or $G_t(v_t)\le r_\alpha(v_t)\le G_t(v_t^-)$. In the first case, $r_\alpha(v_t)-G(v_t)=(G_t-G)(v_t)=tH_t(v_t)$, and $\lim_{t\to0}\frac{r_\alpha(v_t)-G(v_t)}{t}=H(t)$ according to Lemma 2.7.1. In the second case, we have

$$(G_t-G)(v_t)\le r_\alpha(v_t)-G(v_t)\le(G_t-G)(v_t^-)$$

as G is non decreasing, that is, $tH_t(v_t)\le\psi_{G,\alpha}(v_t)\le tH_t(v_t^-)$. Therefore, we have

$$H_t(v_t) - H(v_t) \leq \frac{\psi_{G,\alpha}(v_t)}{t} - H(v_t) \leq H_t(v_t^-) - H(v_t)\,.$$

As H is continuous, $H(v_t) = H(v_t^-)$, and the upper and lower bounds converge to 0 according to Lemma 2.7.1, and (i) is proved.

(ii) is a consequence of the continuity of $\mathcal{U}$ (with respect to its second variable), that of g and that of $\nabla_u r$ with respect to its first variable.

(iii) Let $\alpha = \mathcal{A}(G)$. Let α_t be a sequence of points of $(0,1]$ that converges to α. Let $v = \mathcal{U}(G,\alpha)$ and $v_t = \mathcal{U}(G,\alpha_t)$. By the continuity of U (Proposition 2.7.9), $v_t \to v$ as $t \to 0$. Therefore, applying Taylor's formula to $\psi_{G,\alpha} : u \mapsto r_\alpha(u) - G(u)$ yields

$$\psi_{G,\alpha}(v_t) - \psi_{G,\alpha}(v) \underset{t\to 0}{=} (v_t - v)\dot{\psi}_{G,\alpha}(v)\,(1 + o(1))\ .$$

We have $\psi_{G,\alpha}(v) = 0$ and $\psi_{G,\alpha}(v_t) = r(\alpha, v_t) - r(\alpha_t, v_t)$ by Proposition 2.7.8. As r is C^1 in a neighborhood of (α, v), we have, according to Taylor's formula,

$$r(\alpha_t, v_t) - r(\alpha, v_t) \underset{t\to 0}{=} (\alpha_t - \alpha)\frac{\partial r}{\partial \alpha}(\alpha, v)\,(1 + o(1))\ .$$

As $\alpha = \mathcal{A}$ and $v = \mathcal{U}(G,\alpha)$, $(\alpha, v) \in V$. Therefore $\dot{\psi}_{G,\alpha}(v) = \frac{\partial r}{\partial u}(\alpha, v) - g(v) > 0$ according to Lemma 2.7.6. Finally, we have

$$\lim_{t\to 0} \frac{r(\alpha, v_t) - r(\alpha_t, v_t)}{\alpha_t - \alpha} = -\frac{\frac{\partial r}{\partial \alpha}(\alpha, v)}{\frac{\partial r}{\partial u}(\alpha, v) - g(v)},$$

which concludes the proof. □

Asymptotic FDP.

THEOREM 2.7.13 (Asymptotic distribution of FDP_m). *Let r_α be a rejection curve such that r is C^1 on $(0,1] \times [0,1]$, and $\mathcal{A}$ a level function. Let us denote by $\mathcal{T} : F \mapsto \mathcal{U}(F, \mathcal{A}(F))$ the associated threshold function, where $\mathcal{U}(F,\alpha) = \sup\{u \in [0,1], F(u) \geq r_\alpha(u)\}$.*

Under Conditions C.2 and C.3, if $\mathcal{A}$ is Hadamard-differentiable at G tangentially to $C[0,1]$, then

(i)

$$\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star\right) \rightsquigarrow \frac{\mathbb{Z}(\tau^\star) - \frac{\partial r}{\partial \alpha}(\mathcal{A}(G), \tau^\star))\dot{\mathcal{A}}_G(\mathbb{Z})}{\frac{\partial r}{\partial u}(\mathcal{A}(G), t) - g(t)},$$

(ii)

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}^\star\right) \rightsquigarrow X,$$

where $\mathsf{p}^\star = \frac{\pi_0 \tau^\star}{G(\tau^\star)}$ is the pFDR achieved by procedure $\mathcal{T}$, $\dot{\mathsf{p}}(t) = \frac{\pi_0}{G(t)}\left(1 - \frac{t g(t)}{G(t)}\right)$, and

$$X = \mathsf{p}^\star(1 - \mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \mathsf{p}^\star(1 - \mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)} + \dot{\mathsf{p}}(\tau^\star)\xi(\tau^\star)\dot{\mathcal{A}}_G(\mathbb{Z})\,,$$

with $\zeta(t) = -\frac{\frac{G(t)}{t} - \frac{\partial r}{\partial u}(\mathcal{A}(G),t)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$, $\xi(t) = \frac{-\frac{\partial r}{\partial \alpha}(\mathcal{A}(G),t)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$, *and* $\mathbb{Z} = \pi_0 \mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ *and* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$, *where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$.

Proof of Theorem 2.7.13. Under these assumptions, Condition C.1 holds for $\mathcal{T}$ according to Corollary 2.7.12, with

$$\dot{\mathcal{T}}_G(H) = \frac{H(\tau^\star) - \frac{\partial r}{\partial \alpha}(\mathcal{A}(G),\tau^\star))\dot{\mathcal{A}}_G(H)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$$

Therefore, Theorem 2.3.2 yields $\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star)\right) \rightsquigarrow \dot{\mathcal{T}}_G(\mathbb{Z})$, and

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \frac{\pi_0 \tau^\star}{G(\tau^\star)}\right) \rightsquigarrow X\,,$$

with $X = \mathsf{p}^\star(1-\mathsf{p}^\star)\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\right) + \dot{\mathsf{p}}(\tau^\star)\dot{\mathcal{T}}_G(\mathbb{Z})$ and $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ and $\mathbb{Z}_0$ and $\mathbb{Z}_1$ are independent Gaussian processes such that $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ and $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B} \circ G_1$, where $\mathbb{B}$ is a standard Brownian bridge on $[0,1]$.

Letting $\delta(t) = \frac{1}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$, we have

$$\dot{\mathcal{T}}_G(\mathbb{Z}) = \delta(\tau^\star)\left(\mathbb{Z}(\tau^\star) - \frac{\partial r}{\partial \alpha}(\mathcal{A}(G),\tau^\star))\dot{\mathcal{A}}_G(H)\right).$$

As $\dot{\mathsf{p}}(t) = \mathsf{p}(t)\left(\frac{1}{t} - \frac{g(t)}{G(t)}\right)$, we have

$$\dot{\mathsf{p}}(\tau^\star)\delta(\tau^\star)\mathbb{Z}(\tau^\star) = \mathsf{p}^\star\left(\frac{G(\tau^\star)}{\tau^\star} - g(\tau^\star)\right)\delta(\tau^\star)\frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)}\,,$$

with $\frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)} = \mathsf{p}^\star\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + (1-\mathsf{p}^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}$. Hence letting $\zeta(t) = 1 - \delta(t)\left(\frac{G(t)}{t} - g(t)\right)$, we have

$$X = \mathsf{p}^\star(1-\mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)} + \dot{\mathsf{p}}(\tau^\star)\xi(\tau^\star)\dot{\mathcal{A}}_G(\mathbb{Z})\,,$$

where $\xi(t) = \frac{-\frac{\partial r}{\partial \alpha}(\mathcal{A}(G),t)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$. This concludes the proof since ζ may be written as $\zeta(t) = -\frac{\frac{G(t)}{t} - \frac{\partial r}{\partial u}(\mathcal{A}(G),t)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$. □

2.7.3. Limit distribution for procedures under consideration.

One-stage procedures. In this section $\mathcal{A}(G)$ is fixed. Therefore, only the dependence of r_α u is of importance. In order to lighten the notation we let

$$\dot{r_\alpha} = \frac{\partial r}{\partial u}(\alpha,\cdot)\,.$$

Theorem 2.7.14 (Asymptotic FDP for one-stage procedures). *Let* $\mathcal{T} : F \mapsto \mathcal{U}(F,\alpha)$ *a one-stage procedure such that* r_α *is continuous on* $[0,1]$, *and* C^1 *in a neighborhood of* $\tau^\star = \mathcal{T}(G)$. *Under Condition C.2 and C.3,*

(i)

$$\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star)\right) \rightsquigarrow \frac{\mathbb{Z}(\tau^\star)}{\dot{r_\alpha}(\tau^\star) - g(\tau^\star)}$$

(ii)

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}^\star\right) \rightsquigarrow X\,,$$

with

$$X = \mathsf{p}^\star(1 - \mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star))\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\,,$$

where $\mathsf{p}^\star = \frac{\pi_0\tau^\star}{r_\alpha(\tau^\star)}$, $\zeta(t) = \frac{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - \frac{G(t)}{t}}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$, $\mathbb{Z} = \pi_0\mathbb{Z}_0 + (1-\pi_0)\,\mathbb{Z}_1$ *and* $\mathbb{Z}_0$ *and* $\mathbb{Z}_1$ *are independent Gaussian processes such that* $\mathbb{Z}_0 \overset{(d)}{=} \mathbb{B}$ *and* $\mathbb{Z}_1 \overset{(d)}{=} \mathbb{B}\circ G_1$, *where* $\mathbb{B}$ *is a standard Brownian bridge on* $[0,1]$.

Proof of Theorem 2.7.14. As $\mathcal{T}$ is a one-stage procedure, we have $\mathcal{A} = \alpha$. Therefore, the assumptions for Theorem 2.7.13 hold, with $\xi = 0$. According to Proposition 2.7.8, $G(\tau^\star) = r_\alpha(\tau^\star)$; therefore $\mathsf{p}^\star = \frac{\pi_0\tau^\star}{r_\alpha(\tau^\star)}$, which concludes the proof. □

Proof of Theorem 2.4.2 (BH95). Uniqueness Condition C.3 always holds because r_α is linear, and Condition C.2 holds because it corresponds to Condition C.4. Therefore, Theorem 2.7.14 can be applied, and we have $\zeta(\tau^\star) = 0$ since $\dot{r_\alpha}(\tau^\star) = 1/\alpha = r_\alpha(\tau^\star)/\tau^\star$, and $\mathsf{p}(\tau^\star) = \pi_0\alpha$. Hence,

$$\sqrt{m}\left(\mathcal{T}(\hat{\mathbb{G}}_m) - \tau^\star)\right) \rightsquigarrow \frac{\mathbb{Z}(\tau^\star)}{1/\alpha - g(\tau^\star)}$$

and

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \pi_0\alpha\right) \rightsquigarrow \pi_0\alpha\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star},$$

which concludes the proof because $\operatorname{Var}\mathbb{Z}_0(\tau^\star) = \tau^\star(1-\tau^\star)$. □

Proof of Theorem 2.4.7 (FDR08). The uniqueness Condition C.3, and existence Conditions C.4 and C.7 ensure that there is a unique interior right crossing point $\tau^\star$ between f_α and G, which satisfies $\tau^\star \le \kappa$. Condition C.6 guarantees that $\tau^\star$ is also the only right crossing point between f_α^λ and G. Thus, $[0,\kappa]$ is a neighborhood of $\tau^\star$ in which f_α^λ coincides with f_α and is C^1, with $\dot{f}_\alpha(u) = \frac{\alpha}{(\alpha+(1-\alpha)u)^2}$. Therefore, Theorem 2.7.14 yields $\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}^\star\right) \rightsquigarrow X$, with

$$X = \mathsf{p}^\star(1 - \mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\,,$$

where $\mathsf{p}^\star = \frac{\pi_0\tau^\star}{f_\alpha(\tau^\star)} = \pi_0(\alpha+(1-\alpha)\tau^\star)$ and $\zeta(\tau^\star) = -\frac{G(\tau^\star)/\tau^\star - \dot{f}_\alpha(\tau^\star)}{f_\alpha(\tau^\star) - g(\tau^\star)}$. Letting

$$\overline{\pi_0}(t) = \frac{1-G(t)}{1-t}\,,$$

we have $G(\tau^\star)/\tau^\star = \overline{\pi_0}(\tau^\star)/\alpha$, and $\dot{f_\alpha}(\tau^\star) = \alpha(f_\alpha(\tau^\star)/\tau^\star)^2 = \overline{\pi_0}(\tau^\star)^2/\alpha$, so that $\mathsf{p}^\star = \alpha\pi_0/\overline{\pi_0}\,(\tau^\star)$, and

$$\begin{aligned}\zeta(\tau^\star) &= -\frac{G(\tau^\star)/\tau^\star - \dot{f_\alpha}(\tau^\star)}{\dot{f_\alpha}(\tau^\star) - g(\tau^\star)} \\ &= -\frac{\overline{\pi_0}(\tau^\star)/\alpha - \overline{\pi_0}(\tau^\star)^2/\alpha}{\overline{\pi_0}(\tau^\star)^2/\alpha - g(\tau^\star)} \\ &= -(1-\overline{\pi_0}(\tau^\star))\frac{\overline{\pi_0}(\tau^\star)/\alpha}{\overline{\pi_0}(\tau^\star)^2/\alpha - g(\tau^\star)}\,,\end{aligned}$$

which concludes the proof. □

Proof of Theorem 2.4.10 (BR08(λ)). The uniqueness Condition C.3, and existence Conditions C.8 and C.9 ensure that there is a unique interior right crossing point $\tau^\star$ between b_α^λ and G, which satisfies $\tau^\star \leq \lambda$. Thus $[0,\lambda]$ is a neighborhood of $\tau^\star$ in which b_α^λ is C^1, with $\dot{b_\alpha^\lambda}(u) = \frac{\alpha(1-\lambda)}{(\alpha(1-\lambda)+u)^2}$. Therefore, Theorem 2.7.14 yields $\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}^\star\right) \rightsquigarrow X$, with

$$X = \mathsf{p}^\star(1-\mathsf{p}^\star\zeta(\tau^\star))\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}_1(\tau^\star)}{G_1(\tau^\star)}\,,$$

where $\mathsf{p}^\star = \frac{\pi_0\tau^\star}{b_\alpha(\tau^\star)} = \pi_0(\alpha+(1-\alpha)\tau^\star)$ and $\zeta(\tau^\star) = -\frac{G(\tau^\star)/\tau^\star - \dot{b_\alpha}(\tau^\star)}{\dot{b_\alpha}(\tau^\star) - g(\tau^\star)}$. We have $\dot{f_\alpha}(\tau^\star) = \alpha(1-\lambda)(b_\alpha(\tau^\star)/\tau^\star)^2 = G(\tau^\star)(1-G(\tau^\star))/\tau^\star$, so that

$$\begin{aligned}\zeta(\tau^\star) &= -\frac{G(\tau^\star)/\tau^\star - \dot{b_\alpha}(\tau^\star)}{\dot{b_\alpha}(\tau^\star) - g(\tau^\star)} \\ &= -\frac{G(\tau^\star)/\tau^\star(1-(1-G(\tau^\star)))}{G(\tau^\star)(1-G(\tau^\star))/\tau^\star - g(\tau^\star)} \\ &= -\frac{G(\tau^\star)^2/\tau^\star}{G(\tau^\star)(1-G(\tau^\star))/\tau^\star - g(\tau^\star)}\,,\end{aligned}$$

which concludes the proof. □

Two-stage adaptive procedures.

Proof of Theorem 2.4.12. As pointed out in section 2.4.3, Condition C.3 always holds because r_α is linear, and Condition C.2 holds as soon as $\mathcal{A}(G) > \alpha^\star$. Therefore, Theorem 2.7.13 yields $\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \mathsf{p}^\star\right) \rightsquigarrow X$, with $\mathsf{p}^\star = \pi_0\mathcal{A}(G)$, and

$$X = \mathsf{p}^\star(1-\zeta(\tau^\star)\mathsf{p}^\star)\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} - \mathsf{p}^\star(1-\mathsf{p}^\star)\zeta(\tau^\star)\frac{\mathbb{Z}(\tau^\star)}{G(\tau^\star)} + \dot{\mathsf{p}}(\tau^\star)\xi(\tau^\star)\dot{\mathcal{A}}_G(\mathbb{Z})\,,$$

where $\dot{\mathsf{p}}(\tau^\star) = \frac{\mathsf{p}^\star}{G(\tau^\star)}\left(\frac{G(\tau^\star)}{\tau^\star} - g(\tau^\star)\right)$, $\zeta(t) = -\frac{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - G(t)/t}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$, and $\xi(t) = \frac{-\frac{\partial r}{\partial \alpha}(\mathcal{A}(G),t)}{\frac{\partial r}{\partial u}(\mathcal{A}(G),t) - g(t)}$. Simes' line is defined by $r_\alpha : u \mapsto u/\alpha$. Therefore, we have $\frac{\partial r}{\partial u}(\mathcal{A}(G),t) = \frac{1}{\mathcal{A}(G)}$ and $\frac{\partial r}{\partial \alpha}(\mathcal{A}(G),t) = -\frac{t}{\mathcal{A}(G)^2}$, and $G(\tau^\star) = \frac{\tau^\star}{\mathcal{A}(G)}$ according to Proposition 2.7.8. We have $\zeta(\tau^\star) = 0$, $\xi(\tau^\star) = \frac{\tau^\star/\mathcal{A}(G)^2}{\frac{1}{\mathcal{A}(G)} - g(\tau^\star)}$, and $\dot{\mathsf{p}}(\tau^\star) = \mathsf{p}^\star\frac{\mathcal{A}(\mathcal{G})}{\tau^\star}\left(\frac{1}{\mathcal{A}(G)} - g(\tau^\star)\right)$, which concludes the proof. □

Sto02 procedure. The following Proposition establishes the Hadamard differentiability of the level function of procedure Sto02. The proof is immediate.

PROPOSITION 2.7.15. *For $F \in D[0,1]$, let*

$$\mathcal{A}(F) = \alpha \frac{1-\lambda}{1-F(\lambda)},$$

where $\alpha \in [0,1]$. Under Condition C.11, $\mathcal{A}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, with derivative

$$\dot{\mathcal{A}}_G(H) = \mathcal{A}(G) \frac{H(\lambda)}{1-G(\lambda)}\,.$$

PROPOSITION 2.7.16. *Let $\lambda \in [0,1)$ such that Conditions C.9 and C.11 hold. Then procedures* Sto02(λ) *and* STS04(λ) *are asymptotically equivalent.*

PROOF OF PROPOSITION 2.7.16. Let $\lambda \in [0,1)$. According to Condition C.9, we have $\mathcal{T}^{\mathsf{Sto02}(\lambda)}(G) < \lambda$. Therefore, procedure Sto02 is asymptotically equivalent to the same procedure truncated at λ, that is, the procedure with threshold function defined for $F \in D[0,1]$ by

$$\sup\left\{ u \in [0,\lambda], F(u) \geq \frac{u}{\alpha}\frac{1-\lambda}{1-F(\lambda)} \right\},$$

so that we will work with this truncated version for the remainder of the proof. By definition, the rejection curve of procedure STS04 is larger than that of procedure Sto02. Therefore, we have $\mathcal{T}_m^{\mathsf{STS04}(\lambda)}(F) \leq \mathcal{T}^{\mathsf{Sto02}(\lambda)}(F)$ for any $F \in D[0,1]$. With the same argument we also have $\mathcal{T}^{\mathsf{Sto02}(\lambda)}\left(F - \frac{1}{m}\right) \leq \mathcal{T}_m^{\mathsf{STS04}(\lambda)}(F)$ for any $F \in D[0,1]$. As we have assumed that Condition C.11 holds, Condition C.1 holds for $\mathcal{T}$ according to Proposition 2.7.15, and the result follows from Proposition 2.3.6. □

PROOF OF THEOREM 2.4.15. According to Proposition 2.7.15, and because $\mathcal{A}(G) > \alpha^\star$, Theorem 2.4.12 ensures that

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}^{\mathsf{Sto02}}(\hat{\mathbb{G}}_m)) - \pi_0\mathcal{A}(G)\right) \rightsquigarrow \pi_0\mathcal{A}(G)\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \frac{\dot{\mathcal{A}}_G(\mathbb{Z})}{\mathcal{A}(G)}\right),$$

where

$$\dot{\mathcal{A}}_G(\mathbb{Z}) = \mathcal{A}(G)\frac{\mathbb{Z}(\lambda)}{1-G(\lambda)}$$

Denoting $\overline{\pi_0}(\lambda) = \overline{\pi_0}^G(\lambda)$, this may be written as

$$\sqrt{m}\left(\mathsf{FDP}_m(\mathcal{T}(\hat{\mathbb{G}}_m)) - \frac{\pi_0}{\overline{\pi_0}(\lambda)}\alpha\right) \rightsquigarrow \frac{\pi_0}{\overline{\pi_0}(\lambda)}\alpha\left(\frac{\mathbb{Z}_0(\tau^\star)}{\tau^\star} + \frac{\mathbb{Z}(\lambda)}{1-G(\lambda)}\right)$$

For the calculation of variance, it should be noted that $\operatorname{Var}\mathbb{Z}_0(\tau^\star) = \tau^\star(1-\tau^\star)$ and

$$\begin{aligned}\mathbb{E}\left[\mathbb{Z}_0(\tau^\star)\mathbb{Z}(\lambda)\right] &= \mathbb{E}\left[\mathbb{Z}_0(\tau^\star)\mathbb{Z}_0(\lambda)\right] \\ &= \tau^\star_\wedge\lambda\left(1-\tau^\star\vee\lambda\right),\end{aligned}$$

which concludes the proof. □

Procedure BKY06(λ)[9]. According to Proposition 2.7.8, we have $F(\mathcal{U}(F,\beta)) = \mathcal{U}(F,\beta)/\beta$ for any $F \in D[0,1]$, so that the level function of procedure BKY06 may be written as

$$\mathcal{A}(F) = \frac{\alpha(1-\lambda)}{1-\mathcal{U}(F,\lambda)/\lambda}\,.$$

We now prove the Hadamard differentiability of the level function of procedure BKY06(λ) under Condition C.12.

PROPOSITION 2.7.17. *For $\lambda \in [0,1)$ and $F \in D[0,1]$, let $\mathcal{A}(F) = \frac{\alpha(1-\lambda)}{1-\mathcal{U}(F,\lambda)/\lambda}$. Under Condition C.12, $\mathcal{A}$ is Hadamard-differentiable at G, tangentially to $C[0,1]$, with derivative*

$$\dot{\mathcal{A}}_G(H) = \frac{\mathcal{A}(G)^2}{\alpha(1-\lambda)}\frac{H(\mathcal{U}(G,\lambda))}{1/(\alpha(1-\lambda)) - g(\mathcal{U}(G,\lambda))}\,.$$

PROOF OF PROPOSITION 2.7.17. As Condition C.12 holds, Condition C.4 holds for the BH95 procedure at level λ: $\mathcal{U}$ is Hadamard-differentiable with respect to its first variable at (G,λ), tangentially to $C[0,1]$, with derivative $\nabla_F \mathcal{U}_{G,\lambda}$ defined for any $H \in C[0,1]$ by

$$\nabla_F \mathcal{U}_{G,\lambda}(H) = \frac{H(\mathcal{U}(G,\lambda))}{\frac{\partial r}{\partial u}(\lambda,\mathcal{U}(G,\lambda)) - g(\mathcal{U}(G,\lambda))}\,.$$

As the rejection curve of $\mathcal{U}$ is Simes' line, we have $\frac{\partial r}{\partial u}(\lambda,\mathcal{U}(G,\lambda)) = \frac{1}{\lambda}$. As $\mathcal{A}(F) = \alpha\phi(\mathcal{U}(F,\lambda))$, where $\phi : x \mapsto (1-\lambda)/(1-x/\lambda)$ is derivable for $x \neq \lambda$, with $\phi'(x) = \frac{\lambda(1-\lambda)}{1-x/\lambda}$, the result follows from the chain rule. □

PROOF OF THEOREM 2.4.20. As Condition C.12 holds, this is a direct consequence of Proposition 2.7.17 and Theorem 2.4.12.

□

2.7.4. One- and two-stage adaptive procedures.

PROOF OF THEOREM 2.5.2. As we have assumed that existence Condition C.2 and uniqueness Condition C.3 hold for procedure $\mathcal{T}$, $\tau^\star = \mathcal{T}(G)$ is the only point in $(0,1)$ such that $G(\tau^\star) = c_\alpha(G,\tau^\star)$. Similarly, as existence Condition C.2 holds for procedure $\mathcal{T}_t$ for any $t \in (0,1)$, $\tau(t) = \mathcal{T}_t(G)$ is the only point in $(0,1)$ such that $G(\tau(t))/\tau(t) = c_\alpha(G,t)/t$. Therefore, we have

$$\begin{aligned}
t \leq \tau^\star \quad &\Longleftrightarrow \quad G(t) \geq c_\alpha(G,t) \\
&\Longleftrightarrow \quad \frac{G(t)}{t} \geq \frac{c_\alpha(G,t)}{t} \\
&\Longleftrightarrow \quad \frac{G(t)}{t} \geq \frac{G(\tau(t))}{\tau(t)} \\
&\Longleftrightarrow \quad t \leq \tau(t)
\end{aligned}$$

as $u \mapsto G(u)/u$, is non increasing (due to the concavity of G). As $u \mapsto c_\alpha(G,u)/u$ is non increasing (Condition C.13), we have

$$\begin{aligned} t \le \tau^\star \quad &\iff \quad \frac{c_\alpha(G,\tau^\star)}{\tau^\star} \le \frac{c_\alpha(G,t)}{t} \\ &\iff \quad \frac{G(\tau^\star)}{\tau^\star} \le \frac{G(\tau(t))}{\tau(t)} \\ &\iff \quad \tau(t) \le \tau^\star\,, \end{aligned}$$

and (i) is proved. Let $\lambda \in (0,1)$. If $\lambda \le \tau^\star$, then by (i), the sequence (t_n) is non decreasing, and smaller than $\tau^\star$. It therefore converges to a limit $\ell \in [\lambda, \tau^\star]$, such that $\tau(\ell) = \ell$, that is, $G(\ell) = c_\alpha(G,\ell)$. The uniqueness Condition C.3 ensures that $\ell = \tau^\star$. Conversely, if $\lambda \ge \tau^\star$, then, by (i), the sequence (t_n) is non increasing, greater than $\tau^\star$, and thus converges to $\ell \in [\tau^\star, \lambda]$ such that $\tau(\ell) = \ell$, and we also have $\ell = \tau^\star$. □

Sto02 **and** FDR08**.**

Proof of Theorem 2.5.4. As existence Condition C.4 holds, existence Condition C.11 also holds for procedure Sto02(t), for any $t \in (0,1)$. Therefore, Theorem 2.4.15 ensures that the asymptotic threshold $\tau(t)$ of procedure Sto02(t) is positive, and satisfies $G(\tau(t)) = \frac{\tau(t)}{\alpha}\overline{\pi_0}\,(t)$, where $\overline{\pi_0}\,(u) = \frac{1-G(u)}{1-u}$.

As uniqueness Condition C.3 and existence Conditions C.4 and C.7 hold, Theorem 2.4.7 ensures that the asymptotic threshold $\tau^\star$ of procedure FDR08 satisfies $\tau^\star \in (0,\kappa)$, and satisfies $G(\tau^\star) = f_\alpha(\tau^\star)$, where $f_\alpha : u \mapsto u/(\alpha + (1-\alpha)u)$ is the rejection curve of the FDR08 procedure. For any fixed $\lambda \in (t \wedge \kappa, 1)$, the FDR08$(\lambda)$ procedure defined by the capped threshold function

$$\mathcal{T}_\lambda(F) = \sup\{u \in [0,\lambda], F(u) \ge f_\alpha u\}$$

also has asymptotic threshold $\tau^\star$ according to Proposition 2.4.6, as $\lambda \ge \kappa$. For $F \in D[0,1]$ and $u \in [0,\lambda]$, let

$$c_\alpha(F,u) = \frac{u}{\alpha}\,\frac{1-F(u)}{1-u}\,.$$

As G is concave, $u \mapsto \frac{1-G(u)}{1-u}$ is non increasing, so that c_α fulfills the requirements of Condition C.13. Therefore, as $F(u) \ge f_\alpha(u)$ may be written as $F(u) \ge c_\alpha(F,u)$, the result follows from the application of Theorem 2.5.2 to procedures FDR08(λ) and Sto02(t). □

BKY06(λ) **and** BR08(λ)**.**

Proof of Theorem 2.5.6. As uniqueness Condition C.3 and existence Conditions C.8 and C.9 hold, Theorem 2.4.10 ensures that the asymptotic threshold $\tau^\star$ of procedure BR08(λ) is the unique point in $(0,\lambda)$ such that $G(\tau^\star) = \frac{\tau^\star}{\alpha(1-\lambda)+\tau^\star}$, because the rejection curve b_α^λ of the BR08 procedure equals $\frac{u}{\alpha(1-\lambda)+u}$ for $u \le \lambda$.

Existence Condition C.8 also ensures that $\tau(t)$ exists for any $t \le \lambda$. For $F \in D[0,1]$ and $u \in [0,\lambda]$, let

$$c_\alpha(F,u) = \frac{u}{\alpha}\,\frac{1-F(u)}{1-\lambda}\,.$$

As $1 - G$ is non increasing, c_α fulfills the requirements of Condition C.13. Therefore, as $F(u) \geq b_\alpha(u)$ may be written as $F(u) \geq c_\alpha(F, u)$, the result follows from the application of Theorem 2.5.2 to procedures $\mathsf{BR08}(\lambda)$ and $\mathsf{BKY06}(\lambda)$. □

Proof of Corollary 2.5.7. According to the definition of u_λ as the asymptotic threshold of the $\mathsf{BH95}$ procedure at level λ, the asymptotic threshold $\tau^\star$ of procedure $\mathsf{BR08}(\lambda)$ satisfies $\tau^\star \geq u_\lambda$ if and only if $G(\tau^\star) \leq \tau^\star$. According to the definition of the rejection curve b_α^λ of the $\mathsf{BR08}(\lambda)$ procedure, this is equivalent to $\tau^\star/(\alpha(1-\lambda) + \tau^\star) \leq \tau^\star/\lambda$, that is, to $\tau^\star \geq \lambda - \alpha(1-\lambda)$. □

CHAPTER 3

Intrinsic Bounds to Multiple Testing Procedures

Surrealism

Mary Brinig, *B14n135*, 2002

Joan Miró, *Dones, Ocell*, 1973

Contents

3.1. Introduction

Given a possibly large set of observations corresponding either to a null hypothesis $\mathcal{H}_0$, or an alternative hypothesis $\mathcal{H}_1$, several questions are of interest:

(i) a **binary testing problem**: are there any true alternatives?
(ii) an **estimation problem**: how many hypotheses are true alternatives?
(iii) a **selection problem**: which hypotheses are true alternatives?

These three problems have been studied in the framework of *mixture models*: a p-value of the test of the null hypothesis $\mathcal{H}_0$ against the alternative $\mathcal{H}_1$ is associated with each observation, and the distribution of these p-values is modeled as a mixture of a null and an alternative distribution. Sparse and non sparse settings have been investigated. In *sparse* mixture models, the fraction of true alternatives tends to 0, and the dissimilarity between the distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$ increases as the number m of tested hypotheses tend to $+\infty$ [2, 13, 24, 25, 51]. In non-sparse mixture models, all parameters of the model remain fixed as $m \to +\infty$ [7, 34, 35, 89].

The concept of False Discovery Rate (FDR) described in Chapter 1 has been introduced by Benjamini and Hochberg [7] for the selection problem in the fixed mixture model. FDR control and the BH95 procedure have been successfully applied to sparse settings, by Donoho and Jin [24] for the detection problem, and by Abramovich and Benjamini [1], Abramovich et al. [2] and Donoho and Jin [25] for the selection problem, in which it was demonstrated to enjoy remarkable minimax properties.

A natural question is whether there exist constraints on the performance of a given procedure for the detection, estimation or selection problem, or intrinsic limits to these three problems.

Detection. For the detection problem, such limitations have been determined by Ingster [47, 48], Ingster and Suslina [49], Jin [50] in the case of sparse Gaussian mixtures: they have identified a sharp *detection boundary* that separates situations in which the Likelihood Ratio Test (LRT) asymptotically almost surely correctly detects, from situations in which it asymptotically almost surely fails to detect. Donoho and Jin [24] have characterized the detection boundary of several detection procedures in this setting, including the BH95 procedure.

Estimation. Cai et al. [13] demonstrated the existence of an *estimation boundary* for sparse Gaussian mixtures, which characterizes situations in which the fraction of true alternatives can be consistently estimated.

Selection. For the selection problem, the criticality phenomenon described in Chapter 1 illustrates the existence of a possibly positive lower bound below which no multiple testing procedure can control pFDR. In such "critical" situations the power of the BH95 procedures converges to 0 in probability [16].

In this chapter, we compare these bounds in the sparse and non sparse settings, in the context of FDR control. We focus on the following questions:

How do the shape and regularity of the distribution functions under $\mathcal{H}_0$ and $\mathcal{H}_1$ drive the performance of FDR controlling procedures in the fixed mixture model ? How can the performances of the BH95 procedure in terms of criticality and detection be characterized for general sparse mixture models ?

This chapter is organized as follows. Section 3.2 introduces generic notation for the definition of the detection, estimation, and selection problems in sparse and non sparse settings. In section 3.3 we illustrate the influence of distribution tails on the criticality phenomenon for the fixed mixture model, and unveil a connection between criticality and identifiability of the fraction of true alternatives, resulting in a upper bound for the power of FDR controlling procedures. Section 3.4 is devoted to the estimation problem for the fixed mixture model, that is, the estimation of the fraction π_0 of true null hypotheses: we demonstrate that non-parametric estimators of π_0 typically have slow convergence rates, due to the poor regularity of the distribution function in a neighborhood of 1; this results in slow rates of convergence for procedures that incorporate an estimator of π_0 in order to yield exact FDR control. In section 3.5 the performances of the BH95 procedure in the sparse setting are studied: we discuss a generalization of the definition of criticality to this setting, and propose and interpretation of the detection boundary of the BH95 procedure in terms of pFDR control. Proofs of the main results are gathered in section 3.6.

3.2. Background and notation

3.2.1. Mixture model. The mixture model we consider is more generic than that of Chapter 2, as we allow the distribution of the test statistics and the corresponding p-values to depend on the number m of hypotheses tested. More specifically, for $i \in \{1 \dots m\}$, we let $Y_i = 0$ if hypothesis i is drawn from the null hypothesis $\mathcal{H}_0$, and $Y_i = 1$ if it is drawn from the alternative $\mathcal{H}_1$; X_i denotes the corresponding test statistic. We assume that the random variables $(X_i, Y_i)_{1 \leq i \leq m}$ are identically independently distributed: Y_i is a Bernoulli random variable with success probability ε_m, where ε_m is the unknown proportion of true alternatives; the conditional distribution of X_i given Y_i is denoted by $F_1^{(m)}$ if $Y_i = 1$ and $F_0^{(m)}$ if $Y_i = 0$. The marginal distribution of each X_i is thus

$$F^{(m)} = (1 - \varepsilon_m) F_0^{(m)} + \varepsilon_m F_1^{(m)} .$$

The corresponding densities are denoted by $f_0^{(m)}$, $f_1^{(m)}$ and $f^{(m)} = (1 - \varepsilon_m) f_0^{(m)} + \varepsilon_m f_1^{(m)}$.

This model may be equivalently formulated in terms of p-values rather than test statistics. In our setting, the p-values are uniform on $[0, 1]$ under $\mathcal{H}_0$: we let $G_0^{(m)}(x) = x$ for $0 \leq x \leq 1$. Letting $G_1^{(m)}$ and $g_1^{(m)}$ denote the distribution function and density function of the p-values under $\mathcal{H}_1$, the marginal distribution function and density of the p-values under the mixture are given by $G^{(m)} = (1 - \varepsilon_m) G_0^{(m)} + \varepsilon_m G_1^{(m)}$ and $g^{(m)} = (1 - \varepsilon_m) + \varepsilon_m g_1^{(m)}$.

We essentially focus on *location problems*, that is problems in which the distribution of the test statistic under $\mathcal{H}_1$ is a shift from that of the test

statistic under $\mathcal{H}_0$: $F_1^{(m)} = F_0^{(m)}(\cdot - \mu_m)$ for some amplitude parameter $\mu_m > 0$.

The mixture model will be considered in two different settings. In the *sparse setting*, we let ε_m converge to 0 and the distance between $\mathcal{H}_0$ and $\mathcal{H}_1$ (typically measured by μ_m in the above location model) grow to $+\infty$ as the number m of tested hypotheses tends to $+\infty$. In the *fixed setting*, all parameters of the mixture model are fixed. In order to alleviate notation, the superscript m will be omitted in this setting. The fraction of $1-\varepsilon$ of true null hypotheses will be denoted by π_0.

3.2.2. Three multiple testing problems. We now give a more precise definition of the detection, estimation, and selection problems. The first problem is the test of the null hypothesis $\mathcal{H}_0^D$ that the proportion ε_m is 0, against the alternative $\mathcal{H}_1^D$ that it is positive. It is a binary testing problem:

$$\begin{aligned}\mathcal{H}_0^D: &\quad (X_i)_i \overset{\text{iid}}{\sim} F_0^{(m)}\\ \mathcal{H}_1^D: &\quad (X_i)_i \overset{\text{iid}}{\sim} (1-\varepsilon_m)F_0^{(m)} + \varepsilon_m F_1^{(m)}\end{aligned}$$

The second problem is to estimate the proportion ε_m. The third problem is to select a subset of the m tested hypotheses to be rejected. In this paper the selection procedures (or multiple testing procedures) we consider determine a threshold on the observed p-values or test statistics, and reject all hypotheses which are less significant than this threshold.

The concept of False Discovery Rate (FDR) has been introduced by Benjamini and Hochberg [7] in the context of the selection problem. A related quantity is the positive false discovery rate (pFDR), that is the conditional expectation of the False Discovery Proportion (FDP) given that at least one discovery is made:

$$\mathsf{pFDR}(t) = \frac{(1-\varepsilon_m)t}{G(t)}\,.$$

FDR and pFDR are tightly connected as

$$\mathsf{FDR}_m(t) = \mathsf{pFDR}(t)\mathbb{P}(R(t) > 0)\,,$$

where $R(t)$ denotes the number of rejections at threshold t. In particular FDR and pFDR asymptotically equivalent for procedures with fixed rejection regions because $\mathbb{P}(R(t) > 0) \to 1$, as shown by Storey et al. [93].

3.2.3. Bounds on multiple comparison problems.

Detection in a sparse setting. The Gaussian detection problem in which the test statistics are distributed as $\mathcal{N}(0,1)$ under $\mathcal{H}_0$ and $\mathcal{N}(\mu_m, 1)$ under $\mathcal{H}_1$ has been studied by [47, 48, 50]. In this setting, the Likelihood Ratio Test (LRT) is the most powerful procedure for testing $\mathcal{H}_0^D$ against $\mathcal{H}_1^D$. When $\varepsilon_m = m^{-\beta}$ for some $\beta \in (1/2, 1)$, and $\mu_m = \sqrt{2r\log(m)}$, there is a threshold effect for the LRT: there exists a *detection boundary* $(\beta, \rho^\star(\beta))$ such that the sum of Type I and Type II errors tends to 0 or 1 depending on whether $r > \rho^\star(\beta)$ or $r < \rho^\star(\beta)$.

[24] discuss the performance of several testing procedures in this Gaussian setting, and for other specific location problems. In particular, the BH95 procedure can be indirectly used to solve the detection problem, by rejecting $\mathcal{H}_0^D$ if the BH95 procedure makes any discovery. [7] showed that this testing

procedure has level $\leq \alpha$ for rejecting the joint null hypothesis $\mathcal{H}_0^D$. [24] show that it is asymptotically optimal in the sparse region $3/4 < \beta < 1/2$, and that in the not-too-sparse region $1/2 < \beta < 3/4$, it is outperformed by a procedure named *higher criticism* that was originally proposed by [102], and achieves quasi-optimal detection boundary.

Estimation in a sparse setting. [13] focus on sparse *Gaussian* mixtures, and prove that the region where the detection problem can be solved coincides with the region where the fraction of true alternatives can be consistently estimated. They derive minimax convergence rates in this region, and propose an estimation procedure that achieve the optimal rate. [61] focus on a family of estimators and derive the corresponding estimation boundary; their result is valid for any sparse mixture.

Criticality of the selection problem in a fixed setting. [16] noticed that depending on the distribution function G of the p-values, $\mathsf{pFDR}(t)_{t>0}$ may be bounded away from 0, giving rise to a phenomenon that he called *criticality*: no selection procedure can achieve pFDR smaller than $\beta^\star = \inf_{t>0} \mathsf{pFDR}(t)$. By definition of $\beta^\star$, this phenomenon is intrinsic to the selection problem, not to the procedure.

Criticality reveals an interesting range of situations in which FDR and pFDR are not asymptotically equivalent anymore [18]: given a multiple comparison problem such that $\beta_* > 0$, any procedure that controls FDR at level $\alpha < \beta_*$ necessarily makes no rejection with positive probability:

$$\mathbb{P}(R = 0) = 1 - \frac{\mathsf{FDR}_m}{\mathsf{pFDR}} \geq 1 - \frac{\alpha}{\beta_*} > 0\,.$$

The criticality of the $\mathsf{BH95}$ procedure is investigated in [16]. It is characterized by a threshold value $\alpha^\star$ of the target FDR level, which separates the subcritical case $\alpha > \alpha^\star$ from the supercritical case, $\alpha < \alpha^\star$. In the subcritical case, pFDR and FDR are asymptotically equivalent, and the $\mathsf{BH95}$ procedure has asymptotically positive power.

$$\exists \rho^\star > 0,\ \frac{R_m}{m} \xrightarrow[m\to+\infty]{(P)} \rho^\star\,.$$

In the supercritical case, pFDR and FDR are not asymptotically equivalent anymore, and the power of the $\mathsf{BH95}$ procedure converges to 0 in probability.

3.3. Criticality, distribution tails and identifiability

In this section we consider the fixed mixture model described in section 3.2. π_0 denotes the unknown proportion of true null hypotheses. We assume that the likelihood ratio between $\mathcal{H}_0$ and $\mathcal{H}_1$ is non-decreasing: this assumption means that the alternative hypothesis dominates the null, or, equivalently, that the distribution function G_1 of the p-values under the alternative is concave.

We begin by giving a characterization of criticality for the $\mathsf{BH95}$ procedure, in terms of the behavior of the density g_1 under the alternative at 0, and discuss its application to location models and to the case of Student test statistics (section 3.3.1). Then, noting that the identifiability of π_0 is related to the behavior of g_1 in a neighborhood of 1, we point out a

connection between criticality and identifiability of π_0 for location models (section 3.3.2).

3.3.1. Criticality and tails of test statistics. The definition of the critical value of the BH95 procedure proposed by [16] in this setting, that is, the minimum pFDR that may be attained by this procedure, can be written as follows.

DEFINITION 3.3.1 (Critical value of the BH95 procedure). *Let G be the distribution function of the p-values under the mixture of $\mathcal{H}_0$ and $\mathcal{H}_1$ with proportion $\pi 0$ of true nulls. The critical value of the* BH95 *procedure for the multiple comparison of $\mathcal{H}_0$ against $\mathcal{H}_1$ is*

$$\alpha^\star = \inf_{u\in[0,1]} \frac{u}{G(u)}$$

As G is positive and $G(1) = 1$, $\alpha^\star \in [0, 1]$. The cumulative distribution function of the one-sided p-value under the alternative distribution are given by:

$$\begin{aligned} G_1(u) &= 1 - F_1(-F_0^{-1}(u)) \\ g_1(u) &= \tfrac{f_1}{f_0}\left(-F_0^{-1}(u)\right) \end{aligned}$$

As a consequence G_1 (or, equivalently, G) is concave if and only if the likelihood ratio $\frac{f_1}{f_0}$ of the test statistics is non-decreasing. In this case, $u \mapsto \frac{u}{G(u)}$ is non-decreasing on $[0, 1]$. Thus the critical value $\alpha^\star$ is simply given by

$$\alpha^\star = \lim_{u\to 0} \frac{u}{G(u)}$$

Criticality therefore only depends on the behavior of $\frac{G(u)}{u}$ at 0. As $\lim_{u\to 0} F_0^{-1}(u) = -\infty$, and as the likelihood ratio $\frac{f_1}{f_0}$ is non-decreasing, criticality only depends on the boundedness of $\frac{f_1}{f_0}$ as $t \to +\infty$, as shown by the following characterization.

PROPOSITION 3.3.2 (Criticality when $\frac{f_1}{f_0}$ is non-decreasing). (i) *If $\frac{f_1}{f_0}$ is bounded as $t \to +\infty$, then g_1 has a finite limit at 0 (which we denote $g_1(0)$). A criticality phenomenon occurs, and the critical value is explicitly given by*

$$\alpha^\star = \frac{1}{g(0)} = \frac{1}{\pi_0 + (1-\pi_0)g_1(0)}$$

(ii) *If $\lim_{t\to+\infty} \frac{f_1}{f_0}(t) = +\infty$, then $\lim_{u\to 0} \frac{G(u)}{u} = +\infty$, and $\alpha^\star = 0$. There is no criticality phenomenon, and all target* FDR *levels are attainable.*

We demonstrate that the criticality phenomenon only occurs for heavy-tailed distribution such as the Laplace (double-exponential) distribution, whereas it does not occur for distributions with lighter tails, such as the Gaussian distribution.

EXAMPLE 3.3.3 (Gaussian test statistics). Assume that the test statistics are $\sim \mathcal{N}(0,1)$ under the null hypothesis, and $\sim \mathcal{N}(\mu,1)$ under the alternative (with $\mu > 0$). The likelihood ratio is thus simply given by

$$\begin{aligned} \frac{f_1}{f_0}(t) &= \exp\left(-\frac{1}{2}(t-\mu)^2 + \frac{1}{2}t^2\right) \\ &= \exp\left(-\frac{\mu^2}{2} + \mu t\right) \end{aligned}$$

As this likelihood ratio is non decreasing and not bounded as $t \to +\infty$, the Gaussian location problem is not critical: $\alpha^\star = 0$.

We now investigate the case of Laplace (bilateral exponential) test statistics, which has heavier tails than the Gaussian distribution; this results in a criticality phenomenon.

EXAMPLE 3.3.4 (Laplace test statistics). Assume that the density of the test statistics is $f_0 : t \mapsto \frac{1}{2}e^{-|t|}$ under the null hypothesis, and $f_1 : t \mapsto \frac{1}{2}e^{-|t-\mu|}$ under the alternative. The likelihood ratio of the model is given by $\frac{f_1}{f_0}(t) = e^{|t|-|t-\mu|}$, that is, $\frac{f_1}{f_0}(t) = e^{2t-\mu}$ if $t \leq \mu$, and e^μ if $t > \mu$. The likelihood ratio of this model is therefore bounded, which results in a positive critical value given by

$$\alpha^\star = \frac{1}{\pi_0 + (1-\pi_0)e^\mu}$$

To illustrate this phenomenon, we note (see proof in section 3.6.1) that the distribution function of the p-values under $\mathcal{H}_1$ is given by

$$\begin{cases} ue^\mu & \text{if } 0 \leq u \leq \frac{e^{-\mu}}{2} \\ 1 - \frac{1}{4u}e^{-\mu} & \text{if } \frac{e^{-\mu}}{2} \leq u \leq \frac{1}{2} \\ 1 - (1-u)e^{-\mu} & \text{if } u \geq \frac{1}{2} \end{cases}$$

The distribution function of the p-values under a Laplace mixture with proportion π_0 of true nulls is linear between 0 and $\frac{e^{-\mu}}{2}$, with slope $\pi_0 + (1-\pi_0)e^\mu = 1/\alpha^\star$. Figure 1 illustrates the criticality phenomenon for $\mu = 1$ (left) and $\mu = 2$ (right), for different values of $\varepsilon = 1 - \pi_0$.

The critical value $\alpha^\star$ depends both on the non-centrality parameter μ and the proportion π_0 of true nulls. As $\alpha^\star$ is a decreasing function of μ and π_0, the knowledge of a lower bound on μ and $1 - \pi_0$ can be translated into a lower bound on $\alpha^\star$. For example, suppose that we know that $\mu \leq 2$, and $\pi_0 \geq 0.75$. Then $\alpha^\star \geq \frac{1}{0.75+0.25e^2} = 0.385$, which means that even though π_0 and μ are not exactly known, we know that the BH95 procedure applied in this setting with any target FDR level $\alpha < 0.385$ has power tending to 0 as the number of tested hypotheses tends to $+\infty$.

In the case when π_0 is totally unknown, for a given lower bound of μ, there is still a positive minimal $\alpha^\star$, namely $\underline{\alpha^\star} = e^{-\mu}$, which corresponds to the limit case when all hypotheses come from the alternative (that is, $\pi_0 = 0$ and $G = G_1$). This limit case is represented in figure 1. For example, with $\mu \leq 2$, then $\underline{\alpha^\star} = 0.135$, whatever π_0.

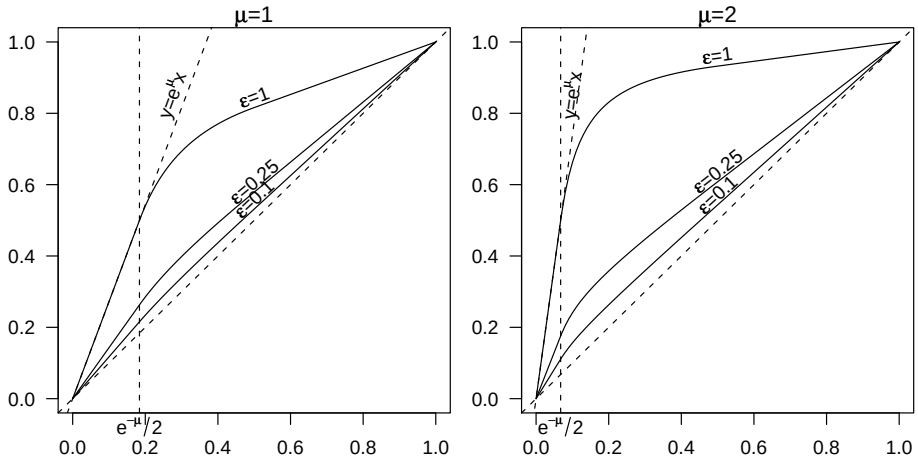


FIGURE 1. *Criticality of the Laplace multiple comparison problem: distribution function of the* p*-values under a Laplace mixture. Left,* $\mu = 1$*; right,* $\mu = 2$*.* $\varepsilon = 1 - \pi_0$*.*

Gaussian and Laplace distributions can be viewed as special cases of a more general class of distribution, introduced by Subbotin [94]:

EXAMPLE 3.3.5 (Subbotin test statistics). Assume that the test statistics under the alternative are given by

$$f_1(t) = \frac{1}{C_\gamma} \exp\left(-\frac{|t-\mu|^\gamma}{\gamma}\right)$$

We focus on $\gamma \geq 1$ because it corresponds to situations in which $\frac{f_1}{f_0}$ is non-decreasing. The Gaussian corresponds to $\gamma = 2$ and the Laplace to $\gamma = 1$. The likelihood ratio of this distribution is given by

$$\begin{aligned} \frac{f_1}{f_0}(t) &= \exp\left(\frac{|t|^\gamma}{\gamma} - \frac{|t-\mu|^\gamma}{\gamma}\right) \\ &= \exp\left(\frac{|t|^\gamma}{\gamma}\left(1 - |1 - \frac{\mu}{t}|^\gamma\right)\right) \end{aligned}$$

As $t \to +\infty$, $|1 - \frac{\mu}{t}|^\gamma \sim 1 - \frac{\gamma\mu}{t}$, so that $\frac{|t|^\gamma}{\gamma}\left(1 - |1 - \frac{\mu}{t}|^\gamma\right) \sim \mu t^{\gamma-1}$, and the behavior of $\frac{f_1}{f_0}(t)$ is driven by the value of γ:

THEOREM 3.3.6 (Subbotin location problem). *Consider the Subbotin location problem with parameter* $\gamma \geq 1$*, and non-centrality* μ *under the alternative. Then:*

(i) $\frac{f_1}{f_0}$ *is non-decreasing;*

(ii) *If* $\gamma > 1$, $\lim_{t\to+\infty} \frac{f_1}{f_0}(t) = +\infty$ *and there is no criticality phenomenon;*

(iii) *If* $\gamma = 1$ *(Laplace),* $\lim_{t\to+\infty} \frac{f_1}{f_0}(t) = e^\mu$ *and there is a positive critical value given by*

$$\alpha^\star = \frac{1}{\pi_0 + (1-\pi_0)e^\mu}\,.$$

Viewing this result, Laplace distributed test statistics appear as borderline cases between criticality-free and criticality-prone situations: the multiple comparison problem is not critical if and only if the tails of the test statistics are lighter than exponential.

Student test statistics. The study of the Student case is motivated by the fact that test statistics are often assumed to have been generated from longitudinal Gaussian observations with unknown variance. We first recall the form of the probability distribution function of the non-central t distribution with k degrees of freedom.

PROPOSITION 3.3.7 (non-central t). *The probability distribution function of non-central t with k degrees of freedom and non centrality parameter δ may be written as*

$$f_1(t) = \frac{\Gamma(k+1)}{2^{\frac{k-1}{2}}\Gamma(\frac{k}{2})\sqrt{k\pi}} \frac{1}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} \exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right] Hh_k\left(-\frac{\delta t}{\sqrt{k+t^2}}\right)$$

where

$$Hh_k(z) = \int_0^{+\infty} \frac{x^k}{k!} e^{-\frac{1}{2}(x+z)^2} dx\,.$$

With this notation the probability distribution function of central t with k degrees of freedom is given by

$$f_0(t) = \frac{\Gamma(k+1)}{2^{\frac{k-1}{2}}\Gamma(\frac{k}{2})\sqrt{k\pi}} \frac{1}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} Hh_k(0)\,,$$

and the likelihood ratio of the model is given by

$$\frac{f_1}{f_0}(t) = \exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right] \frac{Hh_k\left(-\frac{\delta t}{\sqrt{k+t^2}}\right)}{Hh_k(0)}\,.$$

We now demonstrate the criticality of the Student multiple comparison problem.

PROPOSITION 3.3.8 (Student multiple comparison problem). *Consider the multiple comparison problem with test statistics distributed as central Student with k degrees of freedom under $\mathcal{H}_0$, and non central Student with k degrees of freedom and non-centrality parameter δ under $\mathcal{H}_1$. Then:*

(i) *$\frac{f_1}{f_0}$ is non-decreasing*

(ii) *for fixed δ and k, the critical value associated with the multiple comparison of $\mathcal{H}_0$ against $\mathcal{H}_1$ with proportion π_0 or true nulls is*

$$\alpha^\star = \frac{1}{\pi_0 + (1-\pi_0)\frac{Hh_k(-\delta)}{Hh_k(0)}}$$

The fact that $\alpha^\star > 0$ is consistent with the fact that the Student distribution has heavier tails than the Laplace distribution, for which a criticality phenomenon already occurred. Proposition 3.3.8 illustrates how $\alpha^\star$ depends on π_0, δ and k.

Because Student tails become lighter as the number of degrees of freedom increases, an interesting question is whether the criticality phenomenon vanishes if we let the number of degrees of freedom grow to $+\infty$ as the number of tested hypotheses grow to $+\infty$. [17] recently showed that this is the case. Writing the supremum of the likelihood ratio $\frac{Hh_k(-\delta)}{Hh_k(0)}$ as a function $L(k,r)$ of the number k of degrees of freedom, and of the signal to noise ratio $r = \delta/\sqrt{k}$, [17] proved that $L(k,r)$ grows to $+\infty$ whenever $r \to 0$ and $k \to +\infty$, provided that $kr \to +\infty$.

3.3.2. Criticality and identifiability.

Identifiability and achievable FDR control. The question of identifiability of π_0 when $g = \pi_0 + (1-\pi_0)g_1$ is discussed by [35, 55], who demonstrate its importance in the context of FDR control. Recall that the BH95 procedure at level α actually yields $\mathsf{FDR}_m = \pi_0\alpha$: plug-in procedures have therefore been proposed, that apply the BH95 procedure at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0, yielding a larger number of significant hypotheses for the same target FDR level.

When π_0 is not identifiable, such plug-in procedures cannot control FDR at level α exactly: they are bound to remain conservative. In terms of power, all plug-in procedures in the unidentifiable case have power smaller than the Oracle BH95 procedure. Identifiability may thus be interpreted as another intrinsic bound on FDR controlling procedures.

Identifiability and purity. First note that in this setting, the multiple comparison problem is totally determined by (π_0, g_1). [35] make a distinction between the notions of *identifiability of* (π_0, g_1) and *purity of* g_1: purity means that $\inf_{t>0} g_1(t) = 0$, whereas identifiability means that within a given class $\mathcal{F}$ of admissible densities under $\mathcal{H}_1$ for the mixture model, the only way to write $g_1 = (1-b) + bh$, with $0 \le b \le 1$ and $h \in \mathcal{F}$ is to choose $b = 1$ and $h = G_1$.

Under a given parametric assumption, for example, in a Gaussian location model in which test statistics are distributed as standard Gaussian under the null, and are distributed as $\mathcal{N}(\mu, 1)$ for some $\mu \neq 0$, then identifiability does not imply purity [35, 51]. However identifiability and purity are equivalent in our non-parametric estimation setting, as mentioned by [61]: (π_0, g_1) is identifiable if and only if $g_1(1) = 0$.

Criticality is related to the behavior of g_1 at 0, and identifiability is related to the behavior of g_1 at 1. We now point out an important connection between identifiability and criticality for *one-sided p-values* in symmetric location models.

LEMMA 3.3.9 (Density of one-sided location p-values under $\mathcal{H}_1$). *Consider the multiple location problem in which test statistics are distributed as F_0 under $\mathcal{H}_0$, and as $F_1 = F_0(\cdot - \mu)$ under $\mathcal{H}_1$, where $\mu > 0$. Denote by π_0 the proportion of true null hypotheses. Let G_1 be the concave distribution function of* one-sided *p-values under the alternative and g_1 be the corresponding density function. Let $x_\mu = F_0(F_0^{-1}(x) + \mu)$ for any $x \in (0,1)$.*

(i)
$$\forall x \in (0,1), x_\mu > x$$

If F_0 is symmetric, that is, if $F_0(1-x) = F_0(x)$ for any $x \in \mathbb{R}$, then

(ii)
$$\forall x \in (0,1),\ g_1(1-x_\mu) = \frac{1}{g_1(x)}$$

(iii)
$$\forall x \in (0,1),\ g_1(1-x)g_1(x) \leq 1$$

THEOREM 3.3.10 (Criticality and identifiability for one-sided p-values). *Consider the multiple location problem in which test statistics are distributed as F_0 under $\mathcal{H}_0$, and as $F_1 = F_0(\cdot - \mu)$ under $\mathcal{H}_1$, where $\mu > 0$ and F_0 is symmetric. Assume that* one-sided p-values *are computed. Denote by π_0 the proportion of true null hypotheses, and $\alpha^\star$ be the critical value of this multiple comparison problem. Then $\alpha^\star = 0$ if and only if π_0 is identifiable.*

3.4. Estimation of π_0

In the preceding section we studied the connections between criticality, non identifiability of π_0, and the distribution of the test statistics under the null and the alternative hypotheses, summarized by the behavior of the p-value density under the alternative hypothesis. Criticality and non identifiability of π_0 were shown to induce limitations to the intrinsic power of the BH95 procedure, and of plug-in procedures, that consist in applying the BH95 procedure at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0.

In this section we focus on the problem of estimation of π_0 in the non parametric model $g = \pi_0 + (1-\pi_0)g_1$. We assume that π_0 is identifiable and thus may be estimated consistently; that is, we assume that $g_1(1) = 0$ by the preceding section. If this is not the case the results stated here apply to the identifiable part of π_0, which is defined by

$$\overline{\pi_0} = \pi_0 + (1-\pi_0)g_1(1)\,.$$

We begin by proving that the convergence rates of consistent non parametric estimators of π_0 are also related to g_1, the p-value density under the alternative hypothesis, through the regularity of g_1 at 1 (section 3.4.1). Then we investigate the consequences of this property in terms of FDR control, by showing that the convergence rate of FDR controlling procedures that incorporate such an estimator $\widehat{\pi_0}$ is in turn determined by the convergence rate of $\widehat{\pi_0}$ (section 3.4.2). Finally we illustrate the practical consequences of this result by demonstrating that g_1 is typically not regular at 1, even for the Gaussian location model (section 3.4.3).

3.4.1. Convergence rate of $\widehat{\pi_0}$ and regularity of g. We begin by illustrating the connection between convergence rate of $\widehat{\pi_0}$ regularity of g at 1 on a few examples.

Known estimators with convergence rates. To the best of our knowledge, the only non-parametric estimators of π_0 for which convergence rates have been established are those proposed by [89], [97] and [42]. The use of this estimators in the context of multiple testing problems is discussed by [35].

EXAMPLE 3.4.1 (Storey's estimator). Adapting a method originally proposed by [80], [89] defined $\widehat{\pi_0}(\lambda) = \frac{1-\hat{\mathbb{G}}_m(\lambda)}{1-\lambda}$ for $0 \leq \lambda < 1$. As a smooth

functional of the empirical distribution of the p-values, this estimator has the following asymptotic distribution:

$$\sqrt{m}\left(\widehat{\pi_0}(\lambda) - \frac{1-G(\lambda)}{1-\lambda}\right) \rightsquigarrow \mathcal{N}\left(0, \frac{G(\lambda)(1-G(\lambda))}{(1-\lambda)^2}\right)$$

It converges at the parametric rate $1/\sqrt{m}$, and it is asymptotically biased because $\frac{1-G(\lambda)}{1-\lambda} > \pi_0$ for $\lambda < 1$.

EXAMPLE 3.4.2 (Confidence envelopes for the density). [42] derived a finite sample confidence envelope for a monotone density. Assuming that G is concave and that g is Lipschitz in a neighborhood of 1, the resulting estimator converges to π_0 at rate $\left(\frac{\log m}{m}\right)^{-1/3}$.

EXAMPLE 3.4.3 (Spacings-based estimator). [97] proposes an estimator of the minimum of an unknown density based on the distribution of the spacings between observations: he first estimates the location of the minimum, and then the density at this point. Assuming that at the value at which the density g achieves its minimum, g and $\dot{g}$ are null, and $\ddot{g}$ is bounded away from 0 and $+\infty$ and Lipschitz, this estimator converges at a rate slightly slower than $m^{-2/5}$ to the true minimum. In our framework the minimum is necessarily achieved at 1 because g is non-increasing. Thus the first step may be omitted, and the Lipschitz condition becomes unnecessary.

Lower bounds on non-parametric convergence rates. The consistent estimators proposed by [42] and [97] illustrate the fact that the more regular g is assumed to be at 1, the faster convergence rates can be obtained. As $\pi_0 = g(1)$, it seems natural to try to estimate π_0 using *kernel estimators of a density at a point*. We give an explicit connection between the convergence rate of these estimators and the regularity of the density at the point of interest.

A kernel of order $\ell \in \mathbb{N}$ is a function $K : \mathbb{R} \to \mathbb{R}$ such that the functions $u \mapsto u^j K(u)$ are integrable for any $j = 0 \dots l$, and verify $\int_{\mathbb{R}} K = 1$, and $\int_{\mathbb{R}} u^j K(u) = 0$ for $j = 1 \dots \ell$.

DEFINITION 3.4.4 (Kernel estimator of a density). *The kernel estimator of a density g at the point p_0 based on m independent, identically distributed observations $P_1, \dots P_m$ from g is defined by*

$$\hat{g}(p_0) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{P_i - p_0}{h}\right),$$

where $h > 0$ is called the bandwidth of the estimator and K is a kernel.

[101] lower bounds on the convergence rate of kernel estimators of $g(1)$, depending on the regularity of g at 1. If g is k times differentiable at 1, with $g^{(k)}(1) \neq 0$, considering a kernel estimator $\hat{g}(1)$ associated with a k^{th} order kernel and fixed bandwidth h, the asymptotic variance of $\hat{g}(1)$ is of the order of $\frac{1}{mh}$, and the asymptotic bias of $\hat{g}(1)$ is of the order of h^k.

Therefore, $\hat{g}(1)$ is asymptotically biased due to the positive bandwidth h. It is possible to obtain a consistent estimator of $g(1)$ by letting h go to 0 as $m \to +\infty$. The exact risk of the corresponding estimator minimizes

the Mean Squared Error of the estimator; it is thus obtained by balancing asymptotic bias and variance, as shown by the following Proposition.

PROPOSITION 3.4.5 (Optimal bandwidth — k^{th} order kernel estimator [101]). *Assume that g is k times differentiable at 1, with $g^{(k)}(1) \neq 0$. Let $\hat{g}(1)$ be a kernel estimator associated with a k^{th} order kernel. The optimal bandwidth for $\hat{g}(1)$ in terms of Mean Squared Error is of the order of $m^{-(2k+1)}$. The corresponding estimator converges to $g(1)$ at rate $m^{-\frac{k}{2k+1}}$.*

As a consequence, the convergence rate of the optimal kernel estimator of $g(1)$ directly depends on the regularity k of g at 1.

Storey's estimator. The estimator proposed by [89] is a kernel estimator with asymmetric rectangular kernel of order 1, and bandwidth $1-\lambda$. It converges at the parametric rate $1/\sqrt{m}$, and it is asymptotically biased for $\lambda < 1$. In order to make this estimator consistent we let $h = 1-\lambda$ go to 0 as m goes to $+\infty$. The asymptotic distribution of the corresponding estimator is given by the following Proposition.

PROPOSITION 3.4.6 (Asymptotic distribution of $\widehat{\pi_0}(1-h_m)$). *Let*

$$\widehat{\pi_0}(\lambda) = \frac{1-\hat{\mathbb{G}}_m(\lambda)}{1-\lambda}$$

for $0 < \lambda < 1$. Let h_m be a positive sequence such that $h_m \to 0$ and $mh_m \to +\infty$ as $m \to +\infty$. Then

$$\sqrt{mh_m}\,(\widehat{\pi_0}(1-h_m) - \pi_0) \rightsquigarrow \mathcal{N}(0, \pi_0)\,.$$

Proposition 3.4.6 shows that consistency can be achieved at the price of a reduction of the convergence rate. We calibrate h_m such that the Mean squared error is minimum, in order to balance bias and variance: the larger h_m, the smaller asymptotic variance but the larger asymptotic bias. As $\widehat{\pi_0}(1-h_m)$ is a kernel of order 1 only, the optimal bandwidth cannot be derived from Proposition 3.4.5. However if we further assume that the $k-1$ first derivatives of g at 1 are null, we obtain the same optimal bandwidth.

PROPOSITION 3.4.7 (Optimal bandwidth — Storey's estimator). *Assume that g is k times differentiable at 1, with $g^{(l)}(1) = 0$ for $0 \leq l < k$, and $g^{(k)}(1) \neq 0$. The optimal bandwidth in terms of MSE is given by $h_m(k) = C_k m^{-\frac{k}{2k+1}}$, where C_k is an explicit constant that depends on k, π_0, and $g^{(k)}(1)$. Moreover we have*

(i)
$$\text{MSE}(\widehat{\pi_0}(1-h_m(k)) = \frac{2/C_k}{m^{\frac{k}{2k+1}}}$$

(ii)
$$m^{\frac{k}{2k+1}}\,(\widehat{\pi_0}(1-h_m(k)) - \pi_0) \rightsquigarrow \mathcal{N}(0, \pi_0 C_k)\,.$$

3.4.2. Convergence rate of consistent plug-in procedures. The goal of this section is to connect the convergence rate of a given estimator $\widehat{\pi_0}$ of π_0 to the asymptotic FDR controlling capabilities of a procedure that takes $\widehat{\pi_0}$ into account. We are thus interested in the asymptotic properties

of plug-in procedures, that consist in applying the BH95 procedure at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0.

The False Discovery Proportion (FDP) achieved by a broad class of FDR controlling procedures including the plug-in procedure proposed by [89] with fixed λ has been shown by [64] to converge at the parametric rate $1/\sqrt{m}$ to their asymptotic FDR in the subcritical case. Storey's procedure is shown to have asymptotic FDR smaller than α because $\widehat{\pi_0}(\lambda)$ is not consistent for fixed λ.

In this section we consider any estimator $\widehat{\pi_0}$ that converges to π_0 at a rate $\sqrt{mh_m}$, with $h_m \to 0$. The results of this section therefore cover in particular in the case for the estimator $\widehat{\pi_0}(1-h_m)$ for any $h_m \to 0$. We prove that the convergence rate of the FDP of a plug-in procedure based on $\widehat{\pi_0}$ is also of the order of $\sqrt{mh_m}$.

THEOREM 3.4.8 (Asymptotic FDP for consistent plug-in procedures). *Let $\widehat{\pi_0}$ be any estimator of π_0 with asymptotic distribution given by*

$$\sqrt{mh_m}\,(\widehat{\pi_0} - \pi_0) \rightsquigarrow \mathcal{N}(0, v(\pi_0))$$

for some function v. Consider the plug-in procedure based on $\widehat{\pi_0}$, which applies the BH95 *procedure at level $\alpha/\widehat{\pi_0}$, for any $\alpha > \pi_0\alpha^\star$. The asymptotic distribution of the* FDP *achieved by this procedure is given by*

$$\sqrt{mh_m}\,(\mathsf{FDP} - \alpha) \rightsquigarrow \mathcal{N}\left(0, \frac{\alpha^2}{\pi_0^2} v(\pi_0)\right).$$

This result can be combined with the optimal bandwidth choices proposed in Propositions 3.4.5 and 3.4.7.

COROLLARY 3.4.9 (Asymptotic FDP for optimal bandwidth — k^{th} order kernel). *Assume that g is k times differentiable at 1, with $g^{(k)}(1) \neq 0$. Let $\hat{g}(1)$ be the kernel estimator associated with a k^{th} order kernel with optimal bandwidth given by Proposition 3.4.5. Then the* FDP *of the plug-in procedure that applies the* BH95 *procedure at level $\alpha/\hat{g}(1)$ converges to α at rate $m^{-\frac{k}{2k+1}}$.*

COROLLARY 3.4.10 (Asymptotic FDP for optimal bandwidth — Storey's estimator). *Assume that g is k times differentiable at 1, with $g^{(l)}(1) = 0$ for $0 \leq l < k$, and $g^{(k)}(1) \neq 0$, and let $\widehat{\pi_0}^\star = \widehat{\pi_0}(1 - h_m(k))$, where $\widehat{\pi_0}(\lambda)$ is Storey's estimator for fixed $\lambda \in [0,1)$, and $h_m(k) = C_k m^{-\frac{k}{2k+1}}$ the optimal bandwidth defined by Proposition 3.4.7. Then the asymptotic* FDP *of the plug-in procedure that applies the* BH95 *procedure at level $\alpha/\widehat{\pi_0}^\star$ is given by*

$$m^{\frac{k}{2k+1}}\,(\mathsf{FDP} - \alpha) \rightsquigarrow \mathcal{N}\left(0, \frac{\alpha^2 C_k}{\pi_0}\right).$$

3.4.3. Regularity of g_1 at 1. These results motivate the study of the regularity of g at 1. As $g = \pi_0 + (1-\pi_0)g_1$, this is equivalent to the study of the regularity of g_1 at 1. It turns out that even for Gaussian test statistics, this regularity is quite poor.

PROPOSITION 3.4.11 (One-sided Gaussian location problem). *Consider the case when the test statistics follow $\mathcal{N}(0,1)$ under the null hypothesis, and*

$\mathcal{N}(\mu,1)$, under the alternative, where $\mu>0$. Let g_1 be the density function of the corresponding one-sided p-values. Then, as $h\to 0$, there exists a constant C such that

$$g_1(1-h)\leq C\exp\left(-\frac{\mu^2}{2}-\sqrt{2\log(1/h)}\right)$$

COROLLARY 3.4.12. *g_1 is not differentiable at 1.*

PROOF. Recall that $g_1(x)=\frac{f_1}{f_0}\left(-F_0^{-1}(x)\right)$, with $f_0(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $f_1(x)=f_0(x-\mu)$, and $F_0^{-1}(x)=\Phi^{-1}(x)=-\Phi^{-1}(1-x)$. Thus,

$$g_1(x)=\exp\left(-\frac{\mu^2}{2}-\mu\Phi^{-1}(x)\right).$$

Write $x=1-h$, with $h\to 0$. We have $g_1(1-h)=\exp\left(-\frac{\mu^2}{2}+\mu\Phi^{-1}(h)\right)$, with $\Phi^{-1}(h)=-\sqrt{2\log(1/h)}+r(h)$, where r is bounded as $h\to 0$. □

In fact Proposition 3.4.11 implies that for any $\gamma>0$, $g_1(1-m^{-\gamma})$ goes to 0 more slowly than any positive power of $\frac{1}{m}$. For two-sided p-values, the density function g_1 is given by $g_1(x)=\frac{f_1}{f_0}\left(F_0^{-1}(1-x/2)\right)+\frac{f_1}{f_0}\left(F_0^{-1}(x/2)\right)$, so that

$$\begin{aligned} g_1(x) &= e^{-\frac{\mu^2}{2}}\left(\exp\left(-\mu\Phi^{-1}(x/2)\right)+\exp\left(\mu\Phi^{-1}(x/2)\right)\right)\\ &= 2e^{-\frac{\mu^2}{2}}\cosh\left(\mu\Phi^{-1}(x/2)\right). \end{aligned}$$

Hence g_1 is more regular at 1 than for the one-sided case: we have $\dot{g_1}(1)=0$ and $\ddot{g_1}(1)\neq 0$. We can therefore choose $k=2$ in Proposition 3.4.7, and the optimal bandwidth Storey's estimator is $h_m=m^{-1/5}$. The corresponding convergence rate for $\widehat{\pi_0}(1-h_m)$ and the associated FDP is also $m^{-2/5}$. This is rather slow, but still much faster than for the one-sided case. However in the two-sided case $g_1(1)=2e^{-\frac{\mu^2}{2}}$ is positive: thus π_0 is not identifiable, and only the upper bound $\pi_0+2(1-\pi_0)e^{-\frac{\mu^2}{2}}$ can be consistently estimated. These results are illustrated by Figure 2 for the simplest location model: $\mathcal{N}(0,1)$ against $\mathcal{N}(1,1)$.

3.5. FDR control in a sparse setting

We consider the sparse mixture model in which test statistics are distributed as $F_0^{(m)}$ under the null hypothesis $\mathcal{H}_0$, and as $F_1^{(m)}$ under the alternative $\mathcal{H}_1$. We recall that the proportion ε_m of true alternatives is assumed to go to 0 as $m\to+\infty$, hence the term *sparse* mixture model. The marginal distribution of the test statistics is therefore $F^{(m)}=(1-\varepsilon_m)F_0^{(m)}+\varepsilon_m F_1^{(m)}$; the corresponding density functions are denoted by $f_0^{(m)}$, $f_1^{(m)}$, and $f^{(m)}=(1-\varepsilon_m)f_0^{(m)}+\varepsilon_m f_1^{(m)}$.

This mixture model may equivalently be represented in terms of p-values: the marginal distribution of the p-values is denoted by $G^{(m)}=(1-\varepsilon_m)G_0^{(m)}+\varepsilon_m G_1^{(m)}$, where $G_0^{(m)}$ and $G_1^{(m)}$ denote the distribution function of the p-values under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. By definition, $G_0^{(m)}=\mathrm{Id}$. Likewise,

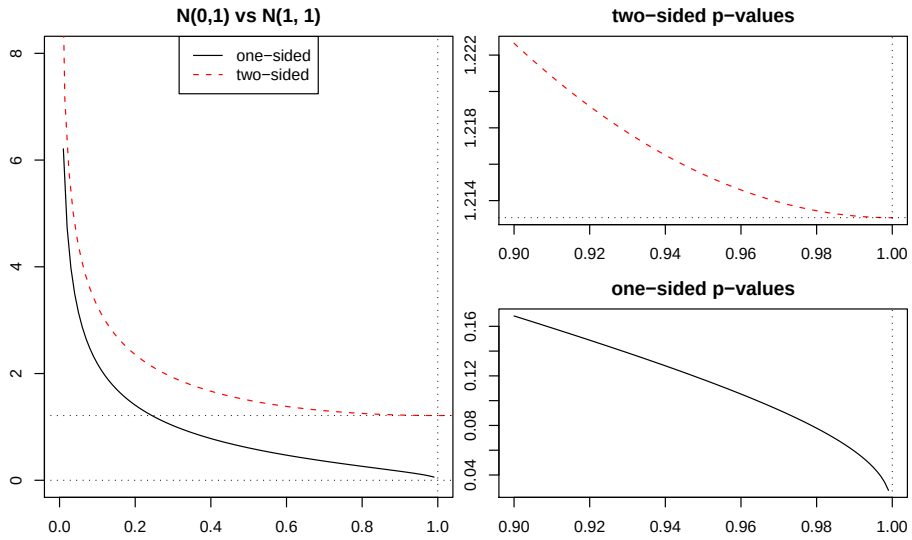


FIGURE 2. *Density of one- and two-sided p-values under the alternative hypothesis for the location model* $\mathcal{N}(0,1)$ *versus* $\mathcal{N}(1,1)$. *Left: one- and two-sided p-values. Top right: zoom on the region* $[0.9,1]$ *for two-sided p-values. Bottom right: zoom on the region* $[0.9,1]$ *for one-sided p-values.*

the density functions of the p-values under $\mathcal{H}_1$ and under the mixture are denoted by $g_1^{(m)}$, and $g^{(m)} = (1-\varepsilon_m) + \varepsilon_m g_1^{(m)}$, respectively.

We begin by studying an extension to this sparse setting of the definition of criticality: we show that the corresponding subcritical and supercritical cases still have different behaviors, even though the interpretation of the subcritical case in terms of power of the BH95 procedure is lost (section 3.5.1). Then we give a sufficient condition for the BH95 procedure to detect in terms of pFDR, and we use this condition to retrieve known detection boundaries for sparse location mixtures (section 3.5.2).

3.5.1. Criticality in a sparse setting. [18] discuss control of the positive False Discovery Excessive Probability pFDEP_α at level $\alpha \in (0,1)$:

$$\mathsf{pFDEP}_\alpha = \mathbb{P}\left(V/R > \alpha | R > 0\right)$$

in a sparse mixture model with unspecified distribution of the test statistics under $\mathcal{H}_0$ and $\mathcal{H}_1$. They propose a definition of criticality for the BH95 procedure in this setting that naturally generalizes Definition 3.3.1.

DEFINITION 3.5.1. *For* $m \in \mathbb{N}$, *let* $\alpha^\star(m)$ *be the critical value of the* BH95 *procedure for the multiple comparison problem parametrized by* ε_m *and* $G_1^{(m)}$. *The critical value of the* BH95 *procedure in this model is defined by*

$$\alpha^\star = \lim_{m\to+\infty} \alpha^\star(m)\,.$$

With this definition, if $\alpha^\star(m) = 0$ for any m, then there is no criticality. This is the case for the Gaussian location model, and more generally for all Subbotin location problems with $\gamma > 1$ studied in section 3.3.

In the fixed setting of sections 3.3 and 3.4, the distinction between the subcritical case ($\alpha > \alpha^\star$) and the supercritical case ($\alpha < \alpha^\star$) had a nice interpretation in terms of asymptotic power of the BH95 procedure [16]: the proportion R_m/m of rejections of the BH95 procedure at level α converged in distribution to a positive value $\rho^\star$ in the subcritical case, whereas it converged in probability to 0 in the supercritical case. Proposition 3.5.2 summarizes the corresponding results in the current sparse setting.

PROPOSITION 3.5.2 (Asymptotic threshold and proportion of rejections of the BH95 procedure). *Consider any multiple testing situation in which the fraction ε_m of true alternatives hypotheses tends to 0 as $m \to +\infty$. For $\alpha \in [0,1)$, consider the mixture model with constant sparsity $\varepsilon_{m'}$. Let $\tau^\star(m')$ and $\rho^\star(m')$ denote the asymptotic threshold and asymptotic proportion of rejections of the* BH95 *procedure at level α in this model. Then*

(i) *$\widehat{\tau}$ and R_m/m converge almost surely to 0 as $m \to +\infty$.*

(ii) *If $\alpha > \alpha^\star$, then $\frac{\widehat{\tau}/m}{\tau^\star(m)} \overset{(P)}{\to} 1$, and $\frac{R_m/m}{\rho^\star(m)} \overset{(P)}{\to} 1$.*

Although Proposition 3.5.2(ii) provides a nice generalization of the behavior of R_m/m in the subcritical case to the current sparse setting, (i) demonstrates that the interpretation in terms of asymptotic power of the BH95 procedure is lost: because $\varepsilon_m \to 0$ as $m \to +\infty$, there is asymptotically no true alternative, so that R_m/m converges almost surely to 0 even in the subcritical case.

3.5.2. Detection and pFDR control. [7] showed that procedure BH_m^D has level $\leq \alpha$ for rejecting the joint null hypothesis $\mathcal{H}_0^D$. We now study the power of the BH_m^D procedure for testing $\mathcal{H}_0^D$ against $\mathcal{H}_1^D$ for generic sparse mixtures. This question has been investigated in [24] in the case of specific location models, namely Gaussian, Chi-squared, and Subbotin location models.

We begin by giving a simple characterization of the detection boundary of sparse mixtures using the Max procedure, which rejects $\mathcal{H}_0^D$ if and only if some p-values are smaller than α/m. This characterization relies on the form of the distribution function $G^{(m)}$ of the p-values under $\mathcal{H}_1^D$. This condition will then provide a sufficient condition for detectability using the BH_m^D procedure, which is then interpreted in terms of pFDR control.

PROPOSITION 3.5.3 (Detection boundary of the Max procedure). *Let $G^{(m)}$ denote the probability distribution function of the* p*-values under $\mathcal{H}_1^D$. Then the condition*

$$\lim_{m \to +\infty} m G^{(m)}\left(\frac{\alpha}{m}\right) \;=\; +\infty$$

is necessary and sufficient for the Max *procedure with level α to have asymptotically full power for separating $\mathcal{H}_1^D$ from $\mathcal{H}_0^D$, that is, to have*

$$\lim_{m \to +\infty} \mathbb{P}_{\mathcal{H}_1^D}\left(\textsf{Max } \textit{rejects } \mathcal{H}_0^D\right) = 1$$

This result provides a connection between the shape of the detection boundary and the characteristics of the mixture model, summarized by the behavior of the marginal distribution function $G^{(m)}$ of the p-values at $\frac{\alpha}{m}$.

Because the BH_m^D procedure does at least as well as the Max procedure, proposition 3.5.3 provides a sufficient condition for detectability using the BH_m^D procedure. This makes it possible to connect the detection boundary of the BH_m^D procedure, pFDR control, and criticality: in the present setting, the pFDR at threshold t may be written as

$$\mathsf{pFDR}_m(t) = \frac{(1-\varepsilon_m)t}{G^{(m)}(t)}$$

THEOREM 3.5.4 (Detection boundary of the BH_m^D procedure). *Let $G^{(m)}$ denote the probability distribution function of the* p*-values under $\mathcal{H}_1^D$. If*

$$\lim_{m\to+\infty} \mathsf{pFDR}_m\left(\frac{\alpha}{m}\right) \quad = \quad 0, \tag{3.5.5}$$

then the BH_m^D procedure with level α has asymptotically full power for separating $\mathcal{H}_1^D$ from $\mathcal{H}_0^D$:

$$\lim_{m\to+\infty} \mathbb{P}_{\mathcal{H}_1^D}\left(\mathsf{BH}_m^D \text{ rejects } \mathcal{H}_0^D\right) = 1\,.$$

Theorem 3.5.4 is interesting because condition (3.5.5) is valid for any mixture model. We now return to the Gaussian, Laplace and Subbotin location problems to demonstrate how naturally and easily the detection boundaries identified in [24] can be derived from condition (3.5.5). We recall that

$$g_1^{(m)}(u) = \frac{f_1}{f_0}\left(-F_0^{-1}(u)\right)\,.$$

The superscript m in the likelihood ratio $\frac{f_1}{f_0}$ is omitted to alleviate notation. Following [24], we choose

$$\varepsilon_m = m^{-\beta},$$

for $\frac{1}{2} < \beta < 1$. The following Proposition gives a sufficient condition for detection using the BH_m^D procedure when $G^{(m)}$ is concave:

PROPOSITION 3.5.6. *Assume that $G^{(m)}$ is concave. If*

$$\lim_{m\to+\infty} \varepsilon_m \frac{f_1}{f_0}\left(-F_0^{-1}\left(\frac{1}{m}\right)\right) = +\infty\,, \tag{3.5.7}$$

then the BH_m^D procedure with level α has asymptotically full power for separating $\mathcal{H}_1^D$ from $\mathcal{H}_0^D$:

$$\lim_{m\to+\infty} \mathbb{P}_{\mathcal{H}_1^D}\left(\mathsf{BH}_m^D \text{ rejects } \mathcal{H}_0^D\right) = 1\,.$$

EXAMPLE 3.5.8 (Gaussian test statistics). In this setting, $-F_0^{-1}\left(\frac{1}{m}\right)$ is of the order of $\sqrt{2\log(m)}$. Following [24], we calibrate μ_m so that the nonzero means are smaller than the largest test statistic under $\mathcal{H}_0$, that is $-F_0^{-1}\left(\frac{1}{m}\right)$:

$$\mu_m = \sqrt{2r\log(m)},$$

with $0 < r < 1$. Recall that for Gaussian test statistics,

$$\frac{f_1}{f_0}(t) \quad = \quad \exp\left(-\frac{\mu_m^2}{2} + \mu_m t\right).$$

Therefore,

$$\begin{aligned}\varepsilon_m \frac{f_1}{f_0}\left(-F_0^{-1}\left(\frac{1}{m}\right)\right) &= m^{-\beta}\exp\left(\frac{\mu_m^2}{2}+\mu_m\sqrt{2\log(m)}\right)\\ &= m^{-r+2\sqrt{r}-\beta}\end{aligned}$$

Hence, by condition (3.5.7), the BH_m^D procedure has full power as soon as

$$r > (1-\sqrt{1-\beta})^2$$

EXAMPLE 3.5.9 (Laplace test statistics). In this setting, we have $-F_0^{-1}\left(\frac{1}{m}\right) = \log(m)$, so we choose

$$\mu_m = r\log(m),$$

with $0 < r < 1$. By definition we have $-F_0^{-1}\left(\frac{1}{m}\right) > \mu_m$, so the likelihood ratio of the model at $-F_0^{-1}\left(\frac{1}{m}\right)$ is $e^{\mu_m} = r$. Therefore

$$\varepsilon_m \frac{f_1}{f_0}\left(-F_0^{-1}\left(\frac{1}{m}\right)\right) = m^{-r+\beta}$$

Hence, by condition (3.5.7), the BH_m^D procedure has full power as soon as

$$r > \beta$$

EXAMPLE 3.5.10 (Subbotin test statistics). Following [61], we choose

$$\mu_m = (\gamma r\log(m))^{\frac{1}{\gamma}},$$

with $0 < r < 1$ for this setting. Recall that

$$\frac{f_1}{f_0}(t) = \exp\left(\frac{|t|^\gamma}{\gamma}\left(1-\left|1-\frac{\mu}{t}\right|^\gamma\right)\right)$$

With $t = -F_0^{-1}\left(\frac{1}{m}\right) > 0$, we have $\frac{t^\gamma}{\gamma} = \log(m)$ and $\frac{\mu_m}{t} = r^{\frac{1}{\gamma}}$, so that

$$\begin{aligned}\varepsilon_m \frac{f_1}{f_0}\left(-F_0^{-1}\left(\frac{1}{m}\right)\right) &= m^{-\beta}\exp\left(\log(m)\left(1-\left(1-r^{\frac{1}{\gamma}}\right)^\gamma\right)\right)\\ &= m^{-\beta+1-\left(1-r^{\frac{1}{\gamma}}\right)^\gamma}\end{aligned}$$

Hence, by condition (3.5.7), the BH_m^D procedure detects correctly as soon as

$$r > \left(1-(1-\beta)^{\frac{1}{\gamma}}\right)^\gamma$$

$\gamma = 1$ and $\gamma = 2$ correspond to the Laplace and Gaussian cases.

3.6. Proofs of main results

3.6.1. Proofs of section 3.3.

Laplace distribution.

LEMMA 3.6.1. *Assume that the pdf of the test statistics is* $f_0 : x \mapsto \frac{1}{2}e^{-|x|}$ *under the null hypothesis, and* $f_1 : x \mapsto \frac{1}{2}e^{-|x-\mu|}$ *under the alternative, with* $\mu > 0$ *(one-sided test). Then*

(i) *The* p*-value is*

$$\begin{aligned} 1 - F_0(x) = \frac{1}{2}e^{(-|x|)} \quad & \text{if } x \geq 0 \\ 1 - \frac{1}{2}e^{(-|x|)} \quad & \text{if } x < 0 \end{aligned}$$

(ii) *The inverse* p*-value is*

$$\begin{aligned} \left(1 - F_0\right)^{-1}(u) = \log\left(\frac{1}{2u}\right) \quad & \text{if } 0 \leq u \leq \tfrac{1}{2} \\ \log\left(2(1-u)\right) \quad & \text{if } \tfrac{1}{2} < u < 1 \end{aligned}$$

(iii) *The cdf of the* p*-values under* $\mathcal{H}_1$ *is*

$$\begin{aligned} G_1(u) = ue^{\mu} \quad & \text{if } 0 \leq u \leq \tfrac{e^{-\mu}}{2} \\ 1 - \frac{1}{4u}e^{-\mu} \quad & \text{if } \tfrac{e^{-\mu}}{2} \leq u \leq \tfrac{1}{2} \\ 1 - (1-u)e^{-\mu} \quad & \text{if } u \geq \tfrac{1}{2} \end{aligned}$$

(iv) *The pdf of the* p*-values under* $\mathcal{H}_1$ *is*

$$\begin{aligned} g_1(u) = e^{\mu} \quad & \text{if } 0 \leq u \leq \tfrac{e^{-\mu}}{2} \\ \frac{1}{4u^2}e^{-\mu} \quad & \text{if } \tfrac{e^{-\mu}}{2} \leq u \leq \tfrac{1}{2} \\ e^{-\mu} \quad & \text{if } u \geq \tfrac{1}{2} \end{aligned}$$

PROOF OF LEMMA 3.6.1. The inverse p-value function directly follows from the p-value function and the pdf of the p-values follows from the cdf, so we only prove (i) (p-value function), and (iii) (cdf of the p-values).

Proof of (i). $1 - F_0(x) = \mathbb{P}(X > x) = \int_{-\infty}^{x} \frac{1}{2}e^{-|t|}It$. Hence for $x < 0$, $1 - F_0(x) = \int_{-\infty}^{x} \frac{1}{2}e^{t}MT = \frac{1}{2}e^{-|x|}$. For $x > 0$, $1 - F_0(x) = \int_{-\infty}^{0} \frac{1}{2}e^{t}at + \int_{0}^{x} \frac{1}{2}e^{-t}dt = 1 - \frac{1}{2}e^{-|x|}$.

Proof of (iii). Let $u \in [0,1]$. The distribution function of the p-values is given by

$$\begin{aligned} G_1(u) &= \mathbb{P}_\mu\left(1 - F_0(x) \leq u\right) \\ &= \mathbb{P}_\mu\left(X \geq (1-F_0)^{-1}(u)\right) \\ &= \int_{1-F_0(u)}^{\mu} f_1(x)dx + \int_{\mu}^{+\infty} f_1(x)dx \\ &= \int_{1-F_0(u)}^{\mu} \frac{1}{2}e^{-|x-\mu|}dx + \frac{1}{2} \end{aligned}$$

For $u < \frac{1}{2}$, $(1-F_0)^{-1}(u) = \log\frac{1}{2u}$ and $(1-F_0)^{-1}(u) \geq \mu \iff u \leq \frac{e^{-\mu}}{2}$.

Hence if $u \leq \frac{e^{-\mu}}{2}$,

$$\begin{aligned} G_1(u) &= \frac{1}{2} - \frac{1}{2}\int_{\mu}^{\log\frac{1}{2u}} e^{-(x-\mu)}dx \\ &= \frac{1}{2} - \frac{1}{2}\left(-e^{-(\log\frac{1}{2u}-\mu)} - (-1)\right) \\ &= ue^{\mu} \end{aligned}$$

If $\frac{e^{-\mu}}{2} < u < \frac{1}{2}$,

$$\begin{aligned} G_1(u) &= \frac{1}{2} + \int_{\log\frac{1}{2u}}^{\mu} \frac{1}{2}e^{(x-\mu)}dx \\ &= \frac{1}{2} + \frac{1}{2}\left(1 - e^{\log\frac{1}{2u}-\mu}\right) \\ &= 1 - \frac{1}{4u}e^{-\mu} \end{aligned}$$

Finally, for $u \geq \frac{1}{2}$, $(1-F_0)^{-1}(u) = \log 2(1-u)$. Thus $(1-F_0)^{-1}(u) \leq \mu \iff u \geq 1 - \frac{e^{\mu}}{2}$, which always holds for $u \geq \frac{1}{2}$ because $\mu > 0$.

Hence for $u \geq \frac{1}{2}$,

$$\begin{aligned} G_1(u) &= \frac{1}{2} + \int_{\log 2(1-u)}^{\mu} \frac{1}{2}e^{(x-\mu)}dx \\ &= \frac{1}{2} + \frac{1}{2}\left(1 - e^{\log 2(1-u)-\mu}\right) \\ &= 1 - (1-u)e^{-\mu} \end{aligned}$$

□

Student distribution. We recall the definition of central and non central t distribution with k degrees of freedom.

DEFINITION 3.6.2 (Student random variable). *Let X be normally distributed with mean δ and variance 1, and Y independently distributed as central χ^2 with k degrees of freedom.*

Then the random variable $T_{k,\delta} = \frac{X}{\sqrt{Y/k}}$ is said to have t distribution (Student distribution) with k degrees of freedom and non-centrality parameter δ.

If $\delta = 0$, $T_{k,0} = T_k$ is simply said to have (central) t distribution with k degrees of freedom.

PROOF OF PROPOSITION 3.3.7. Let $T_{k,\delta} = \frac{Z_\delta}{\sqrt{U/k}}$, where $Z_\delta \sim \mathcal{N}(\delta, 1)$ and $U \sim \chi^2(k)$, with Z_δ and U independent. We have $f_1(t) = \frac{d}{dt}\left(\mathbb{P}\left(T_{k,\delta} \leq t\right)\right) = \frac{d}{dt}\left(\mathbb{P}\left(Z_\delta \leq t\sqrt{U/k}\right)\right)$. As Z_δ and U are independent, we have

$$\begin{aligned}\mathbb{P}\left(Z_\delta \leq t\sqrt{U/k}\right) &= \int_{\mathbb{R}} \mathbb{P}\left(Z_\delta \leq t\sqrt{u/k}\right) f_U(u)du \\ &= \int_{\mathbb{R}} \Phi\left(t\sqrt{u/k} - \delta\right) f_U(u)du\end{aligned}$$

Thus, inverting $\int_{\mathbb{R}}$ and $\frac{d}{dt}$,

$$\begin{aligned}f_1(t) &= \int_{\mathbb{R}} \frac{d}{dt}\left(\Phi\left(t\sqrt{u/k} - \delta\right)\right) f_U(u)du \\ &= \int_{\mathbb{R}} \sqrt{u/k}\phi\left(t\sqrt{u/k} - \delta\right) f_U(u)du \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(t\sqrt{u/k} - \delta\right)^2\right] \frac{1}{2^{k/2}\Gamma(k/2)} u^{k/2-1}e^{-\frac{u}{2}}\mathbf{1}_{u>0}du \\ &= \frac{1}{\sqrt{2k\pi}} \frac{1}{2^{\frac{k}{2}}\Gamma(k/2)} \int_{\mathbb{R}_+} \exp\left[-\frac{1}{2}\left(\left(1 + \frac{t^2}{k}\right)u + \delta^2 - 2\delta t\sqrt{u/k}\right)\right] u^{\frac{k-1}{2}}du\end{aligned}$$

Then, using the transformation $v = \sqrt{(1+\frac{t^2}{k})u}$, we note that

$$\begin{aligned}
&\int_{\mathbb{R}_+} \exp\left[-\frac{1}{2}\left(\left(1+\frac{t^2}{k}\right)u + \delta^2 - 2\delta t\sqrt{u/k}\right)\right] u^{\frac{k-1}{2}} du \\
= &\int_{\mathbb{R}_+} \exp\left[-\frac{1}{2}\left(v^2 + \delta^2 - 2\delta t v \frac{1/\sqrt{k}}{\sqrt{1+t^2/k}}\right)\right] \frac{v^{k-1}}{\left(1+\frac{t^2}{k}\right)^{\frac{k-1}{2}}} \frac{2v}{1+\frac{t^2}{k}} dv \\
= &\int_{\mathbb{R}_+} \exp\left[-\frac{1}{2}\left(\left(v - \delta t \frac{1/\sqrt{k}}{\sqrt{1+t^2/k}}\right)^2 + \delta^2\left(1 - \frac{t^2/k}{1+t^2/k}\right)\right)\right] \times \frac{2v^k}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} dv \\
= &\exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right] \frac{2}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} \int_{\mathbb{R}_+} \exp\left[-\frac{1}{2}\left(v - \frac{\delta t}{\sqrt{k+t^2}}\right)^2\right] v^k dv \\
= &\exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right] \frac{2}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} k! Hh_k\left(\frac{-\delta t}{\sqrt{k+t^2}}\right)
\end{aligned}$$

Thus

$$\begin{aligned}
f_1(t) = &\frac{1}{\sqrt{2k\pi}} \frac{1}{2^{\frac{k}{2}}\Gamma(k/2)} \exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right] \\
&\times \frac{2}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} k! Hh_k\left(\frac{-\delta t}{\sqrt{k+t^2}}\right)
\end{aligned}$$

which completes the proof because $\Gamma(k+1) = k!$ □

The following property of Hh_k is useful to prove that $\frac{f_1}{f_0}$ is non-decreasing

LEMMA 3.6.3.

$$Hh'_{k+1}(z) = -Hh_k(z)$$

PROOF. Let $k \in \mathbb{N}$. As $Hh_{k+1}(z) = \int_0^{+\infty} \frac{x^{k+1}}{(k+1)!} e^{-\frac{1}{2}(x+z)^2} dx$, we have

$$\begin{aligned}
Hh'_{k+1}(z) &= \int_0^{+\infty} \frac{x^{k+1}}{(k+1)!} \left(-(x+z)\right) e^{-\frac{1}{2}(x+z)^2} dx \\
&= \left[\frac{x^{k+1}}{(k+1)!} e^{-\frac{1}{2}(x+z)^2}\right]_0^{+\infty} - \int_0^{+\infty} \frac{(k+1)x^k}{(k+1)!} e^{-\frac{1}{2}(x+z)^2} dx \\
&= 0 - Hh_k(z)
\end{aligned}$$

□

PROOF OF THEOREM 3.3.8. (i) As $t \mapsto \exp\left[-\frac{\delta^2}{2}\frac{1}{1+\frac{t^2}{k}}\right]$ is non-decreasing and $t \mapsto -\frac{\delta t}{\sqrt{k+t^2}}$ is non-increasing, it is sufficient to prove that Hh_k is non-increasing, which follows from lemma 3.6.3 because Hh_{k-1} is positive.

(ii) by proposition 3.3.2 it suffices to note that

$$\lim_{t\to+\infty} \frac{f_1}{f_0}(t) = \frac{Hh_k(-\delta)}{Hh_k(0)}$$

□

Criticality and identifiability.

PROOF OF LEMMA 3.3.9. (i) is obvious because F_0 is increasing. For (ii), recall that

$$g_1(x) = \frac{f_1}{f_0}\left(-F_0^{-1}(x)\right) ,$$

with $f_1(y) = f_0(y-\mu)$. As F_0 is symmetric, we have $F_0^{-1}(1-x) = -F_0^{-1}(x)$, so that

$$g_1(1-x) = \frac{f_1}{f_0}\left(F_0^{-1}(x)\right) .$$

As $F_0^{-1}(x_\mu) = F_0^{-1}(x) + \mu$, we therefore have

$$\begin{aligned} g_1(1-x_\mu) &= \frac{f_1}{f_0}\left(F_0^{-1}(x)+\mu\right) \\ &= \frac{f_0(F_0^{-1}(x))}{f_0(F_0^{-1}(x)+\mu)} \\ &= \frac{f_0(-F_0^{-1}(x))}{f_0(-F_0^{-1}(x)-\mu)} \\ &= \frac{f_1}{f_0}\left(-F_0^{-1}(x)\right) , \end{aligned}$$

which proves (ii). Finally, (iii) is a direct consequence of (i) and (ii) because g_1 is non increasing. □

PROOF OF THEOREM 3.3.10. First note that π_0 is identifiable if and only if $\lim_{x\to 0} g_1(1-x) = 0$, and that $\alpha^\star = 0$ if and only if $\lim_{x\to 0} g_1(x) = +\infty$. Therefore, if $\alpha^\star = 0$, then by Lemma 3.3.9(iii) we have $\lim_{x\to 0} g_1(1-x) = 0$; hence identifiability holds. Conversely, assume that $\lim_{x\to 0} g_1(1-x) = 0$. Note that $x \to 0$ is equivalent to $F_0^{-1}(x) \to -\infty$, which in turn is equivalent to $x_\mu \to 0$ since $x_\mu = F_0(F_0^{-1}(x)+\mu)$. Thus, we also have $\lim_{x\to 0} g_1(1-x_\mu) = 0$, which proves that $\lim_{x\to 0} g_1(x) = +\infty$ by Lemma 3.3.9(ii). □

3.6.2. Proofs of section 3.4.
Storey's estimator with $\lambda \to 1$.

PROOF OF PROPOSITION 3.4.6. We demonstrate that $\widehat{\pi_0}(\lambda_m)$ may be written as a sum of m independent random variables that satisfy the Lindeberg-Feller conditions for the Central Limit Theorem [71]. Let $Z_i^m = \mathbf{1}_{P_i \geq 1-h_m}$, where the P_i are the p-values. Z_i^m follows a Bernoulli distribution with parameter $p_m = 1 - G(1-h_m)$. Denoting

$$Y_i^m = \frac{Z_i^m - \mathbb{E}\left[Z_i^m\right]}{\sqrt{mh_m}}$$

we have

$$\begin{aligned}\sum_{i=1}^{m} Y_i &= \sqrt{mh_m}\,(\widehat{\pi_0}(1-h_m) - \mathbb{E}\,[\widehat{\pi_0}(1-h_m)]) \\ &= \sqrt{mh_m}\,(\widehat{\pi_0}(1-h_m) - \pi_0)\end{aligned}$$

$(Y_i^m)_{1\le i\le m}$ are centered, independent random variables, with $\operatorname{Var} Y_i^m = \frac{\operatorname{Var} Z_i^m}{mh_m} = \frac{G(1-h_m)(1-G(1-h_m))}{mh_m}$, which is equivalent to $\frac{\pi_0}{m}$ as $m \to +\infty$. Therefore,

$$\lim_{m\to+\infty} \sum_{i=1}^{m} \mathbb{E}\left[(Y_i)^2\right] = \pi_0\,.$$

Finally we prove that for any $\varepsilon > 0$,

$$\lim_{m\to+\infty} \sum_{i=1}^{m} \mathbb{E}\left[(Y_i)^2 \mathbf{1}_{|Y_i^m|>\varepsilon}\right] = 0\,.$$

As $Z_i^m \in \{0,1\}$ and $\mathbb{E}\,[Z_i^m] \in [0,1]$, we have $(Y_i^m)^2 \le \frac{1}{h_m}$, and

$$\begin{aligned}\sum_{i=1}^{m} \mathbb{E}\left[(Y_i)^2 \mathbf{1}_{|Y_i^m|>\varepsilon}\right] &\le \frac{1}{h_m}\mathbb{E}\left[\mathbf{1}_{|Y_i^m|>\varepsilon}\right] \\ &= \frac{1}{h_m}\mathbb{P}\Big(\mathbf{1}_{|Y_i^m|>\varepsilon}\Big) \\ &\le \frac{1}{h_m}\frac{\operatorname{Var} Y_i^m}{\varepsilon^2}\end{aligned}$$

by Chebycheff's inequality. As $mh_m \to +\infty$ and $\operatorname{Var} Y_i^m \sim \frac{\pi_0}{m}$ as $m \to +\infty$, the above sum therefore goes to 0 as $mh_m \to +\infty$. The Lindeberg-Feller conditions for the Central Limit Theorem are thus fulfilled, and we have

$$\sum_{i=1}^{m} Y_i \rightsquigarrow \mathcal{N}(0, \pi_0)\,,$$

which concludes the proof because $\sum_{i=1}^m Y_i = \sqrt{mh_m}\,(\widehat{\pi_0}(1-h_m) - \pi_0)$. □

Proof of Proposition 3.4.7. By Proposition 3.4.6, the asymptotic variance of $\widehat{\pi_0}(1-h_m)$ is equivalent to $\frac{\pi_0}{mh_m}$. As we assumed that the $k-1$ first derivatives of g_1 at 1 are null, and that $g_1^{(k)}(1) \neq 0$, a Taylor expansion of $\widehat{\pi_0}(1-h_m) - \pi_0$ ensures that the bias is of the order of h_m^k. The optimal bandwidth is obtained for h_m proportional to $m^{-\frac{1}{2k+1}}$, because this choice balances variance and squared bias. The proportionality constant, which we denote by C_k, is an explicit function of k, π_0, and $g_1^{(k)}(1)$.

By definition, the MSE that corresponds to this optimal choice is twice the variance, and the asymptotic distribution of the corresponding $\widehat{\pi_0}$ is derived from Proposition 3.4.6. □

Asymptotic FDP for plug-in procedures. We study the BH95procedure at level $\alpha/\widehat{\pi_0}$, where $\widehat{\pi_0}$ is an estimator of π_0 that converges to π_0 at rate $\sqrt{mh_m}$, where $h_m \to 0$. This procedure rejects all hypotheses with p-values smaller than

$$\widehat{\tau} = \sup\left\{t \in [0,1], \hat{\mathbb{G}}_m(t) \ge \widehat{\pi_0} t/\alpha\right\}\,.$$

The associated proportion of rejections and proportion of incorrect rejections are given by $\widehat{\rho} = \hat{\mathbb{G}}_m(\widehat{\tau}) = \widehat{\tau}\widehat{\pi_0}/\alpha$, and $\widehat{\nu} = \pi_0 \hat{\mathbb{G}}_{0,m}(\widehat{\tau})$, respectively, where $\hat{\mathbb{G}}_{0,m}$ denotes the empirical distribution function of p-values that correspond to true null hypotheses. The asymptotic threshold of the Oracle BH95procedure is defined by

$$\tau^\star = \sup\left\{t \in [0,1], G(t) \geq \pi_0 t/\alpha\right\} .$$

Note that by the definition of $\widehat{\tau}$ and $\tau^\star$, we have $\hat{\mathbb{G}}_m(\widehat{\tau}) = \widehat{\pi_0}\widehat{\tau}/\alpha$ and $G(\tau^\star) = \pi_0\tau^\star/\alpha$. The following Proposition shows that the convergence rate of $(\widehat{\tau}, \widehat{\nu}, \widehat{\rho})$ is driven by the convergence rate of $\widehat{\pi_0}$.

PROPOSITION 3.6.4. *Let $\alpha > \pi_0\alpha^\star$, and $\widehat{\pi_0}$ be any estimator of π_0 with asymptotic distribution given by $\sqrt{mh_m}\,(\widehat{\pi_0} - \pi_0) \rightsquigarrow \mathcal{N}(0, v(\pi_0))$ for some function v. Then, as $m \to +\infty$,*

$$\begin{pmatrix}\widehat{\tau}\\ \widehat{\nu}\\ \widehat{\rho}\end{pmatrix} - \begin{pmatrix}\tau^\star\\ \pi_0\tau^\star\\ \pi_0\tau^\star/\alpha\end{pmatrix} = \frac{\tau^\star/\alpha}{g(\tau^\star) - \pi_0/\alpha}\begin{pmatrix}1\\ \pi_0\\ g(\tau^\star)\end{pmatrix}(\widehat{\pi_0} - \pi_0)(1 + \mathrm{o}\,(1))$$

PROOF OF PROPOSITION 3.6.4. We begin by noting that $\widehat{\tau}$ converges almost surely to $\tau^\star$. Let $\psi_{F,\gamma} : u \mapsto F(u) - u/\gamma$ for any distribution function F and any $\gamma \in (0,1]$. As $\hat{\mathbb{G}}_m(\widehat{\tau}) = \widehat{\pi_0}\widehat{\tau}/\alpha$ and $G(\tau^\star) = \pi_0\tau^\star/\alpha$, we have $\psi_{G,\alpha/\pi_0}(\tau^\star) = 0$ and $\psi_{\hat{\mathbb{G}}_m,\alpha/\widehat{\pi_0}}(\widehat{\tau}) = 0$. The idea of the proof is to note that

- $\psi_{G,\alpha/\pi_0(\widehat{\tau})}$ converges almost surely to $0 = \psi_{G,\alpha/\pi_0}(\tau^\star)$
- $\psi_{G,\alpha/\pi_0}$ is locally invertible in a neighborhood of $\tau^\star$.

The second point holds because we are in a subcritical situation: $\alpha > \pi_0\alpha^\star$, with $\alpha^\star = \lim_{u\to 0} u/G(u)$. For the first point, note that

$$\begin{aligned}\psi_{G,\alpha/\pi_0}(\widehat{\tau}) &= G(\widehat{\tau}) - \pi_0\widehat{\tau}/\alpha\\ &= (G - \hat{\mathbb{G}}_m)(\widehat{\tau}) + (\hat{\mathbb{G}}_m(\widehat{\tau}) - \widehat{\pi_0}\widehat{\tau}/\alpha) + (\widehat{\pi_0} - \pi_0)\widehat{\tau}/\alpha\,.\end{aligned}$$

The first terms converges to 0 almost surely, the second is identically null, and the third converges almost surely to 0 because $\widehat{\pi_0}$ is consistent. Hence $\widehat{\tau}$ converges almost surely to $\tau^\star$.

We only prove the result for $\widehat{\tau}$, as the proofs for $\widehat{\nu}$ and $\widehat{\rho}$ are quite similar. The idea is that because $h_m \to 0$, the fluctuations of $\hat{\mathbb{G}}_m - G$ are negligible with respect to the fluctuations of $\widehat{\pi_0} - \pi_0$. We have

$$\begin{aligned}G(\widehat{\tau}) - G(\tau^\star) &= (G(\widehat{\tau}) - \hat{\mathbb{G}}_m(\widehat{\tau})) + (\hat{\mathbb{G}}_m(\widehat{\tau}) - G(\tau^\star))\\ &= \bar{\mathbb{G}}_m(\widehat{\tau}) + (\widehat{\pi_0}\widehat{\tau}/\alpha - \pi_0\tau^\star/\alpha)\end{aligned}$$

because $\hat{\mathbb{G}}_m(\widehat{\tau}) = \widehat{\pi_0}\widehat{\tau}/\alpha$ and $G(\tau^\star) = \pi_0\tau^\star/\alpha$, where $\bar{\mathbb{G}}_m = \hat{\mathbb{G}}_m - G$ is the centered empirical process associated with G. Therefore,

$$G(\widehat{\tau}) - G(\tau^\star) = \bar{\mathbb{G}}_m(\widehat{\tau}) + \frac{\widehat{\pi_0}}{\alpha}(\widehat{\tau} - \tau^\star) + \frac{\widehat{\pi_0} - \pi_0}{\alpha}\tau^\star\,.$$

Noting that $\bar{\mathbb{G}}_m$ is of the order of $1/\sqrt{m}$, $\bar{\mathbb{G}}_m(\widehat{\tau}) = \mathrm{o}\,(\widehat{\pi_0} - \pi_0)$. Finally, since $\widehat{\tau} \xrightarrow{a.s.} \tau^\star$ as $m \to +\infty$, we also have $G(\widehat{\tau}) - G(\tau^\star) = (\widehat{\tau} - \tau^\star)(g(\tau^\star) + \mathrm{o}\,(1))$ by Taylor's formula, which concludes the proof. □

PROOF OF THEOREM 3.4.8. By Proposition 3.6.4 and the Delta method, we have

$$\sqrt{mh_m}\left(\begin{pmatrix}\widehat{\nu}\\ \widehat{\rho}\end{pmatrix}-\begin{pmatrix}\pi_0\tau^\star\\ \pi_0\tau^\star/\alpha\end{pmatrix}\right)\rightsquigarrow\mathcal{N}\left(0,V\right),$$

with

$$V=h(\pi_0)\left(\frac{\tau^\star/\alpha}{g(\tau^\star)-\pi_0/\alpha}\right)^2\begin{pmatrix}\pi_0\\ g(\tau^\star)\end{pmatrix}\begin{pmatrix}\pi_0 & g(\tau^\star)\end{pmatrix}.$$

We write $\mathsf{FDP}=\gamma(\widehat{\nu},\widehat{\rho})$, where $\gamma:(x,y)\mapsto x/y$ for any $x\geq 0$ and $y>0$. γ is differentiable at $(\pi_0\tau^\star,\pi_0/\alpha\tau^\star)$, with derivative

$$\begin{aligned}\dot{\gamma}_{\pi_0\tau^\star,\pi_0/\alpha\tau^\star} &= (1/\pi_0/\alpha\tau^\star,-\pi_0\tau^\star/(\pi_0/\alpha\tau^\star)^2)\\ &= \frac{\alpha}{\pi_0\tau^\star}(1,-\alpha).\end{aligned}$$

As $\gamma(\pi_0\tau^\star,\pi_0/\alpha\tau^\star)=\alpha$ the Delta method yields

$$\sqrt{mh_m}\left(\mathsf{FDP}_m-\alpha\right)\rightsquigarrow\mathcal{N}\left(0,w\right),$$

with

$$\begin{aligned}w &= h(\pi_0)\left(\frac{\tau^\star/\alpha}{g(\tau^\star)-\pi_0/\alpha}\right)^2\dot{\gamma}_{\pi_0\tau^\star,\pi_0/\alpha\tau^\star}\begin{pmatrix}\pi_0\\ g(\tau^\star)\end{pmatrix}\begin{pmatrix}\pi_0 & g(\tau^\star)\end{pmatrix}\dot{\gamma}'_{\pi_0\tau^\star,\pi_0/\alpha\tau^\star}\\ &= h(\pi_0)\left(\frac{\tau^\star/\alpha}{g(\tau^\star)-\pi_0/\alpha}\right)^2\left(\frac{\alpha}{\pi_0\tau^\star}\begin{pmatrix}1 & -\alpha\end{pmatrix}\begin{pmatrix}\pi_0\\ g(\tau^\star)\end{pmatrix}\right)^2\\ &= h(\pi_0)/\pi_0^2\left(\frac{\pi_0-\alpha g\tau^\star}{g(\tau^\star)-\pi_0/\alpha}\right)^2\\ &= h(\pi_0)\alpha^2/\pi_0^2.\end{aligned}$$

□

3.6.3. Proofs of section 3.5.

PROOF OF PROPOSITION 3.5.2. (i) By the definition of $\widehat{\tau}$, we have $\hat{\mathbb{G}}_m(\widehat{\tau})=\widehat{\tau}/\alpha$. Thus, we have

$$\widehat{\tau}/\alpha=(1-\varepsilon_m)\widehat{\tau}+\varepsilon_m\hat{\mathbb{G}}_{1,m}(\widehat{\tau}).$$

As $\hat{\mathbb{G}}_{1,m}(\widehat{\tau})\leq 1$, we thus have $\widehat{\tau}/\alpha\leq\widehat{\tau}+\varepsilon_m$, which proves that $\widehat{\tau}$ converges almost surely to 0, as $\alpha<1$. The same holds for R_m/m because $R_m/m=\hat{\mathbb{G}}_m(\widehat{\tau})=\widehat{\tau}/\alpha$.

(ii) Consequence of [18, Lemma S2.1 and S2.3].

□

PROOF OF PROPOSITION 3.5.3. We have

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_1^D}\left(P_{(1)}\geq\frac{\alpha}{m}\right) &= \left(1-G^{(m)}\left(\frac{\alpha}{m}\right)\right)^m\\ &= \exp\left(m\log\left(1-G^{(m)}\left(\frac{\alpha}{m}\right)\right)\right)\end{aligned}$$

We have $G^{(m)}\left(\frac{\alpha}{m}\right)=(1-\varepsilon_m)\frac{\alpha}{m}+\varepsilon_m G_1^{(m)}\left(\frac{\alpha}{m}\right)\leq\frac{\alpha}{m}+\varepsilon_m\to 0$. Therefore, $\lim_{m\to+\infty}\mathbb{P}_{\mathcal{H}_1^D}\left(P_{(1)}\leq\frac{\alpha}{m}\right)=1$ if and only if

$$\lim_{m\to+\infty}mG^{(m)}\left(\frac{\alpha}{m}\right)=+\infty.$$

□

PROOF OF PROPOSITION 3.5.6. As $G^{(m)}$ is concave, we have $\frac{G^{(m)}(\alpha/m)}{\alpha/m} \geq g^{(m)}(\alpha/m) \geq g^{(m)}(1/m)$. As $g^{(m)}(u) = (1-\varepsilon_m) + \varepsilon_m g_1^{(m)}(u)$, condition (3.5.5) in theorem 3.5.4 can thus be replaced by condition (3.5.7). □

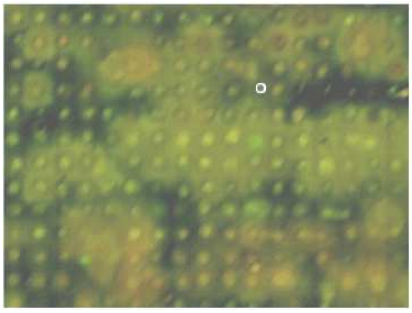
Part 2

Application to microarray data analysis

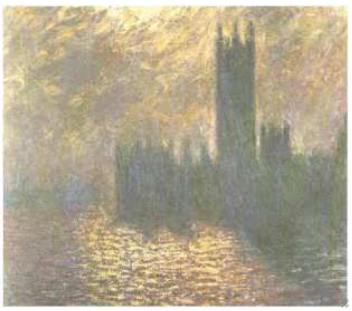
CHAPTER 4

Microarray analysis for cancer resarch

IMPRESSIONISM



Mary Brinig, *B18n030*, 2003



Claude Monet, *Le Parlement, Effet de Brouillard*, 1903

Contents

Cancer is a class of diseases in which abnormal cells proliferate without control, avoid programmed cellular death, and are able to invade other tissues, and eventually spread to other parts of the body through the blood and lymph systems. The main goals of cancer research are biological and clinical: to better understand the biological mechanisms underlying disease pathogenesis, and to improve cancer diagnosis, prognosis and treatment.

Cancer is fundamentally a disease of regulation of tissue growth. In order for a normal cell to be transformed into a cancer cell, genes which regulate cell growth and differentiation must be altered. Genetic changes can occur at many levels, from gain or loss of entire chromosomes to a mutation affecting a single DNA nucleotide, and directly or indirectly result in modifications of genes expressions. This motivates the use of advanced molecular biology techniques such as DNA microarrays for cancer research. Such high-throughput technologies require the development of dedicated statistical methods that are adapted to the dimensionality of these data, as well as to each specific biological or clinical problem of interest.

We begin by a brief description of cancer cell physiology, and of genomic changes that occur in cancers (section 4.1). Then we describe DNA microarray techniques (section 4.2), and give an overview of statistical issues of interest for their analysis (section 4.3). Finally we list the contributions of this thesis in terms of statistical analysis of DNA microarray data (section 4.4).

4.1. Cancer and genes

4.1.1. A few words of molecular biology. Genes may be defined as heritable units of information that drive the physical development and phenotype of an organism by interacting with each other and with the environment. A gene is encoded in a sequence of four chemical compounds called deoxyribonucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). These nucleotides make up a long strand of DNA (deoxyribonucleic acid), which is thus considered as the carrier of genetic information across generations of cells and organisms.

In a classical, textbook view of how genes are associated with the behavior of the cell, the DNA sequence of a gene is transcribed intro mRNA (messenger ribonucleic acid), and this mRNA sequence may be in turn translated into a protein, which consists of a sequence of amino acids. Translation is performed according to a quasi-universal genetic code that maps each triplet of nucleotides to an amino acid. Genome, transcriptome and proteome are three levels of information that refer to the set of genes, messenger RNAs, and proteins, respectively.

Proteins are essential structural components of organisms; they participate in every cell process: for example, chemical reactions are catalyzed by enzymes; cellular transport and communication involve extracellular or membrane proteins; structural proteins maintain cell shape. However, the activity of a cell may not be understood based only on the above simplistic picture of a linear flow of genetic information from DNA to protein for each gene, independently from other genes and the environment. This information flow and the resulting protein activities, which constitute the molecular

phenotype of the cell, are strictly controlled, both by environmental stimuli (which may be external or internal to the cell) and by tissue-specific regulation mechanisms that reflect complex interactions between DNA, mRNA, proteins, and small sequences of non protein-coding RNA (ncRNA), including micro RNA (miRNA), and small interfering RNA (siRNA).

These regulation mechanisms may be classified into three types, depending on the information level at which the target gene is influenced: *transcriptional regulation* occurs before transcription and regroups activation or repression of the expression of specific genes by proteins called transcription factors, and conformational or chemical modifications of DNA called epigenetic modifications. *Post-transcriptional regulation* covers mechanisms through which a given primary gene transcript may be alternatively spliced into several mature transcripts, and mechanisms of repression of gene expression by miRNA. *Post-translational regulation* involves chemical modifications of proteins which turn a protein's activity on or off; for example phosphorylations are catalyzed by kinases, and dephosphorylations are catalyzed by phosphatases.

4.1.2. Genomic changes in cancer cells. Even though there are dozens of cancer types, and many more subtypes, it is now well admitted that cancer cells are characterized by few essential alterations in cell physiology, that guide malignant growth [40]. These alterations are illustrated by Figure 1.

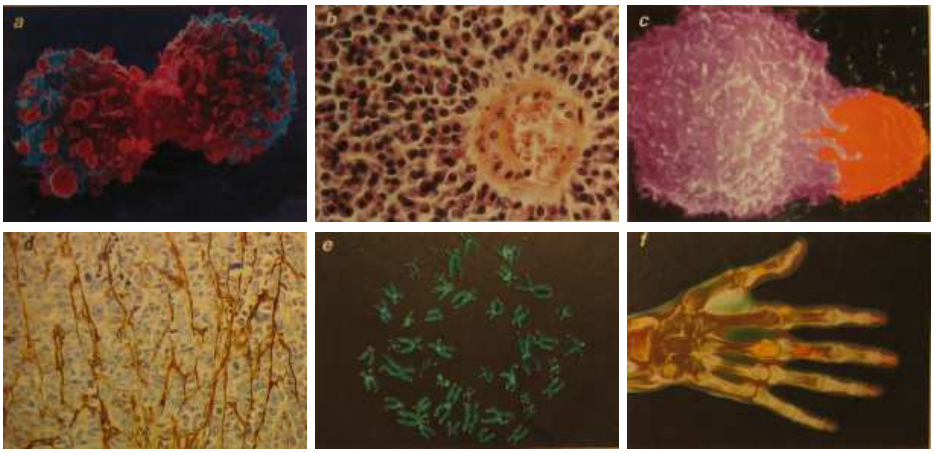


FIGURE 1. *Acquired capabilities of cancer cells: (a) self-sufficiency in growth signals; (b) insensitivity to growth-inhibitory (antigrowth) signals; (c) evasion of programmed cell death (apoptosis); (d) sustained angiogenesis (growth of blood vessels); (e) limitless replicative potential; (f) tissue invasion and metastasis. Images taken from [39].*

Biological evidence suggests that these capabilities are acquired through *genomic alterations* in cancer cells: either aneuploidies, that is, the presence of an abnormal number of chromosomes, or mutations, that is, changes in the nucleotide sequence of genomic DNA. However, these genomic changes are quite rare events in normal cells, due to efficient regulatory mechanisms

that aim at maintaining genomic integrity [60]. This suggests that *genome instability* should be added to the list of acquired capabilities of cancer cells, and considered as an enabling characteristic for the other capabilities to be acquired.

Genomic changes that occur during tumorigenesis involve small or large-scale alterations. *Small-scale alterations* include point mutations, deletions, and insertions, which may occur in the promoter of a gene and affect its expression, or in the gene's coding sequence and alter the function or stability of its protein product. *Large-scale alterations* include the deletion or gain of (a portion of) a chromosome, but also genomic amplification, which occurs when a cell gains many copies (often 20 or more) of a small chromosomal locus, often containing one or more oncogenes, and translocations, that is the abnormal fusion of two separate chromosomal regions.

Two broad categories of genes are affected by these genomic changes: *oncogenes* and *tumor suppressor genes*. Oncogenes may be normal genes which are expressed at inappropriately high levels, such as MYCN, which is amplified[1] in many neuroblastoma (the most frequent pediatric tumor) or altered forms of *proto-oncogenes*, such as mutated Fibroblast growth factor receptor (FGFR3) in bladder cancers. In either case, overexpression of these genes promotes the malignant phenotype of cancer cells because they directly or indirectly control cell proliferation and/or apoptosis. Tumor suppressor genes inhibit cell division or promote apoptosis; for instance TP53 codes for p53, a transcription factor involved in cell cycle regulation.

4.2. Microarray data in cancer research

4.2.1. Overview. DNA microarrays are a molecular biology technique that performs simultaneous measurement of a given level of genomic information (DNA copy number, expression level, or protein activity) for each locus or gene within the genome of a biological sample. It takes advantage of the specific base pairing between DNA nucleotides: adenine with thymine, and cytosine with guanine. This fundamental property of DNA allows sequences which have been extracted from a biological sample (targets) and labeled with fluorescent molecules to hybridize to their complementary sequences (probes), which have been fixed at known locations to a solid surface, the microarray. As a result, it is possible to quantify the amount of DNA bound to each location of the microarray using a scanner that measures the amount of fluorescence.

Figure 2 illustrates the result of a typical microarray experiment. Thousands to millions of different probe sequences are fixed to a microarray, each sequence corresponding to a given location on the genome; each sequence is represented by thousands of probes, in order to increase hybridization probability.

There exist two main types of microarrays, depending on whether probes are spotted on the microarray or synthesized (either *in situ* or at the surface of beads). Each of these two families of microarrays have technical specificities (sketched in Figure 3 for expression microarrays) which must be taken into account for proper statistical analysis of the resulting data.

[1]see Appendix E.

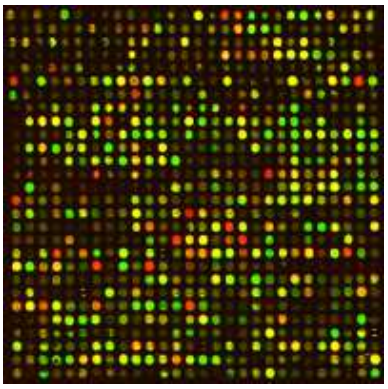


FIGURE 2. *Small portion of a scanned (spotted) microarray (1 out of 48 blocks). Each colored dot corresponds to a spot where thousands of identical sequences have been fixed. Color scale ranges from green (underexpressed genes) to red (overexpressed genes) by yellow (normally expressed genes).*

spotted-probe microarrays: a test and a reference sample are labeled with different molecules that emit energy at two different wavelengths, and hybridized on the same microarray; because it is not possible to know accurately how many DNA fragments have been spotted for each probe, only the *ratio* between the test and the reference intensity levels can be compared across probes;

synthesized-probe microarrays: as the number of sequences synthesized is exactly the same for each probe, *absolute intensity levels* can be compared across probes; in most cases synthesized-probe microarrays are thus single-colored.

FIGURE 3. *Comparison between spotted (left) and synthesized (right) microarrays, in the case of expression microarrays. Image Courtesy of* The Science Creative Quarterly*; artist: Jiang Long.*

Microarrays in cancer research are most frequently used to quantify gene expression levels or DNA copy numbers:

expression microarrays: Expression microarrays, are used to estimate an absolute or relative quantity of messenger RNA for each probe sequence on the microarray. The most widely used expression microarrays are single-color, oligonucleotide microarrays, in particular those commercialized by Affymetrix[2]; many two-color expression array platforms have also been developed;

copy number microarrays: Comparative genomic hybridization (CGH) is a molecular cytogenetic technique based on the competitive hybridization of fluorescently labeled tumor DNA and normal DNA to normal metaphase chromosomes [52]. It permits detecting DNA copy number changes (gains or losses) with a resolution of 10 to 20 megabases (Mb). In the late nineties a microarray-based version of this technique (array-CGH) has been developed [69], which permits detecting copy number changes at a resolution of tens to hundreds of kilobases, depending on the probe density of the array.

More recently, another DNA-level microarray has been developed, that provides high-resolution genome-wide identification of Single Nucleotide Polymorphisms (SNP), that is DNA sequence variations at a single genome locus. SNP arrays are single-color microarrays that may be used to estimate an absolute copy number.

4.2.2. Applications to cancer research. Microarrays turn out to be a technology of choice for both biological cancer research, which aims at understanding tumorigenesis and tumor progression, and clinical cancer research, which aims at improving cancer diagnosis, prognosis and treatment. The use of microarray technologies illustrates tight connections between these two aspects of cancer research (statistical issues are discussed in more detail in section 4.3.2):

- Exploratory and comparative analyses help identifying genes that are specific from a cancer type or subtype: these are either candidate oncogenes or tumor suppressor genes found using CGH-arrays [3] in regions gained and lost, respectively, or over- or underexpressed genes in specific conditions, revealed by expression arrays. This type of analyses also pinpoint groups of co-expressed genes or of samples that share similar expression or copy number profiles [86]. Such genes are potential key players in specific cancer types or subtypes, and as such may enhance global understanding of cancer mechanisms, but also refine cancer diagnosis, and suggest new therapeutic targets;
- Classification analyses aim at building gene signatures or *biomarkers* whose expression or copy number may be used to predict a biological or clinical variable of interest, such as the clinical outcome for a patient if no anticancer drug has been administered (prognostic biomarkers), or the outcome of a specific therapy for a patient

[2]http://www.affymetrix.com

(predictive biomarkers). Functional analysis of these biomarkers may also provide insights into fundamental mechanisms of cancer.

As a practical illustration of the use of microarray technologies for cancer research, we give an example of development of prognostic biomarker for predicting breast cancer relapse, which illustrates one of the directions that are being explored towards personalized medicine.

Today, most women with early stage breast cancer undergo adjuvant chemotherapy (that is, chemotherapy after surgery), whereas most of them could be healed by surgery and localized radiotherapy: the problem is that classical clinical parameters fail to identify those patients who really need chemotherapy (that is, who otherwise would relapse). Improving prediction accuracy of disease outcome is thus of major importance in terms of public health, as it could help avoiding unnecessary chemotherapies and their side effects.

Several studies underlined the potential of gene expression data to predict disease outcome at time of diagnosis. For example, a 70−gene signature (that is, a classifier based on these genes' expressions) has been identified, which predicts whether a given patient is likely to develop metastases within 5 years [107]. This signature is one of the three expression-based prognostic signatures already commercially available for breast cancer.

4.3. Statistical issues in microarray data analysis

4.3.1. Low-level analyses. Statistical expertise is first required *before* data analysis, for the design of experiments, image analysis, and data normalization. The advent of a new type of microarray generally necessitates the development of dedicated methods that take its specificities into account [87].

Design of experiments. This question is often still overlooked both by statisticians and by biologists or clinicians: it is not uncommon that the statistician is asked to compare two cancer subgroups using microarray experiments that have all been performed on day 1 for group 1, and on day 2 for group 2. In such a situation it is impossible to determine whether the observed differences come from the cancer subgroup, or from other parameters related to experimental conditions. More generally, one should make sure *beforehand* that the design of experiments permits questions of interest to be answered. This involves deciding how many replicated experiments are performed, which depends on the availability of biological material, the cost of the experiment, and the (estimated) power of the technology to detect an effect of a given amplitude. Another important aspect is the design of the microarray itself: which probe sequences should be chosen ? How many replicates per probe ? Where should they be located on the array?

Image analysis. The output of a microarray experiment is an image that comes from a scanner (see Figure 2), from which a signal value has to be evaluated for each probe. Image analysis usually involves *spot location* using a technology-specific grid, *segmentation* between regions corresponding to signal and noise, and *quantifiation* of the signal corresponding to each spot.

Normalization. Microarray data have been reported to be poorly reproducible, and are affected by various sources of systematic variation [110,

111]. Hence the need for within and between array normalization, which aims at removing such artefactual variation while preserving the true biological signal, and making signals coming from different experiments comparable. Microarray normalization has been quite an active research field over the last few years, and there is still no consensus on which method should be preferred; one of the reason is that normalization methods rely on assumptions on the data that are not easily checked; it is thus difficult to balance "too little normalization" with "too much normalization". As an illustration, a recent study suggests that GCRMA, one of the most popular normalization method for Affymetrix Genechip® arrays, induces artefactual correlations between genes with low expression levels [58], which may bias downstream analyses. This is closely related to the open question of *filtering out* genes that are unexpressed in a whole microarray expression data set.

4.3.2. High-level analyses. Biological and clinical questions in microarray data analysis involve classical domains of statistics: exploratory analyses, hypothesis testing, classification and regression. However, statistical methods have to be adapted to the "small n, large p" context of microarray data: the number p of variables (genes) generally exceeds the number n of available of observations (biological samples) by two or three orders of magnitude.

Exploratory analyses. These analyses aim at identifying groups of genes that share similar patterns across samples, or *vice versa*; they are needed even when the biological question only involves supervised analyses, as they help understanding the data as a whole, and often help identifying artefactual variations that remained after normalization. Classical methods include distance or model-based clusterings, factorial analyses such as Principal Component or Independent Component Analyses (PCA and ICA), and, more recently, biclustering methods, which aim at finding groups of genes that share similar patterns among a group of samples [15].

Hypothesis testing. Comparative analyses aim at identifying those genes whose measurement (expression level, or DNA copy number) significantly differs between two groups of samples. Performing a statistical test for each gene requires an adapted risk measurement, which triggered the development of the multiple testing techniques studied in the first part of this thesis.

Classification and regression. Constructing *biomarkers* that predict clinical outcome, metastasis-free survival, or response to treatment involves building classifiers and regression models, which should ideally make few errors and be easily interpretable and robust.

The large number of variables requires adapted feature selection methods. Feature selection can be performed *before* building the classifier, either based on their univariate discriminative power or using forward or backward selection heuristics. Alternatively, *penalization* or *regularization*-based approaches add a constraint on the norm of the vector of predictors, so that the regression or classification model may be estimated without prior variable selection. When regularization incorporates a ℓ^1 term such as in the LASSO [100] or in the Elastic Net [112], the estimated model is naturally *sparse*: many estimated coefficients are strictly null.

The small number of observations requires appropriate validation techniques to prevent overfitting, which is usually addressed by *cross-validation*: k-fold cross-validation involves partitioning the original sample in to k subsets, and successively training the model on $k-1$ subsets and evaluating generalization error on the remaining subset. Feature selection steps and estimation of regularization parameters should be performed within the training step in order to ensure a fair evaluation of generalization error.

4.3.3. Integrative approaches. Cells are complex systems whose behavior cannot be understood using only one level of biological information or one technology at a time. A number of *integrative approaches* have emerged over the last few years, with various motivations:

statistical power and robustness: multicentric studies are carried out in order to detect subtle effects, that had previously been missed because of insufficient sample size. Combining data from different platforms raises statistical questions that necessitate dedicated normalization methods [81];

data complementarity: combining different levels of genomic information such as copy number and expression data can help discovering new therapeutic targets [78];

biological knowledge: integrating biological knowledge at the inference step may lead to more biologically interpretable results that separating statistical analysis from biological interpretation [57, 73];

These new approaches are related to the advent of *systems biology*, advocating a *system level approach* to biology that focuses on *interactions* within and between different information levels including genome, transcriptome, proteome, regulation networks and epigenetics.

4.4. Contributions

The first two contributions (sections 4.4.1 and 4.4.2) are generic methods that are now widely used, in particular at Institut Curie. The last two contributions (sections 4.4.3 and 4.4.4) have been designed to address specific biological and clinical questions, respectively.

4.4.1. Normalization of array-CGH data. This work has been done in collaboration with Philippe Hupé.

This project is motivated by the analysis of tumor samples coming from two different platforms: University of California San Francisco (UCSF) [85], and Institut Curie. We demonstrated that the major source of non biologically relevant variation in both platforms was spatial artifacts: either clusters of spots on the microarray, with a discrete signal shift, or a smooth gradient in signal from one side of the microarray to the other. These two effects were not properly corrected by existing techniques.

Spatial segmentation method (Chapter 5). We therefore developed a spatial segmentation method, that involves three steps:

(i) estimation of a spatial trend on the array using two-dimensional LOESS regression [20, 21];

(ii) segmentation of the array into spatial areas with similar trend values using NEM, an unsupervised classification algorithm including spatial constraints [4, 5];
(iii) identification of the areas affected by spatial bias.

This method is of practical use in for genomic studies as it helps preventing the misinterpretation of experimental artifacts as biologically relevant outliers in DNA copy number profiles. The outliers that remain after application of the algorithm can thus be called candidate oncogenes or tumor suppressor genes with increased specificity, without loss of sensitivity. This normalization method has been published in *BMC Bioinformatics* in 2006 [66].

MANOR software. Due to the lack of software dedicated to array-CGH data normalization, we developed MANOR, an R package for Micro-Array NORmalization that includes importation, normalization, visualization, and quality control functions to correct identified sources of variability. The spatial segmentation method we developed is implemented as part of MANOR. This R package is freely available from Bioconductor, an open source and open development software project dedicated to the analysis and comprehension of genomic data [38]. A description of functionalities of MANOR and application examples are given in the *vignette* provided in Appendix A.

Integration to analysis pipelines. Biologists from INSERM Unit U830 developed their own local array-CGH platform; a dedicated analysis pipeline, called CAP for CGH-array Analysis Pipeline, has been implemented by the Bioinformatics team in order to store, analyze, and visualize produced data. I have been involved in the integration of MANOR to CAP. As of June 2008, MANOR has been used for the analysis of over 6000 CGH arrays using CAP, shared by 132 users on 94 research projects.

We decided to implement CAPweb, a web-based version of CAP, which may either be used from our website at `http://bioinfo.curie.fr/CAPweb`, or installed locally for internal use within a specific research center [59]. I participated to the integration of MANOR into CAPweb (Appendix D). As of June 2008, CAPweb has been used for the analysis of over 5000 CGH arrays from our website, shared by 214 users on 468 research projects from 27 countries around the world. CAPweb has been installed locally in 10 academic laboratories, and one private company. Several publications report results that have been obtained using CAPweb [31, 45, 46, 103, 108].

Other uses. The wide range of applicability of the method and software we developed has also been reported by a recent review [53], which indicates that MANOR is "the most suitable algorithm for the correction of spatial biases in microarray experiments[...], relevant also to non-CGH experiments". MANOR has recently been integrated by a research group from University College London (United Kingdom) into PerlMAT, a Perl-based microarray analysis pipeline dedicated to two-color microarrays [63].

4.4.2. Correlating DNA copy number and expression microarrays. This work has been done in collaboration with Pierre Gestraud.

Background. A few recent studies have characterized the overall influence of DNA copy number changes on gene expression (gene dosage effect), using parallel, high-throughput microarray measurements of these two pieces

of information (see [19] and references therein). In this context, there is still a need for an easy-to-use and flexible tool that accommodates various kinds of input data (especially array-CGH, cDNA or SNP arrays for copy number data) in order to quantify this dosage effect.

Algorithm. We have developed GTCA, an R package that implements a statistically sound methodology for Genome Transcriptome Correlation Analysis, including data pre-processing, statistical analysis, visualisation and biological interpretation:

- Data pre-processing: an unambiguous mapping between genome and transcriptome probes according to their position on the genome. Missing DNA copy numbers are inferred by taking advantage of the consistency of the copy number signal along the genome.
- Statistical data analysis: for each probe, a correlation coefficient between DNA copy number and expression data is calculated, as well as an associated p-value. These p-values are then adjusted for multiple comparisons [7, 43] for each chromosome or chromosome arm.
- Data visualisation and interpretation: correlation coefficients and associated (multiple-testing-adjusted) significances can be plotted along the genome together with cytobands, and can also be exported as ranked lists of genes (.rnk), which eases biological interpretation of the results using software like GSEA [95].

A poster describing this algorithm has been presented at the ISMB 2007 conference (Appendix B).

Implementation. This algorithm has been implemented in R, and integrated to VAMP [54], a software developed by the bioinformatics team of Institut Curie which is used by local biologists, clinicians and bioinformaticians. VAMP is devoted to Visualization and Analysis of Molecular Profiles, including DNA copy number and expression profiles (Appendix C). GTCA will soon be submitted to Bioconductor as an R package. It may be used via VAMP on public tumor data sets contained in AcTuDB [44], a public repository for array-CGH tumor data available at:

http://bioinfo.curie.fr/actudb.

4.4.3. Learning cooperative regulation networks. This work has been done in collaboration with Mohamed Elati and Céline Rouveirol [27] (Chapter 6).

Background. Transcription factors are proteins which activate or inhibit the expression of their target genes by binding to specific DNA sequences located in the upstream region of these genes. Reconstructing transcriptional regulation networks is a major challenge towards the understanding of cell behavior, and may also be useful to discover therapeutic targets. Several local approaches have been proposed, that infer a set of regulators for each gene of interest based on a measure of correlation or mutual information between the regulated gene and its potential regulators. However exhaustive search for a given score function is a problem with exponential complexity, which cannot be performed because of the dimensionality of expression data.

Method. We designed a method named LICORN[3], which is described in Chapter 6 and takes advantage of the fact that several transcription factors may be involved in the regulation on one single gene [41]. It associates to each gene a *Gene Regulatory Network* (GRN), which is a pair of sets of transcription factors: an *activator set* and an *inhibitor set*. The method works as follows:

(i) use constrained itemset mining techniques to extract co-activator sets and co-inhibitor sets, based on discretized expression data;
(ii) build a structured set of candidate co-activator sets and co-inhibitor sets for each gene, within which an exhaustive search can be performed efficiently;
(iii) define a score that associates to each gene a *best GRN* among all possible pairs of co-activators and co-inhibitors, and select genes whose score is statistically significant using an appropriate multiple testing procedure;
(iv) estimate the *prediction performance* of the selected GRN using cross-validation techniques.

I have been involved in the last two steps. As we worked with discretized expression data, we chose Mean Absolute Error as a measure of distance between profiles, both for the score in (iii) and for the prediction error in (iv). The statistical significance of the *best GRN* (iii) was assessed by comparing its score to the best score obtained by random permutations of samples in the original gene expression matrix, and we used the conservative approach of Benjamini and Yekutieli [8] in order to ensure that FDR was controlled even though tested hypotheses were not independent. The prediction performance (iv) was assessed using ten-fold cross-validation.

Results. On two standard yeast expression data sets [32, 88], LICORN significantly outperformed Minreg, the state of the art method for unsupervised inference [67]. Focusing only on those genes selected at a given FDR threshold resulted in a further significant decrease in MAE. Biological interpretation of the results showed significant overlap with external biological knowledge in terms of overall network structure, transcription factor-target interactions, and candidate co-regulators.

Implementation. LICORN has been implemented in CaML and is freely distributed at `http://www.lri.fr/~elati/licorn.html`.

4.4.4. Defining true recurrences among ipsilateral breast cancers. This work has been done in collaboration with Marc Bollet and Nicolas Servant [11] (Chapter 7).

Background. Treating early-stage breast cancers with breast-conserving therapy has been demonstrated to yield better quality of life and be more easily accepted than mastectomy[4], while providing equal overall survival [99]. However, patients treated this way run the risk of developing an ipsilateral breast tumor recurrence (IBTR), that is, another tumor on the same breast. In this case, it is of major importance to determine whether the IBTR is a new primary cancer (NP) or a true recurrence (TR) of the first one: when

[3]for LearnIng Cooperative Regulation Networks.
[4]surgical removal of one breast.

it is a new primary, the same treatment as for the primary tumor may be applied, whereas a true recurrence will need more aggressive treatment, because it was not healed by the first treatment.

One of the main challenges is the absence of gold standard definition of NP and TR. The classical *clinical definition* of NP and TR relies on several histopathological characteristics: an IBTR was clinically defined as new primary when the IBTR had occurred in a different location, had a distinct histologic type, or had less aggressiveness features (lower grade, appearance of hormonal receptors) than the initial tumor.

Recently, several studies have suggested to use genomic information to improve this definition: in particular, DNA copy number alterations may be used as markers for clonal relatedness between the primary tumor (PT) and the IBTR. These studies typically perform a hierarchical clustering of PT and IBTR based on DNA copy number alterations, and base the distinction between TR and NP on whether the primary tumor and its local recurrence are neighbors or not on the dendrogram [98].

Motivations. Our study aimed at improving on the current definitions of True Recurrences and New Primaries using SNP arrays, and based on two main ideas:

biological idea: it is not unlikely that two unrelated tumors share genomic alterations, simply because this alteration is a mandatory checkpoint. However, their sharing of *breakpoint locations*, that is, starting and ending points of altered regions, should be a strong indicator of their clonal relatedness;

statistical idea: using the output of a hierarchical clustering to separate NP from TR seems arbitrary and lacks robustness as the corresponding NP/TR classification for a given pair can be changed by the addition or removal of another sample. Instead, basing the definition on a *score* rather than a clustering allows it to be adapted to a desired tolerance in terms of false positives or negatives.

Methods and results. We therefore decided to build a *partial identity score* based on the number of common breakpoints between the IBTR and the PT, each breakpoint being weighted by an estimate of its frequency among and independent set of breast tumors. Its significance was assessed by building *artificial pairs* matching each PT to one of the other IBTR, which allowed us to estimate the distribution for the score under the null hypothesis of no partial identity between the paired tumors.

The quality of the score relied strongly on the precision of breakpoint locations; copy number changes in our SNP data were detected using ITALICS, which outperformed other methods in terms of sensitivity of breakpoint detection, and precision of their location [74].

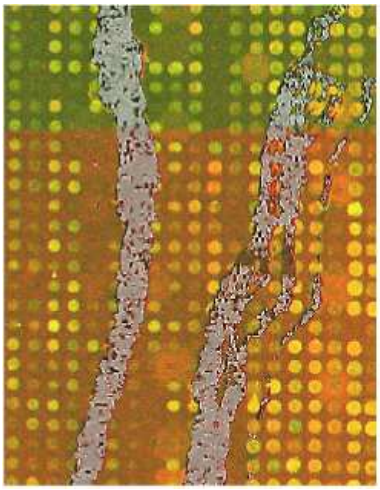
I contributed to the design of the score, and to the resampling-based approach to evaluate its distribution under the null hypothesis. Even though the performance of the score is difficult to assess in absence of gold standard, the score outperformed clinical-based definition in terms of prognosis, that is, in terms of metastasis-free survival.

Further works. This score has been built with the clinical motivation to improve the current definition of new primaries and true recurrences. The new definition is currently used in a biological study that aims at finding genes whose copy number differ between primary tumors whose IBTR is a true recurrence from those whose IBTR is a new primary.

CHAPTER 5

Normalization of DNA copy number microarrays

Art Nouveau



Mary Brinig, *B12n096*, 2002

Gustav Klimt, *Birch Woods* (detail), 1903

BMC Bioinformatics

BioMed Central

Research article

Open Access

Spatial normalization of array-CGH data

Pierre Neuvial*†[1], Philippe Hupé†[1,2], Isabel Brito[1], Stéphane Liva[1], Élodie Manié[3], Caroline Brennetot[3], François Radvanyi[2], Alain Aurias[3] and Emmanuel Barillot[1]

Address: [1]Institut Curie, Service de Bioinformatique, 26, rue d'Ulm, Paris, 75248 cedex 05, France, [2]Institut Curie, CNRS UMR 144, 26, rue d'Ulm, Paris, 75248 cedex 05, France and [3]Institut Curie, INSERM U509, 26, rue d'Ulm, Paris, 75248 cedex 05, France

Email: Pierre Neuvial* - pierre.neuvial@curie.fr; Philippe Hupé - philippe.hupe@curie.fr; Isabel Brito - isabel.brito@curie.fr; Stéphane Liva - stephane.liva@curie.fr; Élodie Manié - elodie.manie@curie.fr; Caroline Brennetot - caroline.brennetot@curie.fr; François Radvanyi - francois.radvanyi@curie.fr; Alain Aurias - alain.aurias@curie.fr; Emmanuel Barillot - emmanuel.barillot@curie.fr

* Corresponding author †Equal contributors

Published: 22 May 2006

BMC Bioinformatics 2006, **7**:264 doi:10.1186/1471-2105-7-264

Received: 15 September 2005

Accepted: 22 May 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/264

Abstract

Background: Array-based comparative genomic hybridization (array-CGH) is a recently developed technique for analyzing changes in DNA copy number. As in all microarray analyses, normalization is required to correct for experimental artifacts while preserving the true biological signal. We investigated various sources of systematic variation in array-CGH data and identified two distinct types of spatial effect of no biological relevance as the predominant experimental artifacts: continuous spatial gradients and local spatial bias. Local spatial bias affects a large proportion of arrays, and has not previously been considered in array-CGH experiments.

Results: We show that existing normalization techniques do not correct these spatial effects properly. We therefore developed an automatic method for the spatial normalization of array-CGH data. This method makes it possible to delineate and to eliminate and/or correct areas affected by spatial bias. It is based on the combination of a spatial segmentation algorithm called NEM (Neighborhood Expectation Maximization) and spatial trend estimation. We defined quality criteria for array-CGH data, demonstrating significant improvements in data quality with our method for three data sets coming from two different platforms (198, 175 and 26 BAC-arrays).

Conclusion: We have designed an automatic algorithm for the spatial normalization of BAC CGH-array data, preventing the misinterpretation of experimental artifacts as biologically relevant outliers in the genomic profile. This algorithm is implemented in the R package MANOR (Micro-Array NORmalization), which is described at http://bioinfo.curie.fr/projects/manor and available from the Bioconductor site http://www.bioconductor.org. It can also be tested on the CAPweb bioinformatics platform at http://bioinfo.curie.fr/CAPweb.

Background

Array-based comparative genomic hybridization (array-CGH) provides a quantitative measure of differences in copy number between two DNA samples [1]. The technique is typically applied to cancer studies because chromosome aberrations frequently occur during tumor progression [2]. Array-CGH facilitates the localization and identification of oncogenes and tumor suppressor genes,

which are likely to be present in chromosomal regions gained and lost, respectively, in cancer cells.

Recent developments in the statistical analysis of array-CGH data have focused on high-level analysis, typically the identification of breakpoints from the genomic profile [3-7], rather than normalization. Most of the normalization techniques used to date for array-CGH data analysis have therefore involved the simple transposition of methods originally designed for expression data [8,9], correcting for differences in the labeling efficiency of the two dyes, spotting effects (block, row, column, or print-tip effects), and local or global intensity dependence of the ratios [10]. As far as we are aware, Khojasteh *et al*. [11] have reported the only method specific to CGH arrays.

Investigation of the systematic sources of variation in the array-CGH data studied showed that the effects affecting expression arrays were negligible with respect to spatial effects of two types. We describe here an algorithm for spatial normalization, which can also be combined with existing normalization methods for handling non-spatial artifacts. We will define and illustrate these two types of spatial effect, and show that such effects are not properly taken into account by traditional normalization techniques.

Two distinct types of spatial artifact

The methods proposed here were originally developed for the analysis of bladder cancer data from tumors collected at Henri Mondor Hospital (Créteil, France) [12], analyzed by hybridization on CGH arrays (F. Radvanyi, D. Pinkel *et al*., unpublished results), including 2464 clones spotted at the University of California San Francisco (UCSF) [13]. They were then adapted to several data sets for CGH arrays produced and hybridized at the Institut Curie, including the breast cancer data (O. Delattre, A. Aurias *et al*., unpublished results) and the neuroblastoma data [14] (which is publicly available [15]) used to illustrate the technique.

We identified two types of spatial effect with fundamentally different natures: *local spatial bias* (Fig. 1(a)) and *continuous spatial gradients* (Fig. 2-1(a)):

Local spatial bias

The array image shows clusters of spots with a discrete signal shift, with the other spots of the array remaining unchanged. These clustered shifted spots on the array image (Fig. 1(a)) have no biological explanation, and correspond to outliers on genomic profiles (Fig. 3(e) and 6(e)). In the data sets studied here, this artifact was found to affect about half of all arrays. We describe it as *local* because it affects only limited areas of the array.

Continuous spatial gradient

The array image shows a smooth gradient in signal from one side of the slide to the other (Fig. 2-1(a)). This artifact leads to genomic profiles with high variability, even between regions with the same DNA copy number. When

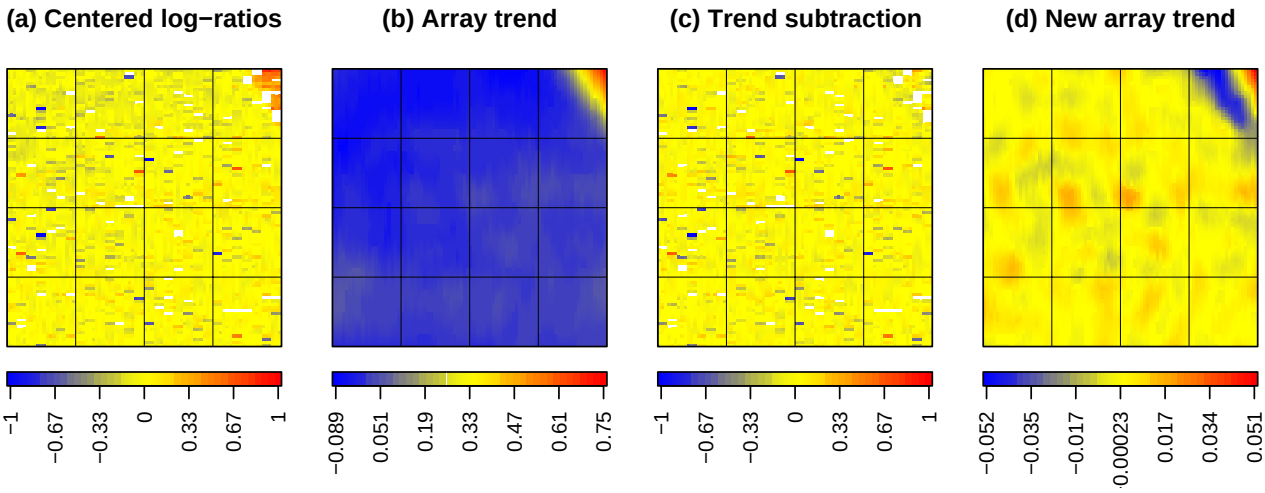


Figure 1

The need for an image segmentation method. An array with areas of local spatial bias (bladder cancer data): a straightforward trend correction method does not address the spatial effect appropriately. (a) Median-centered log-ratios; (b) spatial trend; (c) log-ratios after trend subtraction; (d) remaining spatial trend after subtraction (the color scale is not the same as in (b)). Colors are proportional to signal log-ratios; white dots correspond to missing values.

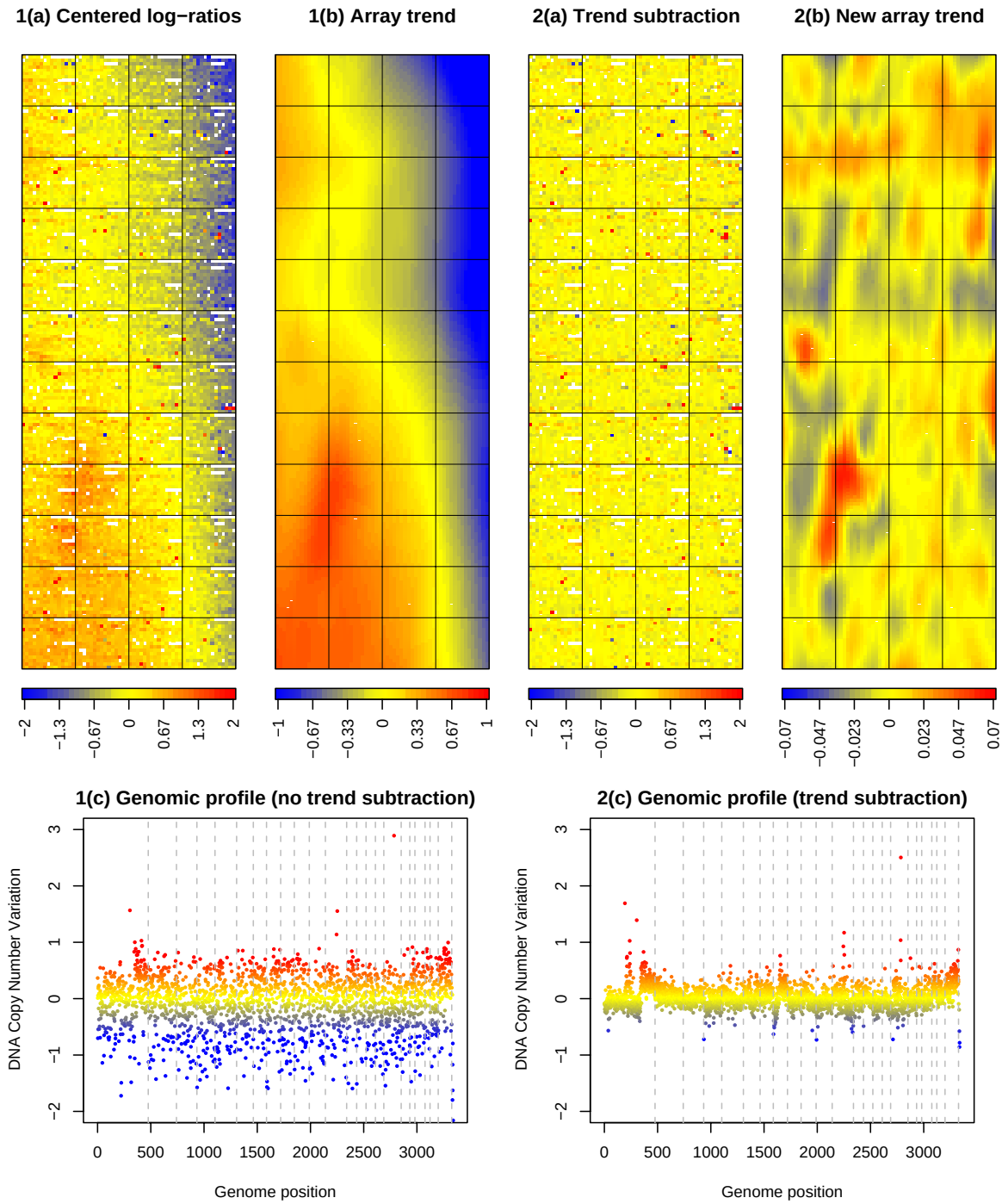


Figure 2

Results of the gradient subtraction step (2dLoess) on a breast cancer array. Correction of the spatial gradient of a breast cancer array: continuous spatial gradients are correctly taken into account by the proposed normalization method. 1(a) Median-centered log-ratios; 1(b) spatial trend; 1(c) genomic profile without spatial normalization; 2(a) corrected log-ratios; 2(b) spatial trend after correction (the color scale is not the same as in 1(b)); 2(c) genomic profile after spatial normalization. The vertical gray dashed lines indicate the separation between chromosomes.

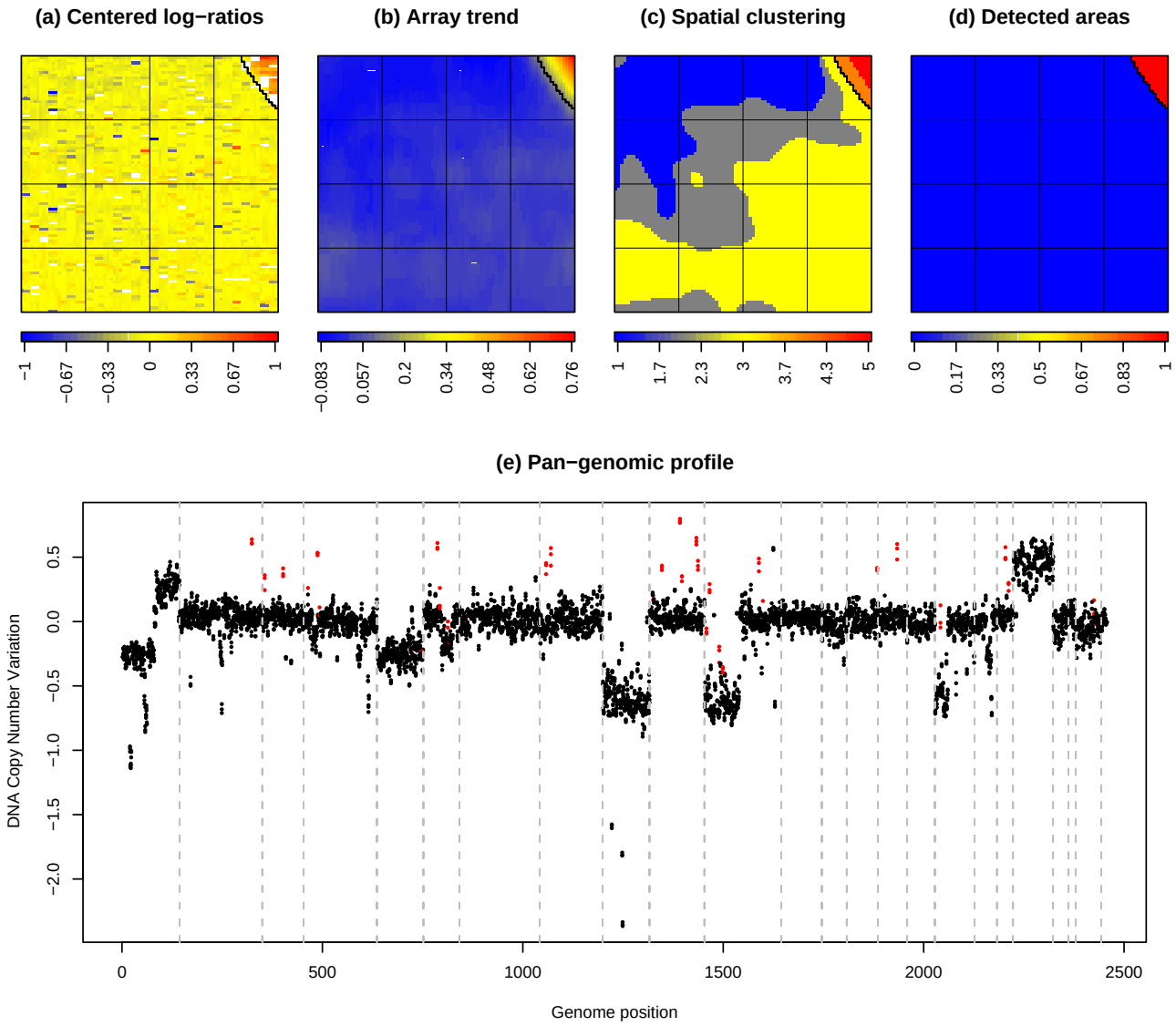


Figure 3

Results of the proposed spatial segmentation method (seg) on a bladder cancer array. Bladder cancer array with local spatial bias accurately detected by the proposed normalization method. (a) Median-centered log-ratios; (b) spatial trend; (c) spatial segmentation; (d) local spatial bias. The border of areas affected by local spatial bias that have been detected in panel (d) are reported on panels (a), (b) and (c) as a black step-function for easy interpretation; (e) genomic profile without spatial normalization (spots detected as local spatial artifacts are marked in red, and the vertical gray dashed lines indicate the separation between chromosomes).

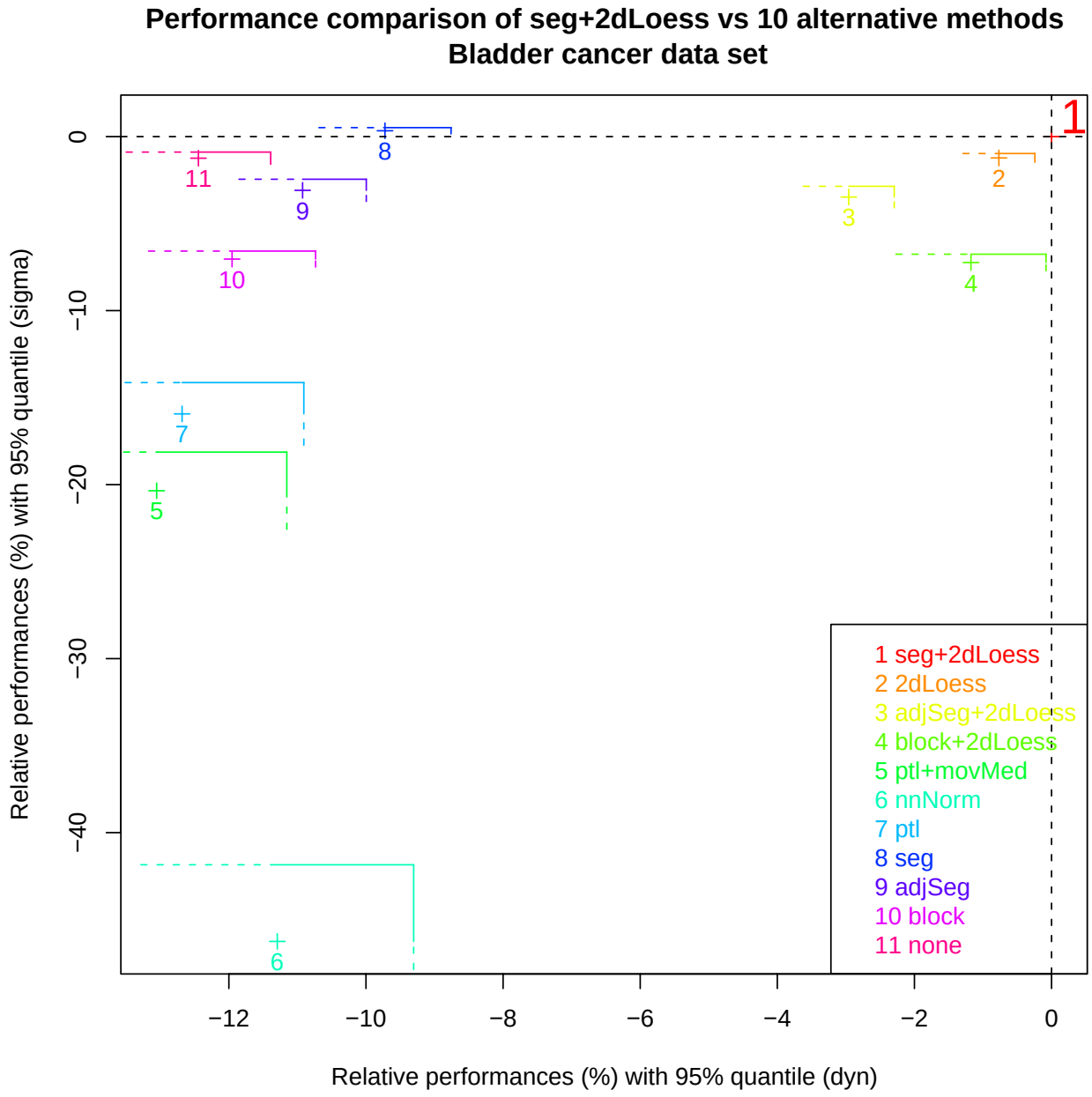


Figure 4

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – bladder cancer data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*, except *seg*, which performs slightly better for *sigma*. Methods 2, 3, and 4, which contain a gradient subtraction step using 2*dLoess*, perform the best against *seg+2dLoess*, as they cluster near the top-right corner of the image. However, *seg+2dLoess* still significantly outperformed these methods for both *sigma* and *dyn*.

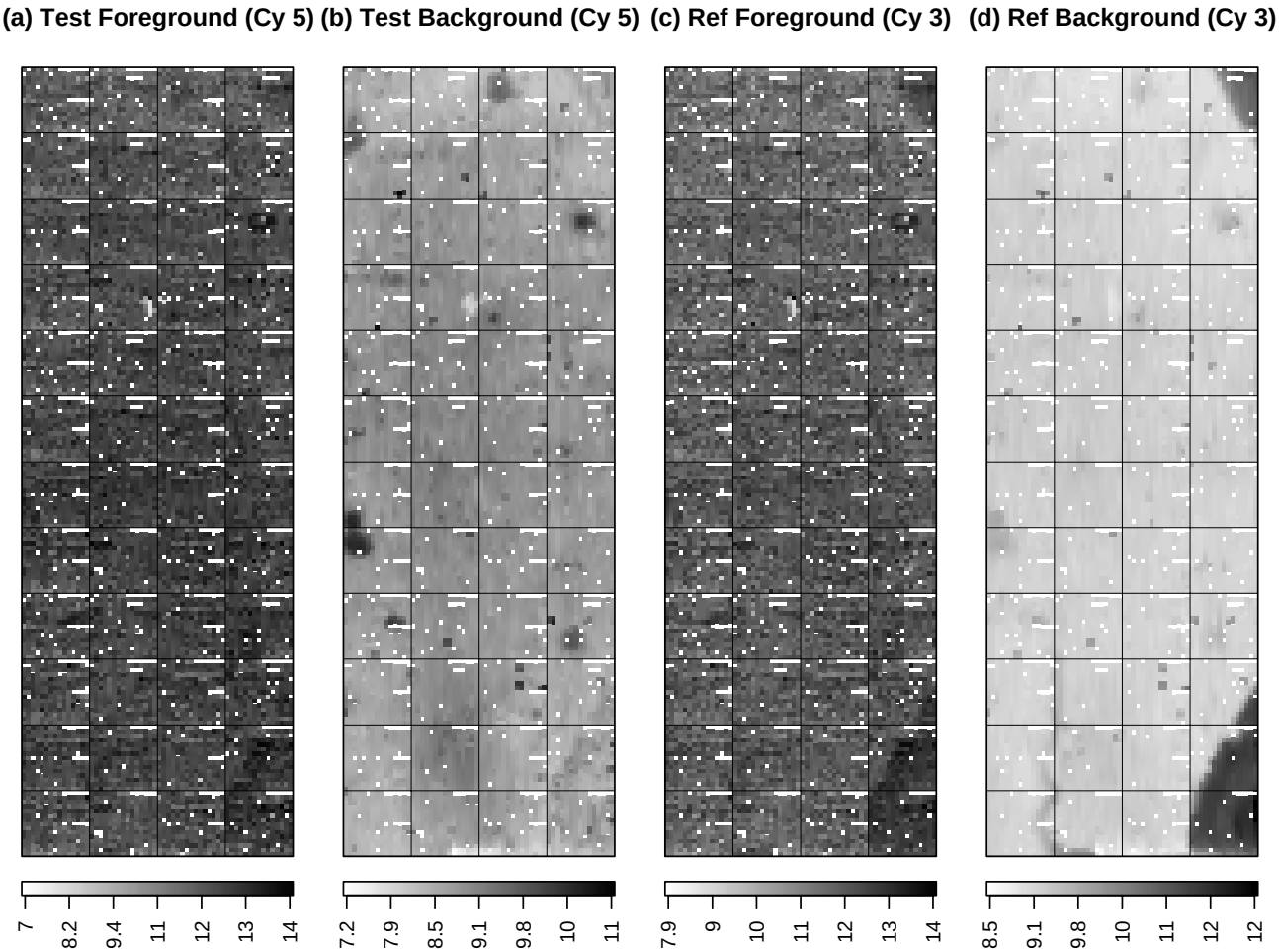


Figure 5
Evidence of local spatial bias on foreground and background raw signals on a breast cancer array. Log-ratios of the four raw signals of a breast cancer array: local spatial biases are easier to detect on a Cy3 background. (a) Test foreground; (b) test background; (c) reference foreground; (d) reference background. Gray-scale level is proportional to signal value.

this effect is observed, it affects all spots to various degrees.

These two types of effect are experimental artifacts of non-biological origin:

- They occur on arrays designed such that neighboring spots on the array correspond to non-neighboring clones in the genome, so there is no obvious biological reason for the clustering of high (or low) signals on the array;

- They are frequently observed on control (normal tissue vs normal tissue) hybridizations, and even on background signals (see Figure 5 for illustration with the breast cancer data set).

The methods proposed are designed to remove or reduce these two types of spatial effect, while preserving the true biological signal.

The need for a spatial segmentation method

The spatial effects described above cannot be attributed to spotting, for two reasons: firstly, they are not limited to array rows, columns or blocks; secondly, they are not reproducible from one array to another, even for arrays taken from batches of slides printed at the same time.

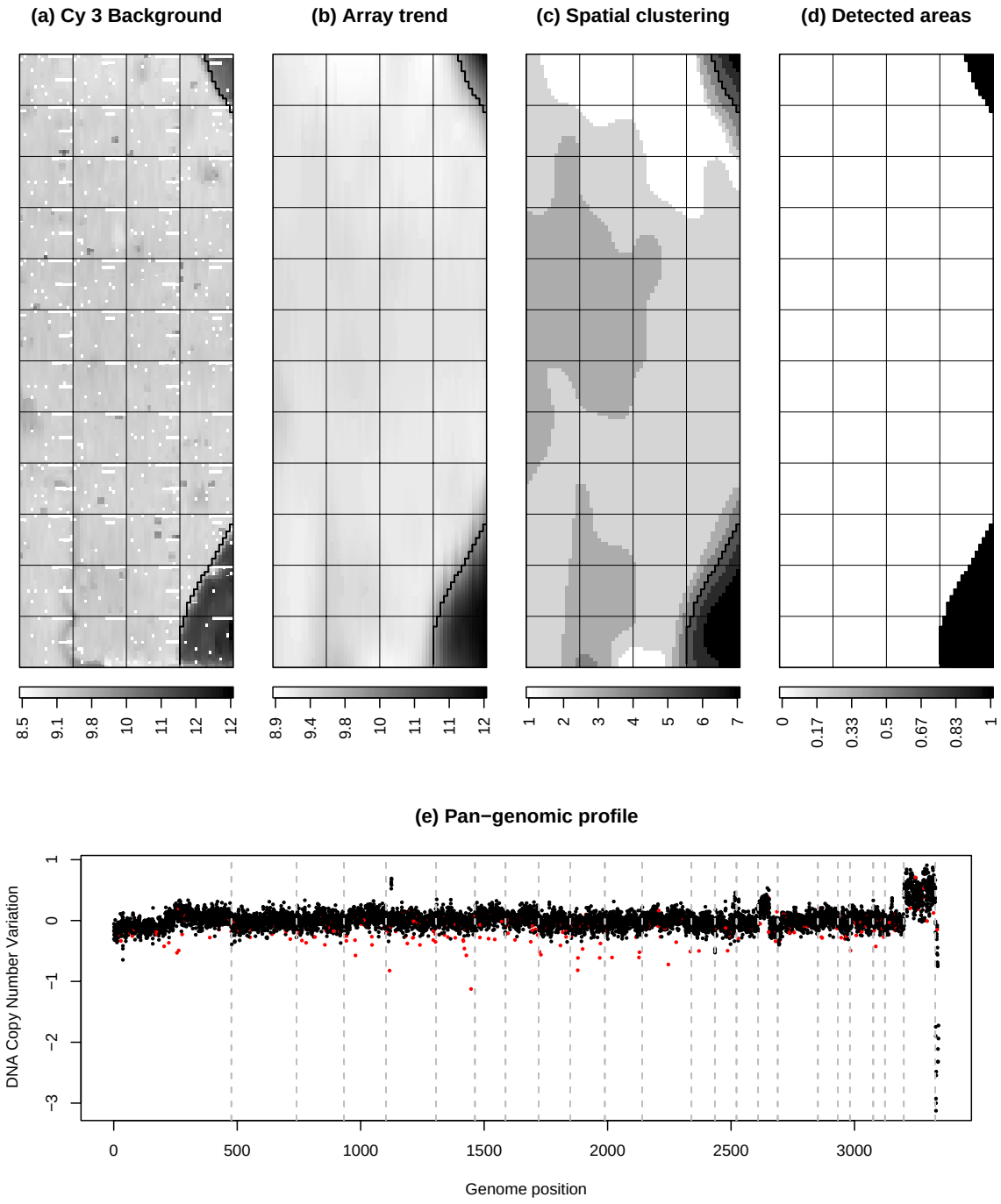


Figure 6

Results of the local spatial normalization step (seg) on a breast cancer array. Breast cancer array with local spatial bias accurately detected by the proposed normalization method. (a) Background signal log-ratios (Cy 3); (b) spatial trend; (c) spatial segmentation; (d) local spatial bias. The border of areas affected by local spatial bias that have been detected in panel (d) are reported on panels (a), (b) and (c) as a black step function for easy interpretation; (e) genomic profile without spatial normalization (spots detected as local spatial artifacts are marked in red, and the vertical gray dashed lines indicate the separation between chromosomes).

Therefore, it is not possible to correct for them properly with the normalization methods generally used for expression arrays, in which "spatial" effects are captured only by row, column, or print-tip group effects. For a method to be appropriate, it must take into account the spatial structure of the array as a whole, and the arbitrary shape of these biased areas.

Several different studies have taken into account spatial effects in expression microarray data and have provided signal correction methods. For example, Workman *et al*. [16] defined a spatial gradient normalization method using a two-dimensional Gaussian function to estimate local background bias in a probe neighborhood. Baird *et al*. [17] proposed a mixed model for cDNA array data, using splines with spatial autocorrelation, assuming the existence of a one-step correlation between adjacent spots in a row or column. Colantuoni *et al*. [18] proposed a method for normalizing the element signal intensities to a mean intensity calculated locally across the surface of a DNA microarray. Others studies have combined intensity-dependent and spatially-dependent effects. Wilson *et al*. [19] have proposed fitting a single LOESS curve on the MA plot and then spatially smoothing the residuals using a median filter to estimate the spatial trend. Tarca *et al*. [20] proposed correcting intensity-dependent and spatially-dependent effects using a feed-forward neural network. Khojasteh *et al*. [11] have compared different CGH array data normalization methods and suggested that a three-step normalization that combines print-tip LOESS with spatial correction using moving median and microplate effect correction gave the best results.

These methods may be suitable for correcting continuous spatial gradients, but they were not designed to detect abrupt changes in signal value across the array, and therefore may not adequately handle local spatial bias: Figure 1 illustrates the need for a spatial *segmentation* method to handle such local spatial effects. From the median-centered log-ratios (a) we estimate a spatial trend (b) by two-dimensional LOESS regression [21,22]; subtracting this spatial trend from the raw values partially corrects the spatial effect (c), but the array trend after correction (d) demonstrates that the spatial effect is undercorrected at the inner border of the biased area, and overcorrected at the outer border, consistent with the observation that signal disturbances vary steeply at the border of the biased area. This systematic overcorrection or undercorrection may lead to misinterpretation in the corresponding genomic profile.

A similar type of spatial effect was reported for expression microarrays by Reimers *et al* [23]. For CGH arrays, this type of effect should be easier to detect and correct, as they have a much smaller range of signal ratio variation than expression microarrays. However, this smaller range necessitates a much greater measurement precision for array-CGH data.

We describe here a spatial segmentation algorithm for the automatic *delineation* and *elimination* of unreliable areas, facilitating the exclusion of local spatial bias from array-CGH data. This algorithm consists of three steps, which are explained in detail in the Methods section:

[step 1]: Estimation of a spatial trend on the array using two-dimensional LOESS regression [21,22]

[step 2]: Segmentation of the array into spatial areas with similar trend values using NEM, an unsupervised classification algorithm including spatial constraints [24,25]

[step 3]: Identification of the areas affected by spatial bias.

A wide variety of microarray techniques based on BACs, cDNAs or oligonucleotides (see [26] for a review) may be used to quantify changes in DNA copy number. From a technical aspect, our method could be applied to any of these microarray types, although we detected local spatial bias only on BAC arrays.

Therefore, we focused on this technology, which has also been the most widely used so far. We provide examples of the implementation of this method and illustrate its performance with three data sets collected on two CGH-array platforms:

- The first data set (bladder cancer data) was produced at the UCSF. In this data set, local spatial effects were observed on 57% of 198 arrays, with a median of 229 affected spots, and no visual evidence of spatial gradients;

- The two other data sets were produced at the Institut Curie, INSERM U509. They consist of a breast cancer data set, in which local spatial effects were observed on 45% of 175 arrays, with a median of 592 affected spots, and a neuroblastoma data set [14,15], with local spatial effects on 23% of 26 arrays, and a median of 551 affected spots.

MANOR: an algorithm combining segmentation and signal correction

In addition to local spatial bias, we also frequently identified continuous spatial gradients, especially in breast cancer data set (Fig. 2-1(a)) and neuroblastoma data set. A straightforward way to correct for spatial gradients (Fig. 2-1(b)) is to subtract from the log-ratios an estimate of the spatial trend on the array (Fig. 2-2(a, b)). The first step of the spatial segmentation algorithm for detecting local spatial bias (step 1) provides such an estimate. This estimate

is calculated using two-dimensional LOESS regression as explained in detail in the Methods section.

In many cases, the CGH arrays were affected by both types of spatial effect: local spatial effects and continuous spatial gradients. In practice, we do not know in advance what type of spatial effect affects a given array. Thus, we propose the following two-step approach:

1. run the spatial segmentation algorithm (*seg*) to identify potential areas of local spatial bias

2. correct spots not excluded during the first step for continuous spatial gradients (*2dLoess*).

This algorithm, implemented in the MANOR package, will be referred to as *seg+2dLoess* in the remainder of this article. The rationale underlying this two-step approach is that arrays affected by continuous spatial gradients only will not be detected as containing local spatial bias by the step *seg*, and will therefore be properly corrected by the step *2dLoess*. This two-step approach is suitable for the spatial normalization of data sets containing both types of spatial effect.

Results and discussion

We have used our method for the spatial normalization of array-CGH data from two different platforms. In this section, we provide information about the practical implementation of the method on these two platforms, and quantitative results comparing our method to ten other normalization techniques. These compare the values of three quality criteria calculated after normalization of each array: the first, *sigma*, estimates the experimental variability between replicates, whereas the others, *smt* and *dyn*, evaluate quality in the context of the estimation of differences in DNA copy number between test and reference samples: *smt* quantifies the smoothness of the signal over the genome, and *dyn* assesses the dynamics of the signal, defined by the signal-to-noise ratio between gained and normal regions; these criteria are defined more formally and explained in detail in the Methods section.

To our knowledge, the ten normalization procedures used for the comparisons cover all the different types of approaches proposed so far and include the methods proposed by Tarca *et al*. [20], Yang *et al*. [10] and Khojasteh *et al*. [11]. These methods are detailed in the Methods section. For each normalization method, we calculated the three quality criteria for each array. When comparing two methods, we calculated a relative performance for each quality criterion, and assessed the significance of this performance using a Student's t-test, as explained in the Methods section. We show that our proposed method outperforms all previously published approaches for the three data sets.

Application to data produced at UCSF

The bladder cancer data set to which our algorithm was applied concerns 198 arrays that were spotted and hybridized at UCSF. These arrays consist of 7392 spots, corresponding to 2464 clones – all of which are BACs (Bacterial Artificial Chromosomes) – with the following design:

- Neighboring clones in the genome are dispersed on the array – a necessary condition for distinguishing between spatial artifacts and real biological information;

- Each clone is replicated three times on the array, and the three replicated spots are adjacent, so a high level of consistency for the three corresponding ratios does not prove that there are no spatial effects.

For this data set, spatial normalization is the last step in the following comprehensive normalization process. After image analysis of the arrays with SPOT 2.0 software [27], we screened for low-quality spots: spots with a foreground reference signal (and foreground DAPI signal) less than 125% of the background reference signal (reference DAPI signal) were discarded, as were clones with a log-ratio standard deviation exceeding 0.1. Clones for which only one of the three replicates was retained after these steps were then also discarded.

Finally, we applied the proposed spatial normalization method *seg+2dLoess* as follows: the spatial segmentation *seg* was applied to the log-ratios of this filtered array, with $K = 5$ and $\beta = 1$ (see Methods for a definition of these parameters and a discussion of how to choose them), followed by the correction for continuous spatial gradients *2dLoess*.

Spatial normalization step

Our segmentation algorithm detected local spatial effects on 113 of 198 bladder cancer arrays (57%); the median proportion of biased areas on these arrays was 3.1%. Figure 3 (top) illustrates the successive steps of the algorithm, from centered log-ratios to array trend, spatial segmentation of the array, and finally the delineation of biased areas. Red dots on the corresponding genomic profile (Figure 3, bottom) correspond to the spots discarded during spatial normalization (on this figure, signal log-ratios have not yet been averaged by clone: *spot-level information* is displayed).

Figure 3 (bottom) illustrates the improvement in data quality achieved with our spatial normalization method: among the apparent outliers (i.e. clones with log-ratio values significantly different from the mean log-ratio value

for the genomic region), it distinguished between experimental artifacts (red dots) and potentially biologically relevant outliers accounting for localized genomic amplifications.

Evaluation of the performance of the seg+2dLoess *method*

For each normalization method (11 methods including ours), we calculated the three quality criteria for each array and performed pairwise comparison of methods using the estimate and significance of their relative performance for each criterion, as explained in detail in the Methods section.

Figure 4 shows the results of comparison of the ten methods with *seg+2dLoess*. For the *dyn* criterion, *seg+2dLoess* significantly outperformed all methods (with all p-values ≤ 0.039), and most significantly methods 5 to 11, that do not include the *2dLoess* step (with all p-values below 8.5×10^{-18}). The *dyn* criterion is particularly important as it assesses the quality of copy number change detection. *seg+2dLoess* also gives significantly better results for the *sigma* criterion than all other methods (with all p-values below 1.1×10^{-8}) except one: *seg* performs significantly better ($p = 7.9 \times 10^{-4}$) but the relative improvement has a limited amplitude (only 0.36%).

For the *smt* criterion, *seg+2dLoess* also significantly outperforms all methods (with all p-values below 8.1×10^{-6}, except *block+2dLoess* for which $p = 0.048$).

Section 1 of the Additional file 1 shows similar plots to Figure 4, but for the *smt* and *dyn* criteria, and for the *smt* and *sigma* criteria. Tables 1 to 3 of the Additional files 2 and 3 summarize the results of all the pairwise comparisons of methods for the three quality criteria.

Taken together, these results show that the seg+2dLoess method outperforms its competitors for the bladder cancer data set.

Application to data produced at Institut Curie, INSERM U 509

The Institut Curie, INSERM U509 has developed its own high-density CGH array; all steps in the production of these chips are performed in Institut Curie laboratories, including array spotting, DNA preparation, hybridization, scanning and image processing. The current version of the array contains 3342 clones, each of which is spotted at least three times on the array, giving a total of 10800 to 11520 spots (including controls).

This array was designed to facilitate distinction between relevant biological effects and experimental artifacts: "empty" spots and spots of water were included as controls, clone replicates were scattered over the array, and the positions of clones on the array are not correlated with their actual positions in the genome. A reliable ratio value can therefore be calculated even if one of the three replicates is flagged. The arrays were scanned using an Axon Genepix 4000b scanner, and images were processed with Genepix Pro 5.1.

We analyzed a breast cancer data set and a neuroblastoma data set from this platform.

For this platform, we applied the proposed spatial normalization method *seg+2dLoess* as follows: the spatial segmentation *seg* was applied to the Background signal as explained in the paragraph below, and the spatial gradients were corrected by *2dLoess* calculated over the log-ratios. A post-processing step that includes spot and clone screening was then applied (allowing us, for example, to discard spots having too low a signal-to-noise ratio, or with poor replicate consistency).

Detail of the spatial segmentation step

Although we can correct the foreground signal for background intensity, a significant proportion of arrays still show localized spatial patterns that cannot be attributed to biological causes. Visual examination of spatial representations of the four signals (foreground and background intensities for test and reference signals) revealed that the bias was much clearer for the background signal of Cy3-labeled samples (Figure 5), which was not the case for bladder cancer data. We therefore applied the spatial segmentation method described above to the background signal of the Cy3 channel, with $K = 7$ and $\beta = 1$ (see Methods for a definition of these parameters and a discussion of how to choose them).

Biased areas of the CGH array are flagged and excluded from subsequent analysis. As clone replicates are not adjacent on the array, at least two of the three replicates generally remain after spatial bias correction, and a reliable ratio value can still be calculated. Figure 6 shows the results of this spatial segmentation step in the case of an array with local spatial bias but no spatial gradients.

Evaluation of the performance of the method seg+2dLoess

As for bladder cancer data, we calculated the three quality criteria for each normalization method and for each array for the breast cancer data set and the neuroblastoma data set. We then compared the methods paiwise using the estimate and significance of their relative performance for each criterion, as explained in detail in the Methods section.

Figures 7 and 8 show the results of comparing the ten methods with *seg+2dLoess* for the *dyn* and *sigma* criteria. *seg+2dLoess* significantly outperforms all other methods

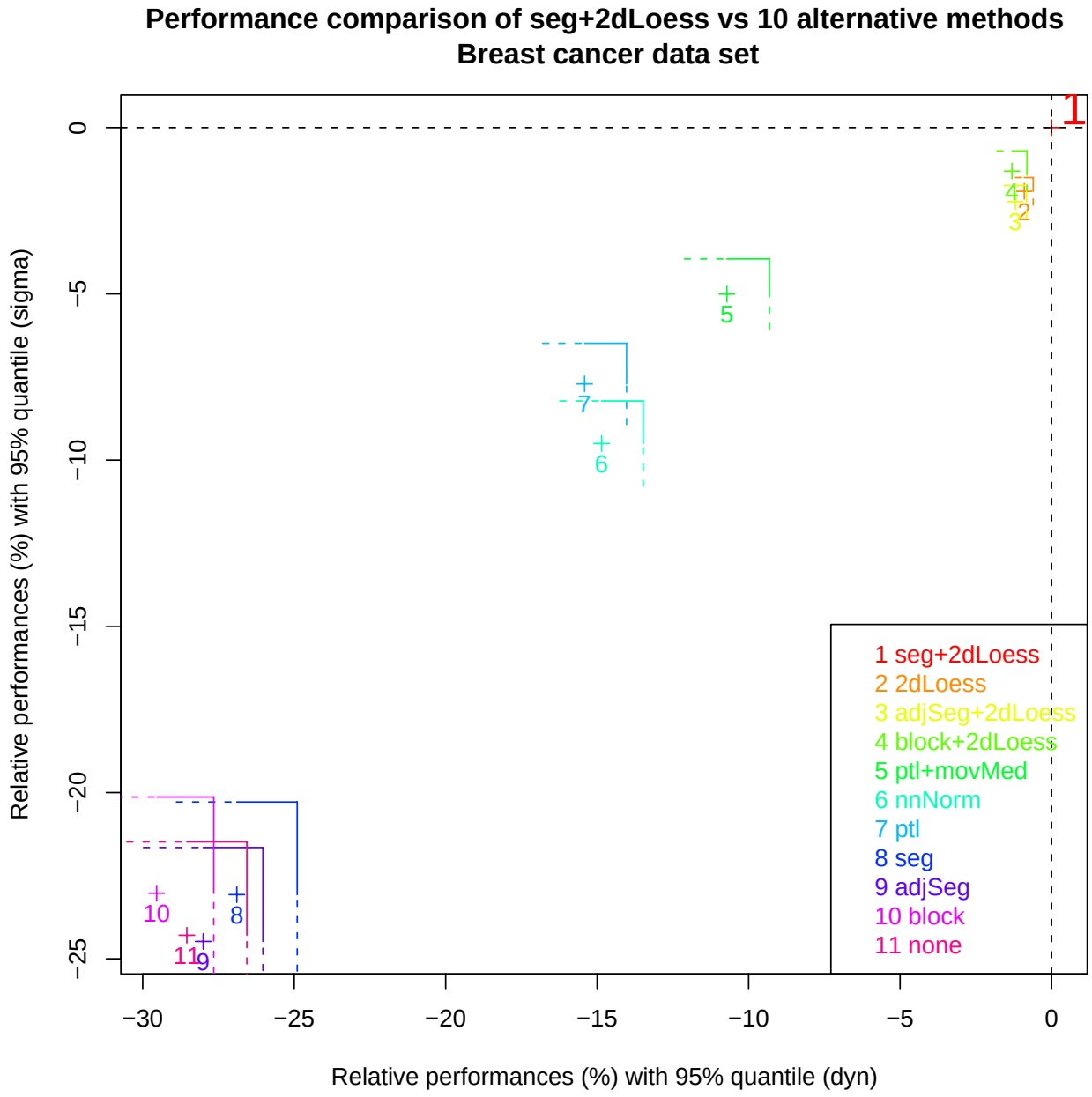


Figure 7

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – breast cancer data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*.

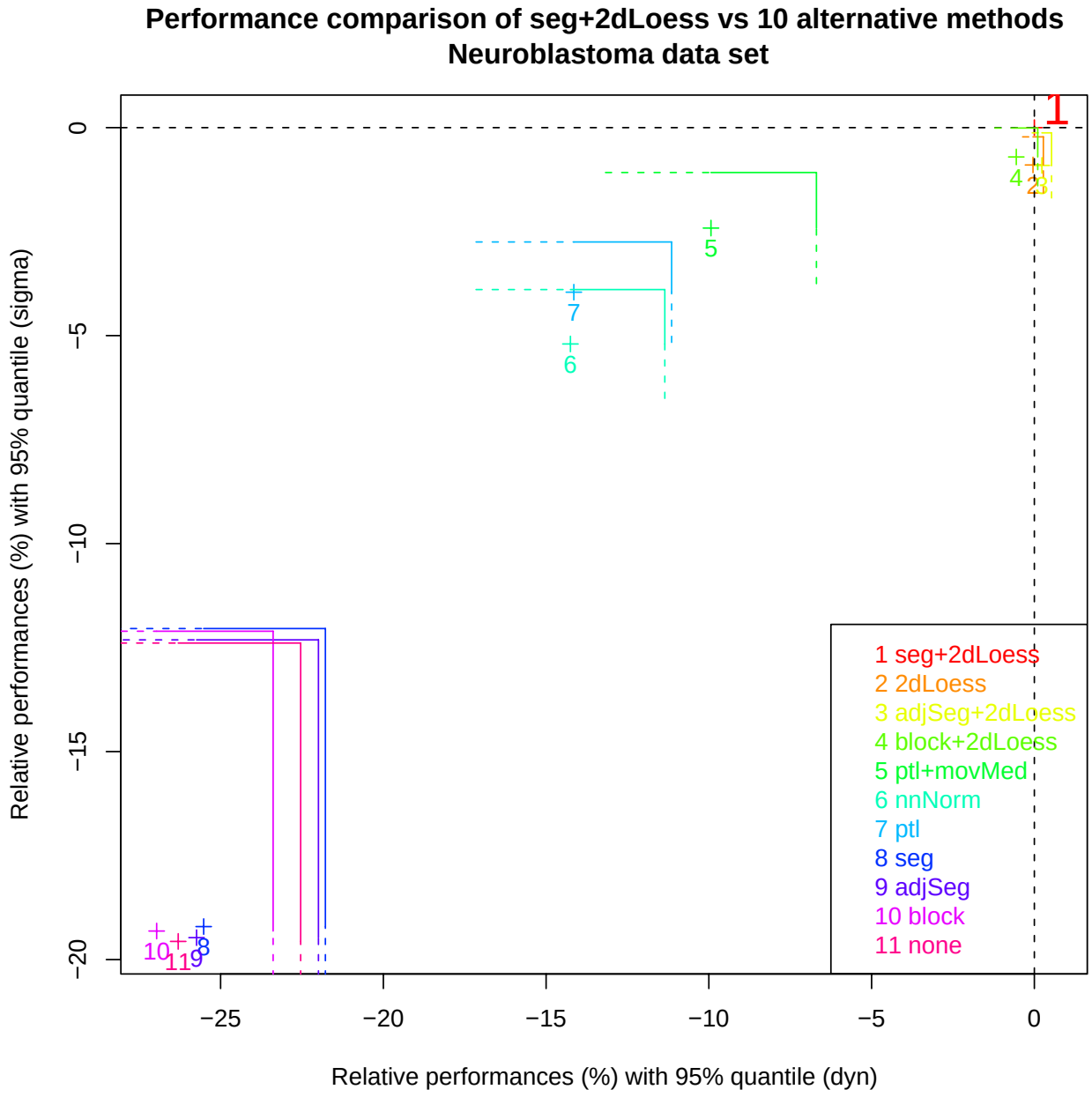


Figure 8

The proposed method (seg+2dLoess) compares favorably to all other normalization methods – neuroblastoma data set. We compared the proposed method (*seg+2dLoess*) to ten methods for two quality criteria: *sigma* and *dyn*. Each color corresponds to the comparison of *seg+2dLoess* with a different method. The proposed method is taken as a reference (red point 1 at (0,0)). For each method *i*, the cross indicates the mean relative performance (see methods section) of the data set for *dyn* (x axis) and in *sigma* (y axis), and the lines give the corresponding 95% quantile of relative performance. For *sigma* (*dyn*, respectively), the methods with a 95% quantile below (left to, respectively) the horizontal (vertical, respectively) dashed black line are significantly outperformed by our proposed method. Here *seg+2dLoess* significantly outperforms all methods for *dyn* and *sigma*, except those containing a gradient subtraction step with *2dLoess*.

for the three criteria on the breast cancer data set (with all p-values below 2.3×10^{-4}).

The neuroblastoma data set gives similar results: *seg+2dLoess* quality criteria are always better than those of the other methods, except for *dyn*, in which *adjSeg+2dLoess* is slightly better (0.22%) but not significantly so ($p = 0.1$). For *smt*, *seg+2dLoess* is only slightly better than *ptl+movMed* and the methods including the *2dLoess* step, but not significantly so for *adjSeg+2dLoess* and *ptl+movMed*. In these cases, the small size of the data set (26 arrays, 6 with local spatial bias) affects the statistical power.

Section 2 and 3 of the Additional file 1 and Tables 4 to 9 of the Additional files 2 and 3 detail and complement these results.

These results show that the *seg+2dLoess* method outperforms the other methods on the two data sets produced on the Institut Curie, INSERM U509 platform. The results also allow the methods to be ranked in terms of performance. Those methods that include a two-dimensional LOESS step are the highest ranked, with the methods proposed by [11,10] and [20], which all include some spatial processing, being next, and the other methods being the lowest ranked (see Figure 7 for example).

Conclusion

We have designed an efficient and automated algorithm for the spatial normalization of BAC array-CGH data, and defined a set of parameters for CGH array data quality assessment. We have shown that our method significantly improves the quality of data from two different BAC-array platforms and outperforms other normalization techniques on three data sets.

The proposed algorithm is particularly suitable for correcting spatial effects not related to array design (row, column, or print-tip group effects): indeed, the arrays studied show two distinct types of such spatial effect (local spatial bias and continuous spatial gradients), which can simultaneously affect any given array. In such cases, using spatial trend correction after spatial segmentation helps to remove or reduce these two types of spatial effect, while preserving the true biological signal.

This method is original in the application of a segmentation algorithm for detecting and removing local spatial bias, preventing the misinterpretation of experimental artifacts as biologically relevant outliers in the genomic profile.

This method was developed for array-CGH experiments, and gave very good results. However, it can be applied to any microarray experiment having the same types of spatial effect.

Availability and requirements

Our method is implemented in the R package MANOR (Micro-Array NORmalization) [28], which is available from the Bioconductor site [29]. It can also be tested on the CAPweb bioinformatics platform [30,31].

Methods

In this section, we provide details of the segmentation method and the other normalization techniques used for comparison, and of the quality criteria proposed. We also discuss the choice of the two parameters of the segmentation algorithm: K and β.

Description of the segmentation algorithm (seg)

The segmentation method consists of three steps:

[step 1]: Estimation of a spatial trend on the array using two-dimensional LOESS regression [21,22]

[step 2]: Segmentation of the array into spatial areas with similar trend values, using NEM, an unsupervised classification algorithm including spatial constraints [24,25]

[step 3]: Identification of the areas affected by spatial bias.

[step 1]: spatial trend estimation

We decided to carry out spatial segmentation based on an estimate of the spatial trend on the array, to optimize the robustness of segmentation. Furthermore, estimation of this trend makes it possible to replace missing values by interpolating the spatial trend.

The trend is estimated by means of a two-dimensional LOESS procedure with three iterative reweighting steps [21,22]. The local estimation is linear and the neighborhood taken into account to fit the local model corresponds to 3% of the total number of points. We use an iterative reweighting procedure to avoid outlier effects. Indeed, in the context of cancer studies, we are investigating changes in DNA copy number, and some clones displaying an amplification or a homozygous deletion may generate extreme but biologically meaningful values, which should not be interpreted as a local spatial bias.

When the spatial trend is estimated from the log-ratios, we first apply a basic correction to these log-ratios to prevent confusion between spatial artifacts and biologically relevant effects. For each chromosome arm, *centered* log-ratios are calculated as follows: the median of the corresponding log-ratio values is calculated and then subtracted from the initial values. The spatial trend is estimated from these centered log-ratios. This method helps to decrease the

impact of true genomic aberrations on the detection of spatial trends in the data, particularly for samples with many, or large genomic alterations, as most of these alterations correspond to the gain or loss of whole chromosome arms.

[step 2]: spatial segmentation

This step aims to identify K clusters corresponding to spots with similar signal levels located close together geographically. This is achieved by Neighborhood Expectation Maximization (NEM) [24,25]. We assume that the data are drawn from a mixed Gaussian density function $f(\mathbf{x}_i \mid \mathbf{\Phi}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k)$ where p_k are the proportions of the mixture model, $f_k(\mathbf{x}_i \mid \theta_k)$ denotes the density function of a Gaussian distribution with parameter $\theta_k = (\mu_k, \Sigma_k)$ and $\Phi = \{p_1, ..., p_k, \theta_1, ..., \theta_K\}$ is the set of parameters to be estimated. The classical EM algorithm considers the following decomposition of the likelihood:

$$L(\mathbf{c}, \mathbf{\Phi}) = \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log p_k f_k(\mathbf{x}_i \mid \boldsymbol{\theta}_k) - \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log c_{ik} \qquad (1)$$

where

$$c_{ik} = \frac{p_k f_k(\mathbf{x}_i \mid \theta_k)}{f(\mathbf{x}_i)} \text{ and } \mathbf{c} = (c_{ik}) \qquad (2)$$

In the mixture model context, [32] pointed out that the EM algorithm is formally equivalent to the alternative maximization of $L(\mathbf{c}, \Phi)$ with respect to $\mathbf{c}$ ("E" step) and with respect to Φ ("M" step). The NEM algorithm is original in that it regularizes the likelihood by means of a term that takes into account the spatial dimension of the problem through the following adjacency matrix:

$$v_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

Here, the neighbors of a point located at coordinates (l, m) are the four points with the following coordinates: (l +1, m), (l - 1, m), (l, m - 1). We define the following quantity:

$$G(\mathbf{c}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} c_{ik} c_{jk} v_{ij} \qquad (3)$$

Thus, instead of maximizing $L(\mathbf{c}, \Phi)$ in the E step, we maximize $L(\mathbf{c}, \Phi) + \beta G(\mathbf{c})$. The value of β controls the weighting of the geographical context in the maximization. The M step remains unchanged.

[step 3]: elimination of local spatial bias

The basic idea is to remove from the array those spatial clusters with signal values significantly higher (or lower) than the unbiased areas of the array. We describe here the situation for positive spatial bias, but the idea can be adapted to negative bias. As local spatial biases cover a limited proportion of the array, we introduced a tuning parameter p_{max}, which corresponds to the maximum proportion of the array image corresponding to local spatial bias. In our experiment, local spatial bias typically applies to less than one quarter of the array, so we used p_{max} = 0.25.

After sorting the clusters identified by NEM by decreasing mean signal, we consider only those clusters with cumulative frequencies lower than p_{max} to be potentially biased, making it possible to define a set of candidate clusters. The mean signal value of the remaining clusters is used as a reference value for the unbiased signal. Each candidate cluster with a mean signal differing from this reference value by more than a given threshold value is considered biased. The other candidates are considered unbiased, unless their mean signal is closer to that of the biased cluster than to that of the reference: such clusters are also considered biased. This threshold was chosen based on the cross-validation of arrays analyzed by experts.

Comparison to other normalization methods

We compared the described methodology with other classical normalization methods. All these methods are listed below:

- A print-tip group method:

block (block normalization): we subtract off the row and column block median log-ratio values for each spot, and adds back the overall block median log-ratio value.

- A print-tip group with intensity dependent effect method:

ptl (print-tip loess): we apply the print-tip LOESS normalization [10] method using the marray R package (1.8.0 release, with default parameters) available from Bioconductor.

- A spatial smoothing method:

2dLoess (correction of continuous spatial gradients): a spatial trend is estimated by two-dimensional LOESS [21,22], which is then substrated from the log-ratio values.

- Two spatial segmentation methods:

seg (segmentation of local spatial bias): we apply the spatial segmentation algorithm described above to automatically eliminate the biased area.

adjSeg (correction of local spatial bias): we apply the spatial segmentation algorithm to automatically delineate the biased area. The median log-ratio value of such an area is then adjusted to the median log-ratio value of the unbiased area.

- A method combining print-tip group and spatial smoothing:

block+2dLoess (block normalization and global correction): we apply the *2dLoess* method on the normalized log-ratio values obtained with *block*.

- Two methods combining intensity dependent effect and spatial smoothing:

nnNorm (neural network normalization): we apply the normalization method described by Tarca *et al*. [20] using the nnNorm R package (1.5.1 release, with default parameters) available from Bioconductor. Briefly, this technique uses a neural network approach to correct the intensity-dependent and spatially-dependent effects.

ptl+movMed (print-tip loess and moving median filter): Khojasteh *et al*. [11] compared different normalization methods and suggested that combining the print-tip LOESS method with spatial correction (using a moving median calculated over a neighborhood of 11 rows by 11 columns) and microplate correction gave the best results. As the microplate information was not available in our data, we discarded the third step and only considered the print-tip LOESS and spatial correction.

- Two methods combining spatial segmentation and spatial smoothing:

adjSeg+2dLoess (correction of local spatial bias and continuous spatial gradients): we apply the *2dLoess* method on the normalized log-ratio values obtained with the *adjSeg* method.

seg+2dLoess (local segmentation and correction of continuous spatial gradients): we apply the *2dLoess* method on the log-ratio obtained with the *seg* method.

- Raw log-ratio values with no normalization (**none**).

Array-CGH data quality assessment

Definition of quality criteria

Evaluation of the quality of the signal ratios of an array facilitates the comparison of different image analyses or normalization algorithms, and makes it possible to quantify the improvement achieved by each step of a given normalization algorithm. We define three criteria for assessing the quality of the analyzed array: the first addresses the issue of overall quality whereas the other two provide quality evaluations for the estimation of differences in DNA copy number between test and reference samples.

sigma The first item provides an estimate of experimental noise. We isolate each clone and calculate the standard deviation of the log-ratio of the corresponding replicates. *sigma* is defined as the median of these standard deviations: the smaller the value of *sigma*, the higher the quality of the array.

The other two criteria are calculated after detection of the altered (gained or lost) regions in the test sample. We used the GLAD algorithm, developed by Hupé *et al*. [4] for this purpose:

smt Within a given DNA copy number region, the ratios of contiguous clones should not differ considerably. The second quality criterion concerns the *smoothness* of the signal log-ratios within such a chromosomal region: signal smoothness is defined as the median absolute difference between log-ratios for contiguous normal clones. If N denotes the set of clones considered normal after DNA copy number estimation, we can calculate

$$smt = \text{median}_{n \in N} |x_{(n)} - x_{(n-1)}|,$$

where $x_{(n)}$ is the value of the log-ratio at the n^{th} clone in genome order.

dyn The last criterion estimates the *dynamics* of DNA copy number variation between test and reference samples. We calculate the discrepancy between the median ratios of the regions considered "gained"(G) and "normal"(N) after DNA copy number estimation, and compare it with signal smoothness, as measured by *smt*:

$$dyn = \frac{\text{median}_{g \in G} x_g - \text{median}_{n \in N} x_n}{smt}$$

If no gained region is detected, we compare "normal" regions with "lost"(L) regions.

smt and *dyn* are not independent parameters and are anti-correlated. However, they quantify related but different ideas, as *smt* estimates the noise level after data normalization whereas *dyn* measures the ability to detect genome alterations after data normalization.

Paiwise comparison of quality criteria

These three criteria help us to decide which of two normalization methods gives the best results for a given array. In this pairwise comparison context, *smt* and *dyn* must be calculated with the same definition of *G*, *N*, and *L* regions for the two normalized arrays. We therefore define consensus *G*, *N*, and *L* regions associated with an array processed with two different normalization methods as the intersection of the two corresponding *G*, *N*, and *L* regions obtained using the two different normalization methods.

In order to test whether method *j* is better than method *i*, we defined a relative performance for each quality criterion as follows:

$$\begin{cases} RP^{sigma}(i,j) &= \dfrac{sigma(i) - sigma(j)}{sigma(i)} \\ RP^{smt}(i,j) &= \dfrac{smt(i) - smt(j)}{smt(i)} \\ RP^{dyn}(i,j) &= \dfrac{dyn(j) - dyn(i)}{dyn(i)} \end{cases}$$

We calculated this relative performance for each array, and assessed its significance by testing the hypotheses $\mathcal{H}_{i,j}: \{RP^{qc}(i,j) < 0\}$ for each quality criterion *qc*, using a Student's unilateral t-test.

In figures 4, 7, and 8, we calculated relative performances *RP*(*seg+2dLoess*, *test*) where *test* corresponds to one of the ten other methods. Hence a *negative value* for *RP* (*seg+2dLoess*, *test*) indicates that our proposed method outperforms the *test* method.

Parameter choice for the segmentation algorithm

The segmentation algorithm includes two parameters: the number *K* of clusters, and the regularization parameter β, which controls the weighting of geographic context in signal segmentation. Our experience suggests that the optimal choice of *K* and β may depend on the array-CGH technology used. We therefore provide guidelines for the choice of suitable parameters of the algorithm. We have investigated two different approaches to the choice of (*K*, β): incorporating a model selection criterion into the algorithm so that an optimal (*K*, β) can be chosen *for each array*, or developing a calibration method to help the user to find relevant sets of parameters for analyzing *a whole data set*. In this section, we discuss these two approaches and justify our choice of the second solution.

The difficulty finding optimal parameters on a per array basis

Choice of the number *K* of components in a mixture model can be addressed using model selection criteria. The basic idea is as follows: as the maximum likelihood estimator of the model increases mechanically with *K* (as model complexity increases with *K*), this method subtracts an increasing function of *K* from the likelihood of the model with *K* components, to prevent model overfitting. Many applications use the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) for this purpose. However, in our framework, *K* and β must be chosen simultaneously, because β also affects the maximum likelihood estimator. As we have no information concerning the quantitative behavior of the maximum likelihood estimator with respect to *K* and β (this complex question is beyond the scope of this paper), the choice of an appropriate penalization remains arbitrary.

We also considered an approach involving the fitting of *K* using model selection criteria and cross-validating the choice of β, but this approach has major drawbacks: first, it strongly increases the complexity of the estimation process, making this method too time-consuming for use as a routine normalization method; second, it makes the normalization method difficult to interpret, because two arrays from the same platform will not be treated with the same parameters.

Guidelines for choosing relevant parameters for analyzing a new data set

Rather than searching for optimal (*K*, β) values for each array, we provide a calibration method making it possible to choose appropriate (*K*, β) values for each data set. The basic principle of the calibration method is comparison of the output of our algorithm run on different (*K*, β) pairs, taken from a pre-defined grid (e. g. $K \in \{2,... 10\}$ and $\beta \in \{0.1,0.2,...2.0\}$).

We considered two different approaches to compare the results of the segmentations and to choose appropriate (*K*, β) values. The first approach involved choosing a (*K*, β) combination that optimizes quality criteria. The second involves expert assessment. An expert examines each array from a representative set and determines whether there is local spatial bias: he or she checks both the array image and the genomic profile to guarantee that the spatial effect is due to an experimental artifact rather than a biological effect. We then select the (*K*, β) combination that gives the best agreement between the expert decision and the algorithm decision. We call this second approach *expert assessment*. We found this second method simpler and more efficient than the first, for a number of reasons, outlined below.

In the first approach, quality criteria are calculated after normalization and DNA copy number assessment, so these three steps have to be carried out for each (*K*, β) combination. Therefore, although this method has the

obvious advantage of not relying on expert assessment, it is time-consuming, and provides only indirect evaluations of the differences between pairs of parameters, which may make the results hard to interpret. Moreover, a much lower level of variation was observed in the values of quality criteria for different (K, β) combinations for a given array than between arrays, so we were unable to identify optimal (K, β) values with this method (data not shown).

In the second approach, we considered two different ways of performing the expert assessment: either identifying arrays displaying local spatial bias (qualitative assessment), or estimating the number of spots that should be discarded (quantitative assessment). We found quantitative assessment to be very poorly reproducible, with large differences between experts, and much more time-consuming than the qualitative method. Therefore, we adopted the qualitative method, which made possible the rapid expert assessment of a larger number of arrays, thus increasing the accuracy of parameter choice.

Based on the qualitative expert assessment of an entire data set or a subset of data, we compare, for each array, the decision of our algorithm (has the algorithm detected a local spatial bias?) with that of the expert. We then calculate the proportion of false positives and false negatives for each combination of the parameters $K \in \{2,...10\}$ and $\beta \in \{0.1, 0.2,... 2.0\}$. Qualitative expert assessment remains highly variable (significant differences between experts), as a substantial proportion of arrays are difficult to classify. Nevertheless, all assessments show the same form of dependence in the error rate in (K, β), and lead to selection of the same parameters (data not shown).

For illustration, we use a subset of arrays on which two different expert assessments agree. The analysis is shown in Figure 9 for breast cancer data (134/179 arrays), and

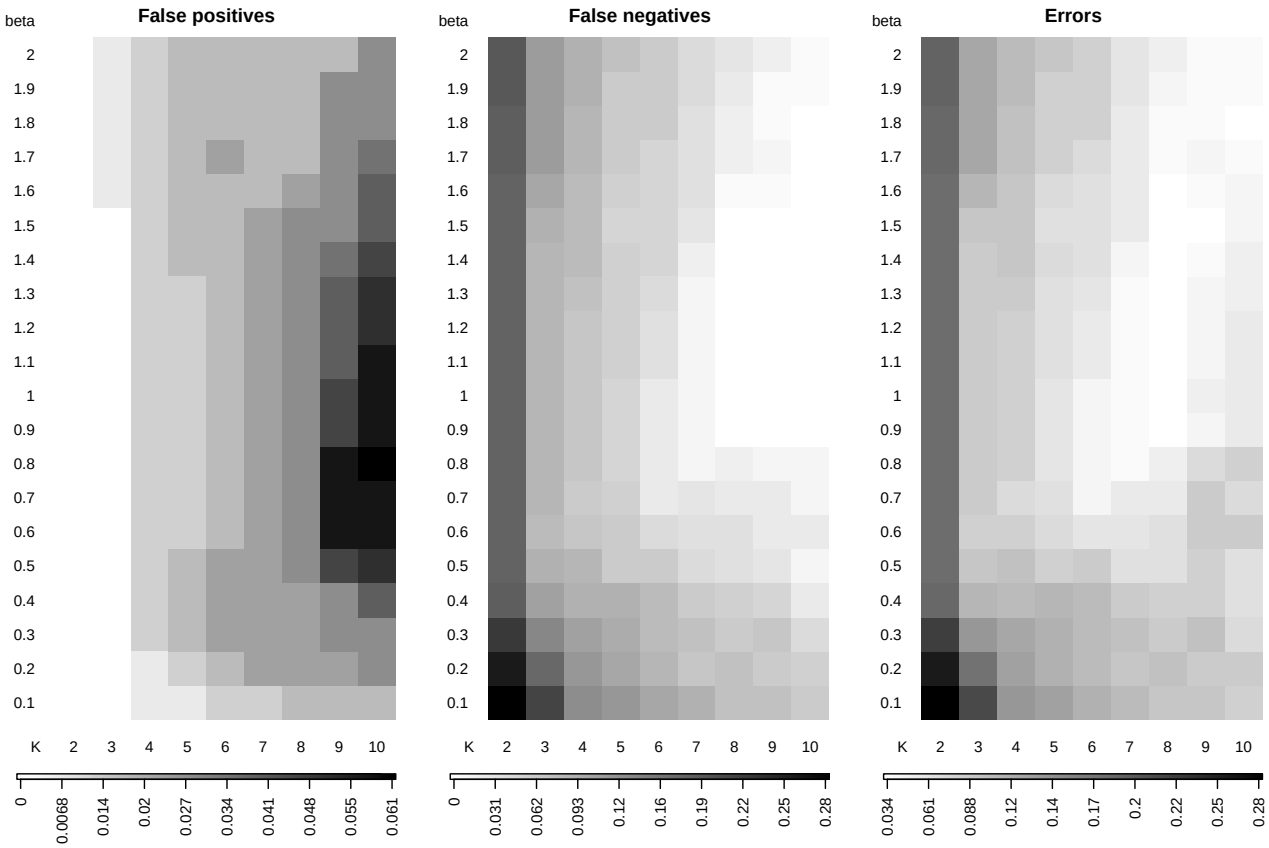


Figure 9

Comparison between qualitative assessment and segmentation results with various (K, β) –breast cancer data set. Thesegmentation algorithm is run with $K \in \{2,...10\}$ (x axis) and $\beta \in \{0.1, 0.2,...2.0\}$ (y axis) and compared with the expert assessment of the breast cancer data set. (a) False positive rate; (b) False negative rate; (c) Total error rate.

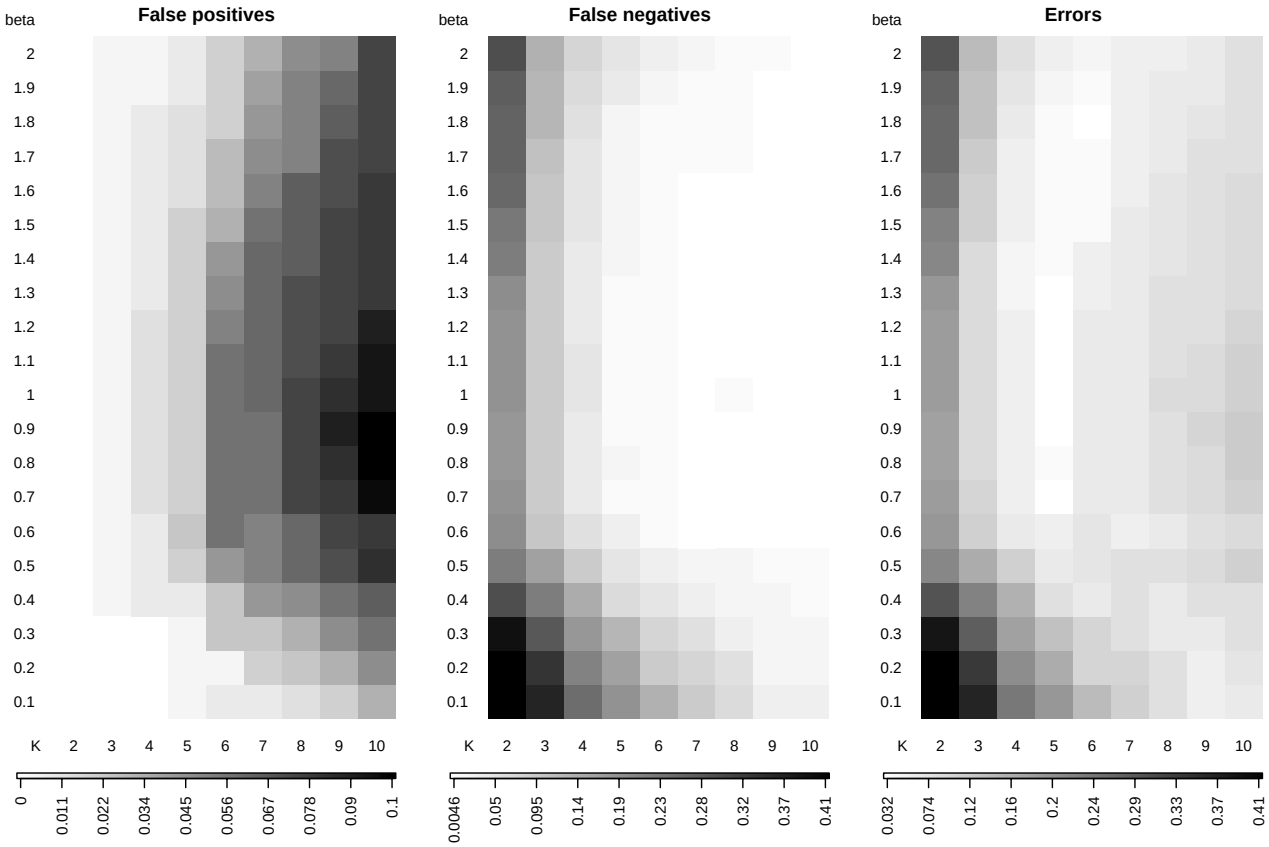


Figure 10
Comparison between qualitative assessment and segmentation results with various (*K*, β) – bladder cancer data set. The segmentation algorithm is run with $K \in \{2,...10\}$ (*x* axis) and $\beta \in \{0.1, 0.2,...2.0\}$ (*y* axis) and compared with the expert assessment of the breast cancer data set. (a) False positive rate; (b) False negative rate; (c) Total error rate.

Figure 10 for bladder cancer data (169/198 arrays). False positives are arrays that experts identified as having no local spatial bias, but which were identified by the algorithm as having local spatial bias. False negatives are arrays that the expert considered to contain local spatial bias, and for which no such areas were reported by the algorithm. Roughly speaking, K controls cluster size, and β influences both the size and spatial coherence of the clusters. As K increases (with fixed β), clusters tend to shrink, leading to an increase in the mean signal value of the highest cluster, making it more likely that this cluster will be identified as a local spatial bias. For fixed K, the highest cluster is slightly more likely to be detected as local spatial bias for intermediate β, corresponding to an extreme cluster with high, homogenous values: for low β this cluster is often quite large and incorporates too small signal values, whereas for very high β, the geographic context is too strong, leading to a highest cluster with heterogeneous signal values.

Drawing figures such as Figure 9 or 10 for any new data set can facilitate the identification of relevant sets of parameters for the segmentation algorithm. In our case, they suggest values of $K = 5$ and β between 0.9 and 1.3 for bladder cancer data set, and $K = 7$ or 8 and β between 0.9 and 1.3 for breast cancer data set. We used $K = 5$, $\beta = 1$ for the bladder cancer data set, and $K = 7$, $\beta = 1$ for the breast cancer data set.

Authors' contributions

PH and EB designed the study. PN and PH designed, coded and validated the spatial normalization algorithm. IB designed and coded the quality criteria. SL performed data integration. PH, PN, IB and EB drafted the manu-

script. EM, CB, FR and AA performed the microarray experiments and validated the spatial normalization algorithm. FR, AA and EB supervised the study. All authors read and approved the final manuscript.

Additional material

Additional File 1

Comparison of method seg+2dLoess with 10 alternative normalization methods. *We compared the method (*seg+2dLoess*) to ten methods for three quality criteria:* sigma, smt *and* dyn. *All images can be described as follows. Each color corresponds to the comparison of* seg+2dLoess *with a different method. The proposed method is taken as a reference (red point 1 at (0, 0)). For each method* i*, the cross indicates the mean relative performance on the data set for the two quality criteria compared, and the lines give the corresponding 95% quantile of the relative performance. The proposed method significantly outperforms, for the quality criterion shown in the y axis (at level 5%), all methods with a 95% quantile below the horizontal dashed black line. Similarly, the proposed method significantly outperformed, for the quality criterion shown in the x axis (at level 5%), all methods with a 95% quantile left of the vertical dashed black line. On most images, methods 2, 3, and 4, which contain a gradient subtraction step using 2*dLoess*, perform the best against* seg+2dLoess*, as they cluster near the top-right corner of the image. However,* seg+2dLoess *still significantly outperforms them for* sigma, smt *and* dyn.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S1.pdf]

Additional File 2

p-*values of the relative performances of 11 normalization methods.* *We compare the results of 11 normalization methods on 3 data sets. Each table gives the significance levels of all pairwise comparisons between these 11 methods, for a given data set and a given quality measurement (*sigma, smt, dyn*). We calculated a relative performance for each array (as explained in the Methods section), and assessed its significance by testing the hypotheses* $\mathcal{H}_{i,j}^{qc}$ *:* $\{RP^{qc}(i, j) < 0\}$ *for each quality criterion* qc*, using a Student's unilateral t-test. The* p*-value associated to* $\mathcal{H}_{i,j}$ *is reported in cell (*i, j*).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S2.pdf]

Additional File 3

Estimates of the relative performances of 11 normalization methods. *We compare the results of 11 normalization methods on 3 data sets. Each table gives the estimates of relative performance of all pairs of methods, for a given data set and a given quality measurement (*sigma, smt, dyn*). We calculated a relative performance for each array, and reported the mean value across all arrays of a given project in the following tables.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-264-S3.pdf]

Acknowledgements

This work was supported by the Institut Curie, the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the IST program from the European Commission through the HKIS project (IST-2001-38153), the Cancéropole Ile de France, and the association *Courir pour la vie, courir pour Curie*.

The construction of the 3.3K BAC-array by Institut Curie, INSERM U509 was supported by grants from the Carte d'Identité des Tumeurs program of the Ligue Nationale Contre le Cancer.

We thank Isabelle Janoueix-Lerosey and Olivier Delattre (Institut Curie, INSERM U509) for making the neuroblastoma data set publicly available.

We thank Nadège Gruel, Virginie Raynal, Gaelle Pierron, Olivier Delattre (Institut Curie, INSERM U509) and Daniel Pinkel (University of California San Francisco) for fruitful discussions.

References

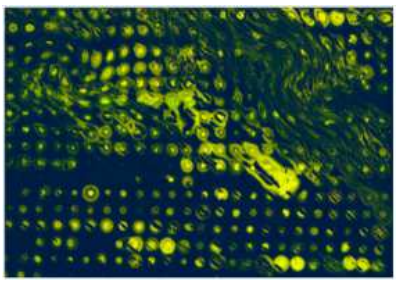
1. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet* 1998, **20:**207-211.
2. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors.** *Nat Genet* 2003, **34:**369-76.
3. Fridlyand J, Snijders A, Pinkel D, Albertson DG, Jain AN: **Application of Hidden Markov Models to the analysis of the array CGH data.** *Journal of Multivariate Analysis* 2004. Special Issue on Multivariate Methods in Genomic Data Analysis
4. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E: **Analysis of array CGH data: from signal ratios to gain and loss of DNA regions.** *Bioinformatics* 2004, **20:**3413-3422.
5. Jong K, Marchiori E, van der Vaart A, Ylstra B, Weiss M, Meijer G: **Chromosomal Breakpoint Detection in Human Cancer.** In *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM, Volume 2611 of LNCS* Edited by: Raidl GR, Cagnoni S, Cardalda JJR, Corne DW, Gottlieb J, Guillot A, Hart E, Johnson CG, Marchiori E, Meyer JA, Middendorf M. University of Essex, England, UK: Springer-Verlag; 2003:54-65.
6. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5:**557-572.
7. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ: **A statistical approach for array CGH data analysis.** *BMC Bioinformatics* 2005, **6:**27.
8. Pollack JR, Sorlie T, Perou CM, Rees A, Jeffreys SS, Lonning P, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a direct role of DNA copy number alteration in the transcriptional program of breast tumors.** *PNAS* 2002.
9. Wang J, Meza-Zepeda LA, Kresse SH, Myklebost O: **M-CGH: Analysing microarray-based CGH experiments.** *BMC Bioinformatics* 2004, **5:**74.
10. Yang YH, Dudoit S, Luu P, Lin DM, Pend V, Ngai J, Speed TP: **Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Research* 2002, **30:**e15:1-e15:11.
11. Khojasteh M, Lam WL, Ward RK, MacAulay C: **A stepwise framework for the normalization of array CGH data.** *BMC Bioinformatics* 2005, **6:**274.
12. Billerey C, Chopin D, Aubriot-Lorton MH, Ricol D, Gil S Diez de Medina, Van Rhijn B, Bralet MP, Lefrere-Belda MA, Lahaye JB, Abbou CC, Bonaventure J, Zafrani ES, van der Kwast T, Thiery JP, Radvanyi F: **Frequent FGFR3 mutations in papillary non-invasive bladder(pTa) tumors.** *Am J Pathol* 2001, **158:**955-1959.
13. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, SL S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet* 2001, **29:**263-4.
14. Janoueix-Lerosey I, Hupé P, Maciorowski Z, La Rosa P, Pierron G, Manié E, Liva S, Barillot E, Delattre O: **Preferential occurence of**

chromosome breakpoints within early replicating regions in neuroblastoma. *Cell Cycle* 2005, **4:**1842-1846.
15. **Replication timing data analysis in Neuroblastoma** [http://microarrays.curie.fr/publications/U509/reptiming]
16. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: **A new non-linear normalization method for reducing variability in DNA microarray experiments.** *Genome Biology* 2002, **3(9):**research0048.1-0048.16.
17. Baird D, Johnstone P, Wilson T: **Normalization of Microarray Data Using a Spatial Mixed Model analysis which includes Splines.** *Bioinformatics* 2004, **20:**3196-3205.
18. Colantuoni C, Henry G, Zeger S, Pevsner J: **Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts.** *Biotechniques* 2002, **32:**1316-1320.
19. Wilson DL, Buckley MJ, Helliwell CA, Wilson IW: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2003, **19:**1325-1332.
20. Tarca AL, Cooke JEK, Mackay J: **A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data.** *Bioinformatics* 2005, **21(11):**2674-2683.
21. Cleveland W, Devlin S, Grosse E: **Regression By Local Fitting.** *Journal of Econometrics* 1988, **37:**87-114.
22. Cleveland WS, Grosse E: **Computational Methods for Local Regression.** *Statistics and Computing* 1991, **1:**47-62.
23. Reimers M, Weinstein JN: **Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases.** *BMC Bioinformatics* 2005, **6:**166.
24. Ambroise C: **Approche probabiliste en classification automatique et contraintes de voisinage.** In *PhD thesis* Université Technique de Compiègne, France; 1996.
25. Ambroise C, Dang M, Govaert G: **Clustering of spatial data by the EM algorithm.** In *Geostatistics for Environmental Applications* Edited by: Soares A, Gomes-Hernandez J, Froidevaux R. Kluwer Academic Publisher; 1997:493-504.
26. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer.** *Nat Genet* 2005:S11-S17.
27. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D: **Fully automatic quantification of microarray image data.** *Genome Res* 2002, **12:**325-332.
28. **MANOR: CGH Micro-Array NORmalization** [http://bioinfo.curie.fr/projects/manor]
29. **Bioconductor: Open software development for computational biology and bioinformatics** [http://www.bioconductor.org]
30. Liva S, Hupé P, Neuvial P, Brito I, Viara E, La Rosa P, Barillot E: **CAPweb : a bioinformatics CGH array Analysis Platform.** *Nucleic Acids Research* 2006 in press.
31. **CAPweb : a bioinformatics CGH array Analysis Platform** [http://bioinfo.curie.fr/CAPweb]
32. Hathaway RJ: **Another interpretation of the EM algorithm for mixture distributions.** *Journal of Statistics and Probability Letters* 1986, **4:**53-56.

CHAPTER 6

Learning cooperative regulation networks

Post-Impressionism



Mary Brinig, *B21n036*, 2003

Vincent van Gogh, *Starry Night*, 1889

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 18 2007, pages 2407–2414
doi:10.1093/bioinformatics/btm352

Systems biology

LICORN: learning cooperative regulation networks from gene expression data

Mohamed Elati[1,2,*,†], Pierre Neuvial[3], Monique Bolotin-Fukuhara[4], Emmanuel Barillot[3], François Radvanyi[2] and Céline Rouveirol[1,†]

[1]LRI, CNRS UMR 8623, bât 490, Université Paris Sud, 91405 F-Orsay, [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 F-Paris, [3]Institut Curie, Service de Bioinformatique, 26 rue d'Ulm, 75248 F-Paris and [4]IGM, CNRS UMR 8621, bât 400/409, Université Paris-Sud, 91405 F-Orsay, France

Received on April 27, 2007; revised on June 27, 2007; accepted on June 29, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: One of the most challenging tasks in the post-genomic era is the reconstruction of transcriptional regulation networks. The goal is to identify, for each gene expressed in a particular cellular context, the regulators affecting its transcription, and the co-ordination of several regulators in specific types of regulation. DNA microarrays can be used to investigate relationships between regulators and their target genes, through simultaneous observations of their RNA levels.

Results: We propose a *data mining* system for inferring transcriptional regulation relationships from RNA expression values. This system is particularly suitable for the detection of cooperative transcriptional regulation. We model regulatory relationships as labelled two-layer gene regulatory networks, and describe a method for the efficient learning of these bipartite networks from discretized expression data sets. We also evaluate the statistical significance of such inferred networks and validate our methods on two public yeast expression data sets.

Availability: http://www.lri.fr/~elati/licorn.html

Contact: mohamed.elati@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Gene regulation in eukaryotes involves many complex mechanisms, most of which are not well understood. With the advent of high-throughput microarray technologies (DeRisi *et al.*, 1997), the expression levels of thousands of genes can be measured simultaneously during various biological processes and for collections of related samples. Considerable effort has been devoted to the analysis of these data sets for the reconstruction of regulatory networks. A family of approaches based on mathematical models of the regulation process has been developed [e.g. Boolean (Liang *et al.*, 1998), Bayesian (Friedman *et al.*, 2000), piecewise-linear (de Jong *et al.*, 2004) and probabilistic Boolean (Bulashevska and Eils, 2005)]. Attempts to learn such models from expression data are hindered by the large number of potential solutions (Chu *et al.*, 2003), and the unrealistically large amount of data required to identify the best solution. In cases of complex formalism for the modelling of regulation, in particular, it has only been possible to reconstruct subnetworks with a few variables. Considerable effort is currently being dedicated to the charting of large-scale gene regulatory networks, relating the expression of a target gene to that of the genes encoding its regulators.

Recent integrative studies have aimed to derive complete yeast gene networks given additional information [e.g. protein–DNA binding from ChIP-chip experiments (Luscombe *et al.*, 2004) or computational analysis of transcription factor binding sites (Middendorf *et al.*, 2004)], with the computational advantage of restricting the number of possible regulators for a given target gene. However, these approaches are difficult to adapt to other organisms, for which the computational detection of *cis*-elements is more difficult, and the experimental detection of binding events is currently limited (e.g. *Homo sapiens*). In contrast, expression data sets are being collected rapidly, and methods based solely on the use of gene expression for network reconstruction are required.

Pe'er *et al.* (2002) have designed the *Minreg* system, a constrained Bayesian network for the reconstruction of large-scale regulatory networks from expression data. The maximal in-degree (i.e. the number of regulators) of target genes and the total number of regulators in the model are limited, so the model focuses on only a small set of global active regulators (AR). The authors made use of these constraints to devise an approximation algorithm for searching for high scoring networks among expression data. The system successfully and robustly identifies the key active regulators, but cannot learn the full detailed network, and may miss interesting regulation relationships: given a current set of active regulators AR, the greedy search of *Minreg* will ignore combinations of co-regulators AR $\cup\{r_1, r_2\}$ if the marginal score values of AR $\cup \{r_1\}$ and AR $\cup \{r_2\}$ are both low, although AR $\cup \{r_1, r_2\}$ may be significant. In such a case, r_1 and r_2 are said to *cooperate* (Nagamine *et al.*, 2005)—i.e. they act

*To whom correspondence should be addressed.
†Present address: LIPN, CNRS UMR 7030 Institut Galilée - Université Paris-Nord F-93430 Villetaneuse, France.

collectively to influence their target genes. Previous computational approaches, due to complexity reasons, have therefore only partly investigated the role of regulator cooperativity. However, such mechanisms have been identified in many organisms (e.g. *Saccharomyces cerevisiae*, *H.sapiens*).

We propose here an original, scalable technique called LICORN (Learning co-operative regulation networks) for deriving cooperative regulations, in which many co-regulators act together to activate or repress a target gene. Many forms of combinatorial logical control may theoretically occur in Boolean or Bayesian models, but we focus here on cooperative regulation patterns that (i) follow the biologically justified activator-repressor model (Woolf and Wang, 2000) (ii) operate on ternary expression level representation (iii) allow for efficient large-scale network computation. LICORN uses an original heuristic approach to accelerate the search for an appropriate structure for the regulation network. It first extracts a global, condensed representation of frequent co-regulator sets using constrained itemset mining techniques (Agrawal *et al.*, 1993). From this representation, a limited subset of candidate co-regulator sets is then efficiently associated with each gene. As this candidate subset is modest in size, exhaustive search for the best gene regulatory network can be performed.

In section 2, we will introduce our model of regulation. Section 3 describes a three-step algorithm for inferring complex combinatorial regulation relationships and a procedure for selecting statistically significant relationships. Finally, in Section 4, we evaluate our system on two yeast data sets.

2 REGULATION MODEL

We represent the regulatory network architecture as a bipartite graph: the top part contains a small number of regulators $\mathcal{R}$ (an estimated 10% of genes in many organisms); the bottom part contains target genes $\mathcal{G}$ (genes, without regulation activity); edges code for a regulatory interaction between regulators and target genes, each edge being labelled with a regulatory mode (i.e. *activator* or *inhibitor*). Like Pe'er *et al.* (2002) and Segal *et al.* (2003), we use a set of candidate regulatory proteins involved in various aspects of gene regulation, including transcription factors, but also signal transduction molecules, to obtain additional information about regulation by considering the levels of expression of signalling molecules with potential indirect effects on transcription.

As in most previous approaches, we chose to convert transcript levels into ternary expression values: -1 (under-expressed), 0 (no change) or 1 (over-expressed). This ternary discretization (see Supplementary Material, Section 1, for more details) is more accurate than a Boolean discretization: it allows for representing both over- and under-expression levels, without making the data representation too complex. Below, the matrix MR stores the expression of regulators in $\mathcal{R}$ and MG the expression of targets in $\mathcal{G}$ for samples from $\mathcal{S}$. For the sake of clarity, we assume that $\mathcal{G}$, $\mathcal{R}$ and $\mathcal{S}$ are arbitrarily ordered and that each target, regulator or sample can be denoted, when it is clear from the context, by its index in $\mathcal{G}$, $\mathcal{R}$ or $\mathcal{S}$.

2.1 Local regulatory program

We model a gene regulatory network (GRN) associated with a target gene g as a pair (A, I), where $A \subseteq \mathcal{R}$ is a co-activator set, and $I \subseteq \mathcal{R}$ is a co-inhibitor set. The cooperative regulators in A (or I), referred to below as *the co-regulator set*, operate collectively as activators or inhibitors of their target gene: for a given sample, they are aggregated in the model through the operator E_AND, which can be interpreted as a logical AND extended to a three-valued logic: $\text{E\_AND}(X) = -1$ if for all $x_i \in X$, $x_i = 1$, $\text{E\_AND}(X) = -1$ if for all $x_i \in X$, $x_i = -1$ and $\text{E\_AND}(X) = 0$ otherwise.

In a simple activator-inhibitor model (Woolf and Wang, 2000), when the level of the activator is high and the level of inhibitor is low, the concentration of the target gene mRNA should be high. Conversely, when the inhibitor concentration is high, and the activator concentration is low, the concentration of the target gene mRNA is low. This qualitative heuristics models expert knowledge concerning regulation control, and was used as the basis for the development of a discrete function called *regulatory program* RP, which, given the combined states of activators A and inhibitors I of g in a sample s computes $\hat{g}_s$ (A, I) the estimated state of g in s as described in Figure 1. The vector of $(\hat{g}_s(A, I))_{s \in S}$ is denoted $\hat{g}$ (A, I).

The main features of our regulation model are therefore the explicit representation of activation and repression relationships for a given target gene, and the representation of cooperative transcriptional regulation.

2.2 Formal problem definition

We can now formally define our inference problem. Given a set of target genes $\mathcal{G}$, a set of regulators $\mathcal{R}$, their discretized expression matrices (MG, MR) over the sample set and an evaluation score h, associating a real number with a candidate GRN, our goal is to find, for each target gene g, the set of regulators that best explains the level of expression

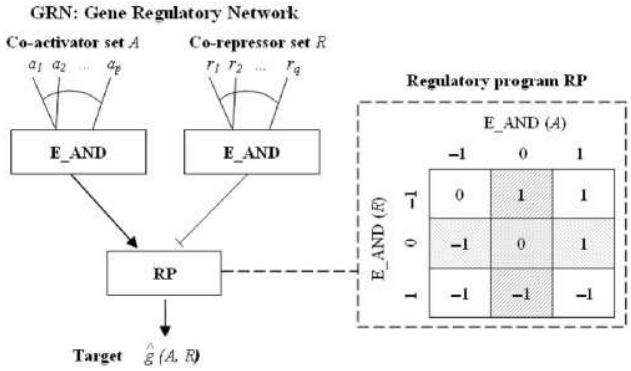


Fig. 1. Definition of the regulatory program RP, which can be interpreted as follows: (i) If GRN contains co-activators only, $\hat{g}$ (A, I) corresponds to the aggregated status of these co-activators. (ii) If GRN contains co-inhibitors only, $\hat{g}$ (A, I) is the inverse of the aggregated status of these co-inhibitors. (iii) Otherwise, $\hat{g}$ (A, I) depends on a combination of the statuses of co-activators and co-inhibitors, as described by the matrix on the right. For example, $\hat{g}$ $(A, I) = 1$ when the co-activators are over-expressed and the co-inhibitors are not.

of g. Finding an optimal GRN—a network minimizing the discrepancy between predicted and observed states for a given gene g—is NP-hard (Pe'er *et al.*, 2002). We will therefore address the problem by adopting a three-step heuristic approach for the detection of cooperative transcriptional regulation.

3 LEARNING ALGORITHM

The first step generates a set of candidate co-regulator sets for all genes of $\mathcal{G}$, such that a candidate co-regulator set is a set of regulators frequently co-expressed in the data. During the second step, for each target gene of $\mathcal{G}$, LICORN efficiently computes a limited set of candidate GRNs and then exhaustively searches for the best one in this set—the activator and inhibitor sets best explaining the target gene status in the sample set. The last step of LICORN is a permutation-based method for the selection of statistically significant GRNs from the inferred GRNs for all target genes.

3.1 Mining global candidate co-regulator sets

3.1.1 Frequent itemset mining The main purpose of *data mining* (Agrawal *et al.*, 1993) is to reveal the relationships between the attributes or *items* of a sparse binary matrix. Sparseness implies very few co-occurrences of items, therefore, most of the counts in the pairwise marginal would be expected to be 0. It is therefore natural to assume that the frequently co-occurring itemsets contain most of the essential information about the data as a whole. A *frequent itemset* is a set of items that appear together in a set of samples (denoted *support*) with a size higher than a user-defined minimum support threshold.

A classical algorithm for mining frequent itemsets is the *Apriori* algorithm (Agrawal *et al.*, 1993). The algorithm relies upon a simple yet fundamental property of the minimum support constraint, namely anti-monotonicity.

DEFINITION 1. (Anti-monotonic property). *A constraint Const is anti-monotonic (with respect to itemset inclusion) if and only if whenever Const is satisfied by an itemset X, Const is also satisfied by all subsets of X.*

Apriori proceeds iteratively, first identifying itemsets of length 1 (1-itemsets). Then, candidate frequent k-itemsets are generated by extending the frequent $(k-1)$-itemsets obtained in the previous iteration. This process is repeated until no more candidate itemsets are found. Considering only candidates obtained by extending existing frequent itemsets allows for an optimized search space exploration. Anti-monotonicity of minimum support guarantees that *Apriori* does not miss any frequent itemset when using this optimized candidate generation.

3.1.2 Candidate co-regulator sets Global candidate co-regulator sets are mined to compute a condensed representation of the discretized expression matrix MR, by looking for all combinations of co-regulators co-occurring frequently in MR. As our input data is three-valued rather than Boolean, each co-regulator set does not have a single support (implicitly a support for value 1 in binary data), but has a support for each value of interest: 1 (denoting over-expression) and −1 (under-expression).

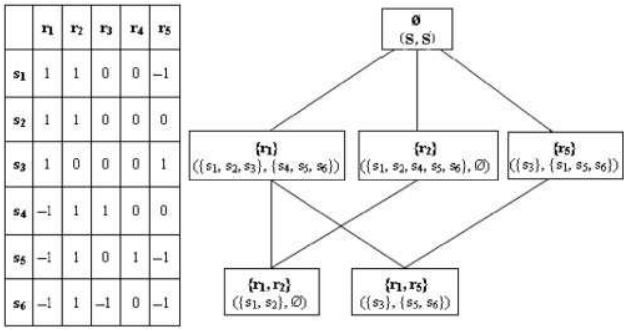


Fig. 2. Given the three-valued expression matrix MR on the left, the right-hand part of the figure shows the sub-lattice of frequent co-regulator sets, with a minimum support of 2 (20% of $|\mathcal{S}|$). Each node of the sub-lattice consists of a co-regulator and its 1- and −1-supports.

DEFINITION 2. (Frequent co-regulator set). *Given the three-valued expression matrix MR, a co-regulator set $C \subseteq \mathcal{R}$ and its 1- and −1-supports, denoted $\mathcal{S}^1(C)$, $\mathcal{S}^{-1}(C) \subseteq \mathcal{S}$ C is frequent if and only if max $(|\mathcal{S}^1(C)|, |\mathcal{S}^{-1}(C)|) \geq T_s$, a user-defined minimum support threshold.*

We have implemented an extension of the *Apriori* algorithm that handles in parallel 1 and −1-supports for building the lattice of frequent itemsets, as shown in Figure 2. At this stage, we opt for a relatively small T_s (20% or less), as the aim is to select candidate co-regulator sets with a low level of stringency, as relevant observed regulations may have medium to low frequency in the data set. This step, the most complex in LICORN, is performed only once in the algorithm.

3.2 Searching for gene regulatory networks

The sub-lattice CL of global frequent co-regulator sets obtained is now used to generate all possible co-regulator sets for each target gene. The criterion for the involvement of a frequent co-regulator set in the regulatory program of a given target gene is hereafter referred to as the *overlap constraint*. Like the co-regulator sets, each gene g has a 1-support $\mathcal{S}^1(g)$ and a −1-support $\mathcal{S}^{-1}(g)$. The overlap constraint (**cov**) checks the size of the intersection between supports of the target gene and a given candidate co-regulator set.

DEFINITION 3. (Overlap constraint). *Given a co-regulator set C, a gene g, and their respective supports $\mathcal{S}^x(C)$ and $\mathcal{S}^y(C)$ for the states $x, y \in \{-1, 1\}$. C in state x co-varies with g in state y, denoted* $\mathbf{cov}(\mathcal{S}^x(C), \mathcal{S}^y(g))$ *if and only if* $\frac{|\mathcal{S}^y(g) \cap \mathcal{S}^x(C)|}{|\mathcal{S}^y(g)|} \geq T_o$, *a user-defined minimum overlap threshold.*

T_o is the lower limit of the proportion of samples in which the target g is over- or under-expressed while the co-regulator set C is over- or under-expressed. In other words, it is the conditional probability $\mathbb{P}$ (E_AND $(C) = x \mid g = y$), with $x, y \in \{-1, 1\}$. Note that T_o should exceed 50%, as a small overlap size makes the definition of the regulatory program meaningless. We distinguish co-regulator sets satisfying **cov** for a given target gene according to their roles: a candidate

co-activator set A for g, is a co-regulator set that positively co-varies with g, and a candidate *co-inhibitor set* I negatively co-varies with g. $\mathcal{A}(g)$ and $\mathcal{I}(g)$ denote, respectively, all candidate co-activator and co-inhibitor sets for g.

$$\mathcal{A}(g) = \left\{A \in \mathrm{CL} | \mathbf{cov}(\mathcal{S}^x(A), \mathcal{S}^x(g)); x \in \{-1, 1\}\right\}$$
$$\mathcal{I}(g) = \left\{I \in \mathrm{CL} | \mathbf{cov}(\mathcal{S}^x(I), \mathcal{S}^{-x}(g)); x \in \{-1, 1\}\right\}$$

CL may be too large and it may therefore be too expensive to generate candidate co-regulator sets of each target gene blindly. As CL is a sublattice partially ordered by $\subseteq$ and given that **cov** is anti-monotonic (see Definition 1) with respect to $\subseteq$, efficient pruning during search is possible: when a coregulator set C does not satisfy **cov**, no superset of C can ever satisfy **cov**. Therefore, large parts of the sublattice need not to be explored.

We can thus compute the set of all candidate GRNs for each target gene g as follows:

$$\mathcal{C}(g) = \{(A,\ I) | A \in \mathcal{A}(g), I \in \mathcal{I}(g) \text{ and } A \cap I = \emptyset\}$$

A candidate GRN for g, or a GRN for short, is an element of $\mathcal{C}(g)$.

3.3 Scoring gene regulatory networks

In the preceding steps, we have built, for each gene g, a relatively small number of candidate regulatory networks, based on the recurrent positive and negative co-variation of candidate co-regulator sets with g. We now define a scoring function to compare the different GRNs inferred for a given gene, and to choose the best one. We propose a resampling approach for estimating the statistical significance of each *best candidate* GRN for each target, and a method for determining which candidates are significant enough to be retained.

3.3.1 Best GRN for each gene We propose a heuristic measurement for comparing discretized expression profiles, in which each candidate GRN associated with a given gene is scored. As discretized expression values are *ordinal variables*, mean absolute error (MAE) is used to measure distance between gene expression profiles: ideally, over-expressed genes should be closer to genes with no change in expression than to under-expressed genes.

$$h_g(A,\ I) = \mathrm{MAE}(g, \hat{g}(A,\ I)) = \sum_{s \in \mathcal{S}} |g_s - \hat{g}_s(A,\ I)|$$

where $g_s = \mathrm{MG}_{sg}$. Note that $0 \leq \mathrm{MAE} \leq 2$. The best candidate GRN for gene g is then defined as

$$\mathrm{GRN}^*(g) = \underset{(A,I) \in \mathcal{C}(g)}{\mathrm{Argmin}}\ \ h_g(A,\ I)$$

3.3.2 Significance estimation Our scoring function h allows us to define a best GRN for each gene, but the scores of the best GRN associated with two different genes may not be directly comparable, as different genes have different probabilities of being under- or over-expressed in the study. Moreover, a GRN is selected because the expression of activator and inhibitor sets co-varies in a recurrent fashion with expression of the gene of interest. Most distances are therefore necessarily small, and a small distance for a given gene does not guarantee that the best GRN is statistically significant. We use statistical hypothesis testing to evaluate how unusually low the score of the best GRN is with respect to the scores that would have been observed if there was no biological relationship between regulators and target gene expression.

The absence of a biological relationship between the target and candidate regulators in the GRN is checked, using random permutations of the samples in the gene expression matrix MG. $B = 1000$ randomized matrices $\mathrm{MG}^{(b)}$ are generated, each corresponding to a particular permutation of the samples. For each permutation b, we infer for each gene g a set $\mathcal{C}_b(g)$ of candidate GRNs, and select the best candidate GRN_b^* from this set, as described above. The statistical significance (P-value) of gene g is estimated as the proportion of permutations for which the best score is lower than that obtained with real data:

$$P(g) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{h_g(\mathrm{GRN}_b^*) \leq h_g(\mathrm{GRN}^*)\}}$$

3.3.3 Correction for multiple hypothesis testing Selecting the genes for which the best candidate GRN is significant based on these P-values consists of a multiple hypothesis testing problem, which can be addressed using the false discovery rate paradigm (FDR) introduced by Benjamini and Hochberg (1995). The idea is to control the expected fraction of false positives (i.e. the FDR) among those GRNs selected. We used the FDR control procedure proposed by Benjamini and Yekutieli (2001), which provides strong FDR control for any kind of dependence between test statistics.

4 RESULTS AND DISCUSSION

As a proof of concept, we used LICORN for the mining of gene regulatory networks separately on two different gene expression data sets for *S.cerevisiae*. The Gasch data set (Gasch *et al.*, 2000) measures the response of yeast to 173 stress conditions for 6152 genes. The Spellman data set (Spellman *et al.*, 1998) consists of a series of 73 microarray experiments measuring gene expression during the cell cycle for 6178 genes. These two expression matrices were discretized into three states -1, 0 and 1: for the Gasch data set, discretized values reflect the expression levels of each gene in each experimental condition; for the Spellman data set, discretized values reflect *expression changes* between consecutive time points. Discretization thresholds, as described in the Supplementary Material (Section 1), were chosen so as to yield balanced frequencies of 1, -1 and 0 in the data set. No gene selection was performed at this step: the discretized matrices still contain 6152 and 6178 genes, respectively.

We used a set of 475 regulators compiled by Middendorf *et al.* (2004), consisting of 237 known and putative transcription factors and 250 known and putative signalling molecules, with an overlap of 12 genes of unknown function. A large amount of biological knowledge on yeast is available: function information, contained in the *Saccharomyces Genome Database (SGD)* (Cherry *et al.*, 1998), documented regulations in the *YEASTRACT* database (Teixeira *et al.*, 2006), protein–protein interactions in the *BioGRID* database

(Stark *et al.*, 2006) and data about DNA-binding to transcriptional regulators in ChIP-chip experiments (Harbison *et al.*, 2004; Lee *et al.*, 2002). Thus, the transcriptional networks identified for these two data sets can be checked by comparison with various sources of information.

4.1 Performance evaluation

4.1.1 Objective measurement of prediction performance Above, we used MAE as a measure of the discrepancy between actual gene expression and the gene expression inferred from the activity of regulators through the GRN. The prediction error of a particular gene *on a given sample set* $\mathcal{T}$ is defined as the MAE within $\mathcal{T}$:

$$e(g) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |g_t - \hat{g}_t|$$

Averaging individual prediction errors across all selected genes leads to the following *global measure of prediction error* of the model:

$$e = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} e(g)$$

For a prediction measure to be objective, it must be evaluated on a validation set that has not been used to build the predictor. Cross-validation involves partitioning the observed population $\mathcal{S}$ into K subgroups $\mathcal{S}_1, \ldots, \mathcal{S}_K$. For $k = 1, \ldots, K$, the predictor is built on the *training population* $\mathcal{S} \setminus \mathcal{S}_k$, and its performance is evaluated on the *test population* $\mathcal{S}_k$. In practice, 10-*fold cross-validation* ($K = 10$) is often considered, as this method provides a fair estimate of the prediction error at a reasonable computational cost (10 training runs with $\frac{|\mathcal{S}|}{10}$ observations each).

4.1.2 Results Using 10-fold cross-validation, we compared four methods: (i) a majority vote in which the predicted gene expression value in the test set is simply the most frequent expression value for this gene in the training set; (ii) a re-implementation of the *Minreg* system, as previously described (Pe'er *et al.*, 2002). We limited running time by filtering out the least informative genes—those remaining almost unchanged in more than 65% of samples—and we have set the maximal in-degree of target genes in the networks to 2 (iii) LICORN algorithm without selection of significant GRNs and (iv) LICORN algorithm with selection of significant GRNs at the 0.05 FDR level.

We used the same 10 cross-validation subgroups to evaluate each of the methods, to facilitate comparisons of performance. The significance of the difference between the prediction rates of two methods on these subgroups was assessed using a paired *t*-test. Cross-validation results are given in Figure 3 for the Gasch data set. Similar results were obtained for the Spellman data set (Supplementary Material, Section 2). It should be noted that the ranking of the methods was the same for all folds, for both data sets. LICORN significantly outperformed *Minreg*, with a *P*-value in paired *t*-tests of 1.6×10^{-8} for the Spellman data set, and 6.7×10^{-9} for the Gasch data set. Focusing only on those GRNs selected at a given FDR threshold resulted in significant further decrease in MAE: LICORN with FDR < 5% outperformed LICORN, with $P = 1.3 \times 10^{-10}$ for the Spellman data set, and 5.2×10^{-13} for the Gasch data set.

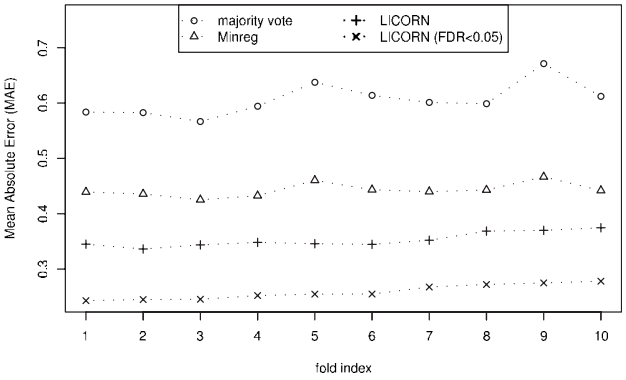


Fig. 3. Results of the 10-fold cross-validation on the Gasch data set: comparison of the MAE for all GRNs for the test set for each fold. We recall that $0 \leq \text{MAE} \leq 2$. Folds were sorted in increasing order of MAE for the method 'LICORN (FDR<0.05)'.

4.2 Biological analysis

We applied LICORN as described in the Section 3, and retained only those GRNs (gene regulatory networks) identified as significant with a 5% FDR level by the Benjamini and Yekutieli (2001) procedure. We chose the 5% level empirically: it is stringent enough to guarantee that the overwhelming majority of selected GRNs are true discoveries, but relaxed enough for almost half the genes to be retained: for the Gasch data set, 2795 GRNs (of 5703 GRNs) were identified as significant, whereas for the Spellman data set, 2792 GRNs (of 5677 GRNs) were identified as significant. We show some examples of learned GRNs in the Gasch and Spellman data sets in the Supplementary Material, Section 3. We discuss below the structural organization of the learned GRNs. We then provide two kinds of biological evidence to support the inferred GRNs: (i) documented regulation and high-throughput ChIP-chip data sets for confirming transcription factor-target interactions; (ii) protein–protein interactions and functional evaluation for confirming co-regulator cooperativity.

4.2.1 Overall network structure Analysis of the structure and organisation of the inferred networks revealed several notable features. In both stress response GRNs (Gasch data set) and cell cycle GRNs (Spellman data set), we found about 10 000 interactions between regulators and target genes. On average, each target is regulated by three regulators in both data sets. Regulators in stress response conditions have a greater influence than cell cycle regulators, as they target more genes simultaneously (on average 30 targets versus 23 targets). We have shown that the distribution of the outgoing connectivity is best approximated by a power-law equation (Supplementary Material, Section 4.2). This allowed us to detect regulator hubs (Lee *et al.*, 2002) with high out-going connectivity (e.g. the heat shock and osmolarity stress regulator *PPT1* regulates 300 target genes). For most regulators in both data sets, a linear dependence was observed between

the number of target genes regulated by a given regulator and the number of co-regulators of that regulator (Supplementary Material, Section 4.3). However, regulators from the Spellman data set have a higher number of co-regulators than do regulators from the Gasch data set (on average 16 co-regulators versus 12 co-regulators), indicating that there are more cooperative associations between regulators in cell cycle GRNs.

All these results, fully detailed in the Supplementary Material (Section 4), are consistent with recent advances (Balaji *et al*., 2006; Guelzim *et al*., 2002; Luscombe *et al*., 2004) concerning the characterization of topological transcriptional network features in yeast and provide the first evidence of the relevance of inferred GRNs.

4.2.2 Evaluating transcription factor-target interactions Firstly, as expected, we found that the transcription factors frequently occurring in GRNs inferred from the Gasch data set (e.g. MSN4, XBP1, YAP1, CAD1) played a major role in the response to stress and that many frequent transcription factors in the Spellman data set (e.g. MBP1, FKH1, XBP1, SWI4, ACE2) were involved in the cell cycle. In addition, the *SGD* annotations, concerning the role (activator/inhibitor) of transcription factors, when available, corresponded to the role most frequently assigned within the GRNs inferred, for both data sets (Supplementary Material, Section 5.1). We also showed that LICORN-inferred TF-target interactions have a significant overlap with condition-specific TF-target interactions obtained by Luscombe *et al*. (2004) with their recent integrative method when applied on the same data sets (Supplementary Material, Section 5.2).

The chromatin immunoprecipitation (ChIP) method profiles the binding sites for each transcription factor throughout the entire genome. We compared our results for the Gasch and the Spell-man data sets respectively with those for a stress-response (Harbison *et al*., 2004) and a cell cycle (Lee *et al*., 2002) ChIP-chip data sets. For each condition, we then checked the overlap of both sets of prediction with more than 12 000 demonstrated TF-target relationships described in diverse studies, organized in the *YEASTRACT* knowledge base (Teixeira *et al*., 2006).

In Figure 4, we show the relative overlap between the three sets of identified interactions. For the Gasch data set, 47% of the relationships predicted by LICORN were confirmed by *YEASTRACT*, and 29% of these relationships were also confirmed by the ChIP-chip predictions. Overall, 25% of LICORN predictions were confirmed by ChIP-chip predictions, and 17% of ChIP-chip predictions were confirmed by LICORN. Similar proportions were obtained for predictions based on the Spellman data set: 50% of the relationships predicted by LICORN were confirmed by *YEASTRACT*, and 32% of these relationships were also confirmed by the ChIP-chip predictions. Overall, 26% of the LICORN predictions were confirmed by ChIP-chip predictions and 20% of the ChIP-chip predictions were confirmed by LICORN. The agreement between LICORN results and regulation documented in YEASTRACT is consistent with the agreement between ChIP-chip predictions and this database. For both data sets, ~40% of the regulations learned by LICORN were not supported

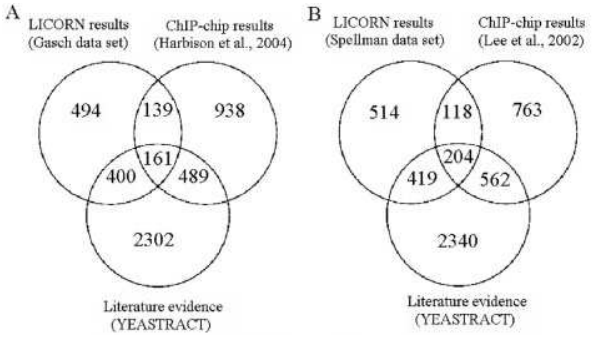


Fig. 4. LICORN interaction predictions and ChIP-chip interaction results (with *P*-values <0.001), compared with experimental evidence concerning regulation collected from *YEASTRACT*: (**A**) The number of TF-target interactions, for the 82 TFs shared by LICORN-inferred GRNs from the Gasch data set, the stress-response ChIP data set (Harbison *et al*., 2004) and *YEASTRACT*. (**B**) Number of TF-target interactions for the 69 transcription factors shared by LICORN-inferred GRNs from the Spellman data set, normal growth ChIP data set (Lee *et al*., 2002) and *YEASTRACT*.

by YEASTRACT or by ChIP-chip experiments. There are several possible explanations for this: (i) the usual noise in expression data and the 5% FDR yield a number of false discoveries (ii) large portions of the underlying true network remain unknown and some of these interactions, currently unsupported experimentally, may enable researchers to propose new hypotheses potentially corresponding to new regulation relations.

Evaluation of the non-documented candidate genes under control of a specific TF in GRNs learned from the Gasch data set revealed some biological connections. These connections were obvious for the GAT1 TF which is know to be a transcriptional activator of genes involved in nitrogen catabolite repression (Coffman *et al*., 1995) and associated in our results to several non-documented genes among which DAL3 and YLR164W. Both genes are associated directly or indirectly with nitrogen utilization (Scherens *et al*., 2006; Yoo and Cooper, 1991). More interestingly, YAP1, a transcription factor required for oxidative stress tolerance (Schnell *et al*., 1992) was found to be associated with the EAF3 and TPP1 genes. Both these genes have functions classified as *DNA repair biological process*. EAF3, a chromatin acetylase component (Eisen *et al*., 2000), is probably involved in transcription-coupled repair, a DNA repair mechanism associated with chromatin modifications (Teng *et al*., 2005). TPP1 repairs endogenous damage to double-stranded DNA (Vance and Wilson, 2001). As oxidative stress is known to induce damages in proteins, lipids and DNA, it seems logical that YAP1, in addition to controlling genes necessary to cope with oxygen reactive species, also induces the transcription of genes involved in DNA repair. Finally, this method can be used to identify less direct connections that are nonetheless biologically sound. An example is provided by the BAS1 TF, which is involved in regulating the basal and induced expression of genes of the purine and histidine biosynthesis pathways (Daignan-Fornier and Fink, 1992). Among the non-documented genes predicted to be controlled by BAS1 we found, DPH5, encoding a

Table 1. List of the 20 most frequent co-regulator pairs involved in the GRNs learned from the Gasch data set

Co-regulators	NT	BG	Shared *GO-Slim* terms
LSG1 PPT1	93	N*	Ribosome biogenesis and assembly protein biosynthesis
TPK1 TPK2	74	Y	Response to stress, RNA metabolism
RAP1 PPT1	72	N	RNA metabolism
PDE1 GLC8	49	N	Protein biosynthesis organelle organization and biogenesis
LSG1 YVH1	44	Y	Organelle organization and biogenesis ribosome biogenesis and assembly
MSN4 TPK1	42	Y	Response to stress
XBP1 TOS8	39	N*	Translation
GIS1 TPK1	35	Y*	Response to stress
PHO2 BAS1	32	Y	RNA metabolism
MSN4 USV1	30	N*	Cell wall organization and biogenesis
BCY1 TPK1	29	Y	Morphogenesis, response to stress
PPT1 YVH1	29	N	Ribosome biogenesis and assembly
BMH2 TPK1	27	Y	Protein catabolism, response to stress
CLB6 CLB5	23	Y	Cell cycle, DNA metabolism
MSN4 TPK2	23	Y	RNA metabolism
GCN20 GCN1	22	Y	Sporulation
XBP1 USV1	20	N*	Organelle organization and biogenesis
FAR1 CLN2	19	Y	Cell wall organization and biogenesis protein modification
PPT1 RAS1	19	N	Response to stress
BMH2 BMH1	19	Y	Carbohydrate metabolism

For each co-regulator pair, the number of targets (NT) and the existence or otherwise of known protein–protein interactions in the *BioGRID* database (BG) are indicated, together with the list of GO-Slim terms significantly shared by more than 40% of their target genes. (*) indicates that the co-regulators found were identified in the results of Segal *et al.* (2003).

methyltransferase required for the synthesis of a modified histidine residue (Mattheakis *et al.*, 1992) and TRM1, encoding an N2,N2-dimethylguanosine-specific tRNA methyltransferase (Ellis *et al.*, 1986). Although not directly involved in the purine or histidine biosynthetic pathways, their functions depend on these pathways. It is therefore reasonable to assume that they may be co-regulated with the well-identified BAS1 target genes, consistent with the coupling of main pathways with secondary ones. These interesting new possibilities require experimental testing.

4.2.3 Evaluating candidate co-regulators We evaluated cooperativity between the co-regulators inferred by LICORN, based on two assumptions: (i) the existence of protein–protein interactions between co-regulators implies participation in the same regulatory mechanism, and (ii) targets contributing to similar biological process are regulated by the same control mechanism. In Table 1, we have listed the 20 most frequent regulator pairs in co-activator or co-inhibitor sets in GRNs learned from the Gasch data set. For each co-regulator pair, we have checked whether the two co-regulators are known to interact (protein or genetic interaction), based on information in the *BioGRID* database. In total, 60% of the co-regulator pairs in the list were reported to interact in *BioGRID*. This proportion is high and confirms the validity of LICORN predictions, with *P*-value close to 0 ($<10^{-15}$). For all co-regulator pairs, we found GO-Slim terms (high-level GO terms that represent the major biological processes in *S.cerevisiae*) significantly shared by at least 40% of the target genes, using the *GO-Slim* mapping tool of the *SGD* (Cherry *et al.*, 1998), demonstrating the functional robustness of the co-regulators inferred by LICORN.

The pairs of co-regulators in Table 1 include the known heat shock and osmolarity stress regulators TPK1, PPT1 and USV1, which occur at high frequency. This observation correlates well with the results obtained by Segal *et al.* (2003) and Middendorf *et al.* (2004) for the Gasch data set. Segal *et al.* (2003) identified these proteins as the master regulators for this data set, as they occurred in more than 5 of the 50 inferred modules of co-regulated genes and their regulators. Four of the eight co-regulator pairs not found in *BioGRID* were identified by Segal *et al.* (2003). Moreover, Segal *et al.* (2003) did not identify some of our confirmed co-regulators (e.g. TPK1-TPK2, GCN20-GCN1 and CLB6-CLB5), as in cases in which several regulators are involved in the same regulatory event, this method typically identifies only one representative of the group.

Finally, we obtained similar results for the list of the 20 most frequent co-regulator pairs involved in the GRNs learned from the Spellman data set (see Supplementary Material, Section 6.1). We also found significant agreement between the extent of cooperative associations between regulators and physical interactions between regulatory proteins during the yeast cell cycle, as reported by de Lichtenberg *et al.* (2005). More details are given in the Supplementary Material (Section 6.2). These results confirm those of recent studies (Balaji *et al.*, 2006; Nagamine *et al.*, 2005) connecting regulator cooperativity and protein–protein interactions.

5 CONCLUSION

We provide here a model for cooperative regulation and an algorithm, LICORN, for the inference of cooperative regulation from gene expression data. We used a permutation-based procedure selecting the most statistically significant regulation networks and have shown that this selection step improves prediction performance in a 10-fold cross-validation framework. Moreover, validation on two yeast data sets showed that LICORN was a powerful *data mining* tool for the analysis of gene expression. The results obtained with this algorithm were consistent with published experimental results. The labelled relationships (activation/inhibition) found with our method do not require post-treatment analysis for interpretation, unlike the combinatorial interactions learned with Bayesian network algorithms (Friedman *et al.*, 2000; Pe'er *et al.*, 2002).

Cooperative regulation patterns cannot be identified by clustering or pairwise methods (Woolf and Wang, 2000), and are only partly revealed by constrained Bayesian or decision tree-based techniques, such as those used in previous studies (Middendorf *et al.*, 2004; Pe'er *et al.*, 2002; Segal *et al.*, 2003). Rather than selecting regulators independently, LICORN efficiently reduces the search space for the candidate regulators of the targets to the sub lattice of frequent co-regulators. This decreases the number of regulator combinations to be evaluated, and LICORN does not require strong a priori selection criteria based on uncertain or incomplete information, such as

DNA-binding data (Middendorf *et al*., 2004). LICORN thus speeds up the inference of gene regulatory networks including co-activator and co-inhibitor sets. LICORN also avoids the use of 'gene modules' (Segal *et al*., 2003) for factorizing the search for the best regulation network. Modularity may be an organizing principle of regulatory networks, but it may be too coarse for the learning of specific regulatory programs (LICORN learns a regulation network for each gene). Instead, partial overlap of the regulator sets for a set of target genes, once inferred, can be used as an alternative measurement of the distance between genes.

Future work should focus on extending the LICORN model, to increase accuracy and generalization. For instance, LICORN can be extended to the learning of other classes of combinatorial regulation, in which several co-activator or co-inhibitor sets may function independently, or in which regulatory relationships may link the regulators themselves. This requires care, to avoid problems of over-fitting, given the small size of the training sets available. Finally, the gene regulatory networks learned by LICORN from expression data can be enriched by integrating various gene networks from diverse data sources (motif networks, ChIP-chip data, protein–protein interactions, functional category, etc.). This suggests the use of a logical representation for gene networks, and the use of adapted integrative algorithms, such as those developed in *Inductive Logic Programming* (Fröhler and Kramer, 2006).

ACKNOWLEDGEMENTS

We thank Ch. Froidevaux for her constant support, Ch. Battail for fruitful discussions and the anonymous referees for their pertinent suggestions. We also thank J. Sappa from Alex Edelman and Associates for careful reading of the manuscript. This work was supported by the CNRS, the Institut Curie, the Plan Pluri-Formation Bioinformatique et Génomique and the IFR Génome. M. Elati and F. Radvanyi are members of the Equipe Oncologie Moléculaire, labellisée par La Ligue Nationale Contre le Cancer. M.E. was supported by a fellowship from the French Ministry of Foreign Affairs. P.N. was supported by a fellowship from the association Courir pour la vie, courir pour Curie.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. *et al*. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the International Conference on Management of Data,* pp.207–216.

Balaji,S. *et al*. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol*., **360**, 204–212.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc*., **57**, 289–300.

Benjamini,Y. and Yekutieli,Y. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1198.

Bulashevska,S. and Eils,R. (2005) Inferring genetic regulatory logic from expression data. *Bioinformatics*, **21**, 2706–2713.

Cherry,J. *et al*. (1998) SGD: Saccharomyces Genome Database. *Nucleic. Acids Res*., **26**, 73–79.

Chu,T. *et al*. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147–1152.

Coffman,J. *et al*. (1995) Genetic evidence for Gln3p-independent, nitrogen catabolite repression-sensitive gene expression in Saccharomyces cerevisiae. *J. Bacteriol*., **177**, 6910–6918.

Daignan-Fornier,B. and Fink,G. (1992) Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proc. Natl Acad. Sci*., **89**, 6746–6750.

de Jong,H. *et al*. (2004) Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull. Math. Biol*., **66**, 301–340.

de Lichtenberg,U. *et al*. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.

DeRisi,J. *et al*. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,A. *et al*. (2000) The yeast NuA4 and Drosophila MSL complexes contain homologous subunits important for transcriptional regulation. *J. Biol. Chem*., **276**, 3484–3491.

Ellis,S. *et al*. (1986) Isolation and characterization of the TRM1 locus, a gene essential for the N2,N2-dimethylguanosine modification of both mitochondrial and cytoplasmic tRNA in Saccharomyces cerevisiae. *J. Biol. Chem*., **261**, 9703–9709.

Friedman,N. *et al*.Using bayesian network to analyze expression data. *Comput. Biol*., **7**, 601–620.

Fröhler,S and Kramer,S. (2006) Logic-based information integration and machine learning for gene regulation prediction. In *Proceedings of the 9th International Conference on Molecular Systems Biology*.

Gasch,A. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Guelzim,N. *et al*. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet*., **31**, 60–63.

Harbison,C. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Liang,S. *et al*. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium in Biocomputing*, 18–29.

Luscombe,N.M. *et al*. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Mattheakis,L. *et al*. (1992) DPH5, a methyltransferase gene required for diphthamide biosynthesis in Saccharomyces cerevisiae. *Mol. Cell. Biol*., **12**, 4026–4037.

Middendorf,M. *et al*. (2004) Predicting genetic regulatory response using classification. *Bioinformatics*, **20**, 232–240.

Nagamine,N. *et al*. (2005) Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic. Acids Res*., **33**, 4828–4837.

Pe'er,D. *et al*. (2002) Minreg: inferring an active regulator set. *Bioinformatics*, **18**, 258–267.

Scherens,B. *et al*. (2006) Identification of direct and indirect targets of the gln3 and gat1 activators by transcriptional profiling in response to nitrogen availability in the short and long term. *FEMS Yeast Res*., **6**, 777–791.

Schnell,N. *et al*. (1992) The par1 (yap1/snq3) gene of saccharomyces cerevisiae, a c-jun homologue, is involved in oxygen metabolism. *Curr. Genet*., **21**, 269–273.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.

Spellman,P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic. Acids Res*., **34**, 535–539.

Teixeira,M. C. *et al*. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic. Acids Res*., **34**, 446–451.

Teng,Y. *et al*. (2005) Histone acetylation, chromatin remodelling, transcription and nucleotide excision repair in s. cerevisiae: studies with two model genes. *DNA Repair*, **4**, 870–883.

Vance,J. and Wilson,T. (2001) Uncoupling of 3' phosphatase and 5' kinase functions in budding yeast: characterization of S. cerevisiae DNA 3' phosphatase (TPP1). *J. Biol. Chem*., **276**, 15073–15081.

Woolf,P. and Wang,Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics*, **3**, 9–15.

Yoo,H. and Cooper,T. (1991) The ureidoglycollate hydrolase (dal3) gene in saccharomyces cerevisiae. *Yeast*, **7**, 693–698.

CHAPTER 7

Defining true recurrences among ipsilateral breast cancers

Surrealism

Mary Brinig, *B14n101*, 2002

Salvador Dalí, *The Persistence of Memory*, 1931

ARTICLE

High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers

Marc A. Bollet, Nicolas Servant, Pierre Neuvial, Charles Decraene, Ingrid Lebigot, Jean-Philippe Meyniel, Yann De Rycke, Alexia Savignoni, Guillem Rigaill, Philippe Hupé, Alain Fourquet, Brigitte Sigal-Zafrani, Emmanuel Barillot, Jean-Paul Thiery

Background To distinguish new primary breast cancers from true recurrences, pangenomic analyses of DNA copy number alterations (CNAs) using single-nucleotide polymorphism arrays have proven useful.

Methods The pangenomic profiles of 22 pairs of primary breast carcinoma (ductal or lobular) and ipsilateral breast cancers from the same patients were analyzed. Hierarchical clustering was performed using CNAs and DNA breakpoint information. A partial identity score developed using DNA breakpoint information was used to quantify partial identities between two tumors. The nature of ipsilateral breast cancers (true recurrence vs new primary tumor) as defined using the clustering methods and the partial identity score was compared with that based on clinical characteristics. Metastasis-free survival was compared among patients with primary tumors and true recurrences as defined using the partial identity score and by clinical characteristics. All statistical tests were two-sided.

Results All methods agreed on the nature of ipsilateral breast cancers for 14 pairs of samples. For five pairs, the clinical definition disagreed with both clustering methods. For three pairs, the two clustering methods were discordant and the one using DNA breakpoints agreed with the clinical definition. The partial identity score confirmed the nature of ipsilateral breast cancers as defined by clustering of DNA breakpoints in 21 of 22 pairs. The difference in metastasis-free survival of patients with new primary tumors and those with true recurrences was not statistically significant when tumors were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% confidence interval [CI] = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; new primary tumor vs true recurrence, hazard ratio = 2.8, 95% CI = 0.6 to 13.7), but the difference was statistically significant when tumors were defined using the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Conclusions DNA breakpoint information more often agreed with the clinical determination than CNAs in this population. The partial identity score, which was calculated based on DNA breakpoints, allows statistical discrimination between new primary tumors and true recurrences that could outperform the clinical determination in terms of prognosis.

J Natl Cancer Inst 2008;100:48–58

Breast-conserving therapy is the preferred treatment for patients with early-stage breast cancer (1). It offers equal local control and overall survival (2) and superior psychosocial outcomes (3,4) compared with modified radical mastectomy. However, an ipsilateral breast cancer recurrence can be traumatizing and can lead to death (2).

When an ipsilateral breast cancer develops, the new tumor can either be a true recurrence—that is, a regrowth of clonogenic cells that were not removed by surgery or killed by radiotherapy—or a new primary tumor that arises from the remaining breast tissue (5). Several definitions have been used to distinguish true recurrences from new primary tumors. Initially, these distinctions were based

Affiliations of authors: Département d'oncologie radiothérapie (MAB, AF), Service de Bio-informatique (NS, PN, GR, PH, EB), Département de Transfert (CD, JPM, JPT), Département de Biologie des tumeurs (IL, BSZ), Service de Biostatistiques (YDR, AS), and Centre National de la Recherche Scientifique, Unité Mixtes de Recherche 144 (CD, PH), Institut Curie, Paris, France; Institute of Molecular and Cell Biology Biopolis A*STAR, Singapore (JPT).

Correspondence to: Marc A. Bollet, MD, Département d'oncologie radiothérapie, Institut Curie, 26, rue d'Ulm, 75248 Paris cedex 05, France (e-mail: marc.bollet@curie.net).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djm266

on the location of the ipsilateral breast cancer (ie, the farther from the initial primary tumor, the more likely it is to be a new primary tumor) and on shared common histopathologic criteria (eg, type, grade, and hormone receptor status) (6–10). In the quest for additional ways to distinguish new primary breast tumors from true breast cancer recurrences, biologic studies of clonal relationships between the new and original tumor have also been performed. These studies have relied on ploidy (5,11), loss of heterozygosity (12–14), p53 analysis (15), or X chromosome inactivation (16) or have been based on DNA copy number alterations (CNAs) (17–19). CNA data can be obtained by high-resolution techniques, such as array-based comparative genomic hybridization or single-nucleotide polymorphism (SNP) arrays (20). One of the most commonly used ways to look at clonal relatedness using pangenomic data is to perform an unsupervised hierarchical clustering that organizes primary breast tumors and ipsilateral breast cancers on the basis of their overall genomic similarity (18,19). These measures of similarity are summarized in a dendrogram, in which the pattern and length of the branches reflect the relatedness of the samples in terms of DNA CNAs.

Changes in DNA copy numbers occur at chromosomal locations called breakpoints. We hypothesized that the precise locations of these breakpoints could serve as markers for clonal relatedness and that we could distinguish true recurrences from new primary tumors by the number of common breakpoints in the ipsilateral breast cancer and the primary tumor. In this study, we first aimed to test the added value of examining the clustering of breakpoints (over CNAs) when determining the nature of the ipsilateral breast cancer. Second, we aimed to develop a score to quantify the partial identity between two tumors according to their clonal relatedness (determination of the partial identity score). Third, we examined prognosis in terms of metastasis-free survival. In each case, these methods were compared with the clinical determination of the nature of the ipsilateral breast cancer.

CONTEXT AND CAVEATS

Prior knowledge

Detecting changes in DNA copy number using single nucleotide polymorphism arrays has been a useful tool in distinguishing new primary breast tumors from recurrences.

Study design

Comparison of hierarchical clustering of DNA copy number and DNA breakpoints, an identity score based on the DNA breakpoint information, and clinical characteristics to accurately designate ipsilateral breast tumors as new primary tumors or true recurrences in breast tumor pairs from 22 patients.

Contributions

For 14 of the pairs, all methods agreed on the designation of the ipsilateral breast cancer as a new primary tumor or a true recurrence; however, for five pairs and three pairs, both clustering methods and clustering by DNA breakpoints, respectively, agreed with the clinical definition. For 21 pairs, the partial identity score confirmed the designation of the tumor as defined by both clustering methods. Patients with recurrences had poorer metastasis-free survival than patients with new primary tumors, according to the partial identity score, but this difference was not statistically significant using the clinical definition.

Implications

The partial identity score may outperform clinical determination for the prognosis of ipsilateral breast cancers.

Limitations

Freshly frozen tissue samples that contain a large number of cells from both the initial primary tumor and the ipsilateral tumor are needed to perform the DNA breakpoint analyses.

Subjects and Methods

Selection of Patients

Specimens from patients with primary breast cancers and ipsilateral breast cancers were selected from freshly frozen samples of the Institut Curie tissue bank according to the following criteria: the primary tumor was either ductal or lobular invasive breast carcinoma; the patient was 49 years or younger at diagnosis of the initial tumor; all patients were premenopausal; and there was no previous history of cancer, except for one nonmelanoma skin cancer. All patients had been treated at the Institut Curie by breast-conserving surgery, including dissection of the axillary lymph nodes in most patients, followed by radiotherapy to the breast with or without a boost to the tumor bed (external beam radiotherapy or brachytherapy) and/or to the regional lymph node–bearing areas if indicated and, when required, systemic treatment as part of their initial management. For all tumors, histopathologic characteristics were reviewed by one pathologist (B. Sigal-Zafrani).

To ensure that the data would be informative, we restricted genomic analyses to tumors (primary and recurrences) in which at least 50% of cancer cells had been assessed by hematoxylin, eosin, and saffron staining of sections from snap-frozen samples. This study reports a series of 22 patients with assessable pairs of primary breast tumors and ipsilateral breast cancers.

To evaluate the genomic features of a population with similar breast cancers, 44 control patients from the pool of patients with primary tumors who met the above selection criteria were matched to the case patients in accordance with their age at diagnosis and adjuvant treatment. The control patients had not experienced an ipsilateral breast recurrence within the time span of the local recurrence of the index patient.

This research was approved by the institutional review boards of the Institut Curie. No patient refused the use of her tumor specimens for research purposes.

Clinical and Histologic Studies

The histologic/biologic properties of the breast cancers were determined by subjecting tissue sections to immunohistochemical analysis for the estrogen receptor (clone 6F11, 1:200 dilution; Novocastra, Newcastle Upon Tyne, England) and progesterone receptor (clone 1A6, 1:200 dilution; Novocastra) antibodies. Tumors were considered to be positive for these receptors if at least 10% of the invasive tumor cells in a section showed nuclear staining (21).

In accordance with theories of the clonal evolution of tumor cell populations, ipsilateral breast cancers were clinically defined as true recurrences if they had the same histologic subtype (ductal or

lobular) and a similar or increased growth rate, similar or loss of dependence on either estradiol or progesterone, and similar or decreased differentiation as the initial tumor (22).True recurrences also had to share with their primary tumors the same breast quadrant. Thus, new primary tumors were clinically defined as such when the ipsilateral breast cancer had occurred in a different location, had a distinct histologic type, or had less aggressiveness features (lower grade, appearance of hormonal receptors) than the initial tumor.

Genomic Studies

Total genomic DNA was extracted from tissue samples using a variation of the standard phenol : chloroform protocol (23). Genomic DNA was quantified by spectrophotometry using a ND-1000 Spectrophotometer (NanoDrop, Wilmington, DE), and quality was assessed by 0.8% agarose gel electrophoresis.

Genomic DNA from each sample was prepared for microarray hybridization using the GeneChips Mapping 50K Xba Assay Kit (Affymetrix Inc., Santa Clara, CA). Briefly, 250 ng of total genomic DNA was digested with the restriction enzyme XbaI and ligated to an adaptor sequence (XbaI adaptator: 5′-ATTATGAGCACGAC AGACGCCTGATCT-3′ and 5′-CTAGAGATCAGGCGTCTG TCGTGCTCATAA-3′) that recognizes the cohesive four base pair (bp) region (3′-GATC-5′). A generic primer (5′-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3′) that recognizes the adaptor sequence was used to preferentially amplify adaptor-ligated DNA fragments 250–2000 bp in size by the optimized polymerase chain reaction (PCR) conditions, according to the manufacturer's instructions. The amplified DNA was then fragmented by DNase treatment and hybridized to the Affymetrix GeneChips Human Mapping 50K array Xba 240 (Affymetrix), according to the manufacturer's instructions. Washing, staining, and scanning of chips were performed using materials and methods provided by the manufacturer. The pangenomic profiles of the 22 pairs of primary tumors/ipsilateral breast cancers are available on ACTuDB (24) (http://bioinfo.curie.fr/actudb/). Human mapping 50K array Xba 240 annotations and sequence files are available on the Affymetrix website (http://www.affymetrix.com/support/technical/byproduct.affx?product=100k).

Metastasis-Free Survival

Metastasis-free survival was estimated by the Kaplan–Meier method (25) and compared between the groups of patients defined as having been diagnosed with either a true recurrence or a new primary tumor using the log-rank test. The confidence interval (CI) of the hazard ratio was obtained using a semiparametric Cox model (26).

Statistical Methods

Copy Number Alteration Determination. SNP data were gathered from the pangenomic profile and analyzed using the iterative and alternative normalization of copy number SNP array (ITALICS) algorithm with default parameters, which simultaneously normalizes the genomic profile and detects the biologic signal. Briefly, ITALICS alternatively estimates the biologic signal (ie, the DNA copy number at each SNP locus) with the gain and loss analysis of DNA algorithm (27) and normalizes the data to correct the nonrelevant effects (CG content and fragment length of PCR products, oligonucleotide CG content, and SNP effect). These two steps are repeated iteratively to improve the biologic signal estimation until no more improvement is seen. ITALICS outperforms other methods of normalization. The result of this process is a segmented genomic profile that consists of regions of homogeneous DNA and information on their corresponding copy numbers. Each region is given a smoothing value (ie, the median of the SNP copy numbers within the region) and a status (ie, gain, normal, or loss).

We defined a breakpoint as 1) a SNP locus located at a change of status (eg, normal/gain or gain/loss) or as 2) a SNP locus located at a change of smoothing value that occurred within a region of gain or loss, thus defining different levels of gain or loss among these regions. Additional breakpoints were also added at the extremities of the chromosome to take into account their gain or loss whenever applicable. Because some breakpoints could be due to copy number variations that occur in healthy individuals, breakpoints arising in the copy number variable regions in the HapMap collection (28) were excluded. The visualization and further analysis of the data was performed through a graphic user interface, Visualization and analysis of array CGH, transcriptome and other molecular profiles (29).

Hierarchical Clustering. *Similarity between genomic profiles.* We considered two measures of similarity among the genomic profiles of a primary tumor and ipsilateral breast cancer. First, we used the Pearson correlation between their CNA profiles. Second, we used a measure M that is derived from the percent concordance proposed by Waldman et al. (18) and adapted from Dice's formula (30) and corresponds to the number of common breakpoints divided by the mean number of breakpoints in either a primary tumor or an ipsilateral breast cancer. M is computed as follows, for a (i,j) pair.

$$M_{i,j} = \frac{\#(S_i \cap S_j)}{1/2 \times (\#S_i + \#S_j)},$$

in which S_i and S_j are the subsets of breakpoints present in the SNP arrays of the primary tumor, i, and ispilateral breast cancer, j. An example of M is given in Supplementary Fig. 1 (available online).

Two tumors had common breakpoints if the following conditions were fulfilled: 1) the changes in copy number occurred at the exact same locus and 2) the changes in copy number were of the same nature (ie, either an increase or a decrease in numbers) in the two tumors.

Assessing clonal relatedness from a dendrogram. We assumed that clonal unrelatedness was revealed by the clustering apart of the two tumors (primary tumor and ipsilateral breast tumor) from the same patient, reflecting that they were more similar to carcinomas of other patients than to each other. In contrast, the clustering together of two tumors from the same patient indicated clonal relatedness among them. For both measures of similarity (Pearson coefficient and M measure), we used Ward's criteria (31) as an agglomerative method in the hierarchical clustering.

Partial Identity Score. *Score definition.* To distinguish true recurrences from new primary tumors, we developed a partial identity score

Table 1. Patient and tumor characteristics of the 22 patients whose tumors (both PT and IBC) had exploitable SNP arrays*

Pair	Age, y	Family	Prob	BRCA1	BRCA2	pT	pN	Surgical margin, mm	Radiotherapy dose, Gy		No. of cycles of chemotherapy†
									Whole breast	Tumorectomy bed	
P1	23.1	0	20	0	2	1	0	≥4	54	54	4
P2	42.1	1	NA	NA	NA	1	0	≥4	50	50	0
P3	42.6	0	NA	NA	NA	1	0	≥4	54	54	0
P4	48.2	1	44	0	0	1	0	≥4	50	50	0
P5	45.5	0	NA	NA	NA	1	1	≥4	50	60	4
P6	35.7	0	8	0	0	2	0	≥4	51	66	4
P10	46.2	0	NA	NA	NA	2	0	0–3	50	70	0
P11	49.0	1	95	0	1	2	0	≥4	50	64	0
P12	48.9	1	NA	NA	NA	1	0	≥4	52	52	0
P13	45.0	0	NA	NA	NA	2	0	≥4	51	67	6
P14	43.6	0	NA	NA	NA	1	0	0–3	50	50	0
P15	46.1	0	NA	NA	NA	1	0	≥4	50	65	0
P16	48.4	0	NA	NA	NA	1	0	≥4	50	66	0
P18	27.9	1	82	0	0	2	0	0–3	50	70	4
P19	49.1	0	NA	NA	NA	2	0	0–3	51	65	4
P20	47.1	0	NA	NA	NA	2	1	0–3	45	65	4
P21	46.3	0	NA	NA	NA	1	0	DCIS	50	70	0‡
P22	35.0	0	NA	NA	NA	2	2	≥4	50	75	6‡
P23	30.8	0	NA	NA	NA	2	0	≥4	50	66	4
P24	47.7	0	NA	NA	NA	1	1	≥4	50	60	6
P25	43.0	0	NA	NA	NA	1	0	0–3	45	60	0‡
P26	30.5	0	NA	NA	NA	NA	1	≥4	52	70	4‡

* PT = primary tumor; IBC = ipsilateral breast cancer; SNP = single nucleotide polymorphism; Family = family history of breast cancer in the first two degrees (0 = no, 1 = yes); Prob = age-specific risk estimates of breast cancer according to the Claus Model (32); BRCA1 and BRCA2 = mutation found in BRCA1 and BRCA2 (0 = not found, 1 = deleterious, 2 = possibly deleterious, NA = not available); pT = histologic tumor classification according to Union Internationale Contre le Cancer (UICC) (33); pN = histologic lymph node classification according to UICC; DCIS = ductal carcinoma in situ.

† Chemotherapy consisted of 5-fluorouracil, anthracyclines, and cyclophosphamide.

‡ Patients were treated with tamoxifen for 5 years.

that is based on the *M* measure of similarity described above. The score reflects the number of common breakpoints among the ipsilateral breast cancer and the primary tumor. In addition, because very frequent breakpoints may be less informative than frequent ones in estimating the clonal relatedness between two tumors, the added value of each breakpoint was weighted according to its frequency among the samples of 44 control patients. The partial identity score (PS) was thus

$$\mathrm{PS}_{i,j} = \frac{\sum_{k\in(S_i\cap S_j)} (1-F_k)^2}{1/2\times[\sum_{k\in S_i}(1-F_k)+\sum_{k\in S_j}(1-F_k)]},$$

in which F_k represents the frequency of appearance of the breakpoint k calculated in the series of the 44 control breast cancers. An example of a partial identity score is given in Supplementary Fig. 1 (available online).

Statistical testing for partial identity. The partial identity score was calculated for all 462 possible "artificial pairs" (462 = 22 × 21, because each of the 22 primary tumors could be artificially paired with the ipsilateral breast cancer of the 21 other patients, *see* Table 3 notes). The distribution under the null hypothesis, H0, of no partial identity between the two tumors was estimated using all 462 possible artificial pairs. We rejected H0 with a type I error fixed at 5%, that is, we considered that a local recurrence shared partial identity with a primary tumor when the score was higher than the upper 5th percentile in the distribution of artificial pairs. The score was then calculated for the "natural pairs," that is, a primary tumor and its ipsilateral breast cancer occurring in the same patients (*see* Table 3 notes). Ipsilateral breast cancers from pairs with scores higher than this cutoff, that is, with shared partial identity, were considered to be true recurrences.

Robustness of the score. The robustness of the partial identity score was assessed by randomly selecting two subgroups of 15 and 7 patients from the population of 22 breast cancer patients. The first subgroup of 15 patients was used to compute the scores of the artificial pairs and to record the cutoff score corresponding to the 95th percentile. This score was then used to determine the status of each of the natural pairs in the seven patients of the other subgroup. To make the comparison statistically sound, each process was repeated 1000 times. The variation of the cutoff scores was assessed by box plot representation. The consistency of the ipsilateral breast cancer status was calculated as the percentage of extractions when the status of this pair was respectively a true recurrence or a new primary tumor.

All statistical tests were two-sided. *P* values less than .05 were considered to be statistically significant.

Results

Clinical and Histologic Features

The clinical and tumor characteristics of 22 patients whose tumors had exploitable SNP arrays were analyzed (Tables 1 and 2). According to clinical and histologic criteria (Table 2), nine of the 22 ipsilateral breast cancers were new primary tumors and the other

Table 2. Histologic characteristics of the primary tumors and their ipsilateral breast cancers: distinctions between new primary tumors and true recurrences according to clinical criteria or clustering methods*

	Primary tumors					Ipsilateral breast cancers					New primary tumors or true recurrences				
Pair	Type	Grade	ER	PR	Time, y	Location	Type	Grade	ER	PR	CNA	BKP	Clinical	Divergence	Score
P1	D	3	0	40	6.5	1	D	2	90	15	TR	NP‡	NP	CNA	0.020
P2	D	2	90	40	5.3	1	L	1	90	70	TR	NP‡	NP	CNA	0.000
P3	D	3	30	80	3.1	1	D	3	60	90	TR	TR‡	TR	No	0.465
P4	L	1	90	80	3.5	1	L	2	90	80	TR	TR‡	TR	No	0.278
P5	D	2	90	40	2.0	1	D	2	80	90	TR	TR‡	TR	No	0.555
P6	L	1	90	100	3.1	1	L	2	70	70	NP	NP‡	TR	Clinical	0.104
P10	L	3	80	95	5.0	0	D	2	70	40	NP	NP‡	NP	No	0.059
P11	L	3	0	0	6.3	1	D	3	0	0	NP	NP‡	NP	No	0.029
P12	L	2	90	50	2.9	0	L	2	90	0	TR	TR†	NP	Clinical	0.116
P13	D	2	20	85	4.6	1	D	2	95	20	TR	TR‡	TR	No	0.240
P14	L	2	90	60	2.5	1	L	2	0	100	TR	TR‡	TR	No	0.310
P15	D	2	100	80	3.3	1	D	2	70	100	NP	TR‡	TR	CNA	0.127
P16	D	2	80	30	3.8	1	D	1	20	70	TR	TR‡	NP	Clinical	0.317
P18	D	3	0	0	2.2	1	D	2	80	50	NP	NP‡	NP	No	0.004
P19§	D	3	0	0	3.0	1	D	3	0	0	TR	TR‡	TR	No	0.325
P20	D	3	0	0	1.4	0	D	3	0	0	TR	TR‡	NP	Clinical	0.139
P21	D	2	80	0	4.2	1	D	2	70		TR	TR‡	TR	No	0.360
P22§	D	2	20	50	3.5	1	M	3	15	0	TR	TR‡	NP	Clinical	0.394
P23	D	3	0	0	0.8	1	D	3	0	0	TR	TR‡	TR	No	0.341
P24§	D	3	0	0	1.0	1	D	3	0	0	TR	TR‡	TR	No	0.311
P25§	D	3	75	70	2.2	1	D	3	70	15	TR	TR‡	TR	No	0.375
P26	D	3	0	0	1.8	1	D	3	0	0	TR	TR‡	TR	No	0.519

* Type = histologic type (D = ductal, L = lobular, M = micropapillary); Grade = histologic grade; ER = estrogen receptor; PR = progesterone receptor; Location (1 = IBC at the index quadrant, 0 = IBC at a different quadrant); CNA = cluster according to copy number alterations; BKP = cluster according to breakpoints; Clinical = definition according to clinical criteria; NP = new primary tumor; TR = true recurrence.

† NP according to the partial identity score.

‡ Agreement with the definition by the partial identity score.

§ The ipsilateral breast cancers of these pairs received chemotherapy before surgery.

13 were true recurrences. Ipsilateral breast cancers occurred at a median time of 3.1 years after the initial breast cancer diagnosis (range = 0.8–6.5 years). In three of 22 (14%) patients, ipsilateral breast cancers occurred in a different quadrant than the initial tumor; all of these were defined clinically as new primary tumors.

Genomic Studies

The pangenomic profiles of a primary tumor and its ipsilateral breast cancer revealed common breakpoints, with a precision within a SNP that can be used as markers of their clonal relatedness. Pair 5 is given as an illustration (Fig. 1).

The median number of breakpoints per array was statistically significantly higher for ipsilateral breast cancers (median = 71, range = 21–433) than for primary tumors (median = 52, range = 4–646) ($P = .001$) (Table 3). The mean number of common breakpoints per pair was also statistically significantly higher for natural pairs (mean = 18.8, SD = 18.8) than for artificial pairs (mean = 4.1, SD = 3.1) ($P = 0.5 \times 10^{-6}$).

Clustering by Copy Number Alterations or Breakpoints

According to hierarchical clustering by DNA CNAs (Fig. 2) and by breakpoints (Fig. 3), five and six ipsilateral breast cancers, respectively, were new primary tumors. The two clustering methods and the clinical definition agreed for 14 pairs (Table 2). However, for five pairs (P6, P12, P16, P20, P22), the clinical definition disagreed with both clustering methods and, for three others (P1, P2, P15), the clustering by breakpoints disagreed with that by CNAs but agreed with the clinical definition. The recurrences in pairs 1 and 2 were identified as true recurrences by the CNA clustering but as new primary tumors by the clinical definitions because of the reappearance of estrogen receptors in the pair 1 ipsilateral breast cancer and different histologic type (ductal instead of lobular carcinoma) in pair 2. In pair 15, CNA clustering did not find a true recurrence, whereas the clinical definition did. No statistically significant differences in clinical and histologic characteristics between the patients diagnosed with new primary tumors or true recurrences were observed by breakpoint information, apart from a suggestion for patients with new primary tumors to be younger and to have a more frequent family history of breast cancer (Supplementary Table 1, available online).

Partial Identity Score

According to the partial identity score reported for each pair in Table 2, 15 ipsilateral breast cancers were true recurrences and seven were new primary tumors (Fig. 4). With a type I error set at 5%, the partial identity score disagreed with clustering by breakpoints in pair 12 only; the clinical definition was new primary tumor because of a change in tumor location. When the score was determined according to Waldman's percent of concordance without either weighing the influence of the coexistence of breakpoints according to their frequency in a similar population or excluding

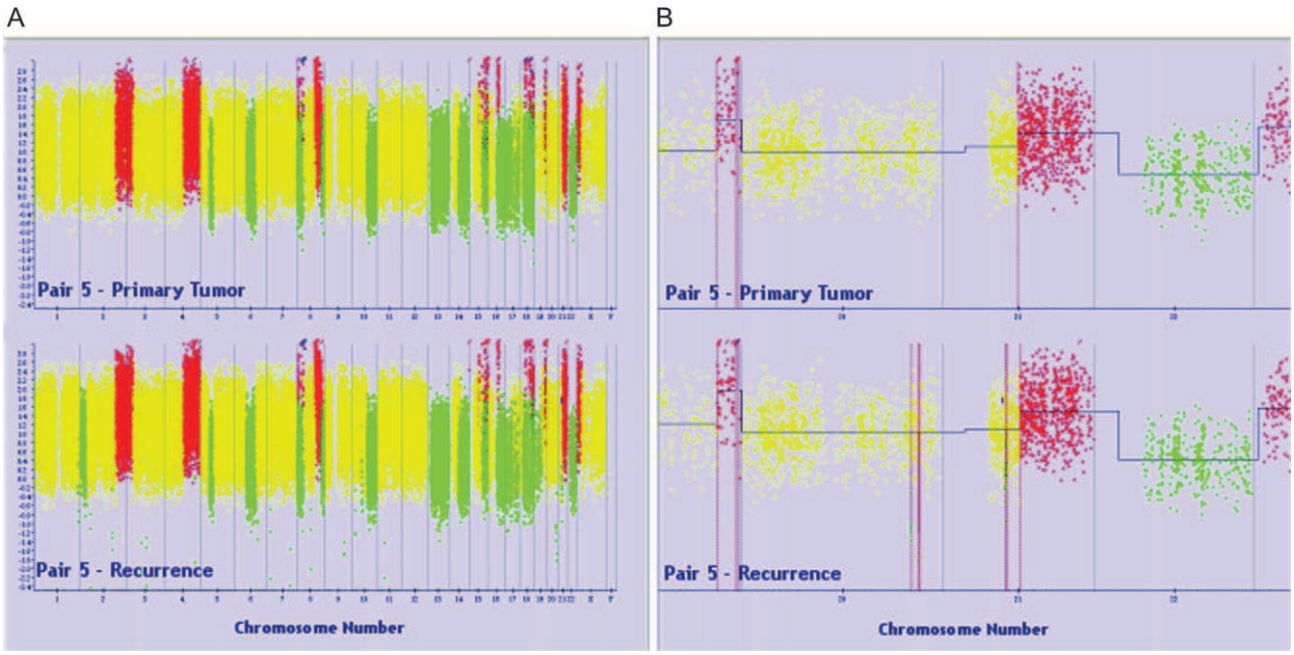


Fig. 1. Genomic profiles of tumors of pair 5 to illustrate the finding of common breakpoints within a single nucleotide polymorphism (SNP). A genomic profile represents the ordered values of the DNA copy numbers obtained as described in "Subjects and Methods". Each **dot** represents the number of DNA copies at each SNP position. Regions with gains are in **red**, with losses in **green**, with no DNA copy number alterations in **yellow**. **A**) Pangenomic profiles. **B**) Profiles of chromosomes 20, 21, and 22. Top primary tumor of pair 5; bottom, ipsilateral breast cancer of pair 5. The **blue horizontal line** represents the smoothing line and the **dotted vertical line** the breakpoint position.

the breakpoints that occur in the copy number variable regions in the HapMap collection, the attribution of the status of three pairs (20 changed from a true recurrence to a new primary, whereas 6 and 12 became true recurrences) and two pairs (10 and 12 changed from new primaries to true recurrences) changed, respectively.

The status of all pairs was confirmed by the 1000 random extractions (Supplementary Table 2, available online). The mean cutoff value was 0.1203 (SD = 0.0102) (Supplementary Fig. 2, available online). The cutoff used to determine the status of the 22 ipsilateral breast cancers, which was defined using all 462 artificial pairs, was 0.1212.

Prognostic Value of the Determination of the Nature of the Ipsilateral Breast Cancer

Patients who were diagnosed with true recurrences had lower metastasis-free survival than those diagnosed with new primary tumors (Supplementary Fig. 3, available online). The difference in metastasis-free survival in the two groups was not statistically significant when they were defined based on clinical and histologic characteristics (5-year metastasis-free survival: 76%, 95% CI = 52% to 100% for new primary tumors and 38%, 95% CI = 17% to 83% for true recurrences; $P = .18$; primary tumors vs true recurrences, hazard ratio = 2.8, 95% CI = 0.6 to 13.7). However, metastasis-free survival was different when the groups were defined according to the partial identity score (5-year metastasis-free survival: 100% for new primary tumors and 29%, 95% CI = 11% to 78% for true recurrences; $P = .01$).

Discussion

DNA breakpoint information was more often in agreement with the clinical definition than that from CNAs to define true recurrences among ipsilateral breast cancers in this population. We developed a partial identity score that is based on DNA breakpoints, which allowed statistical discrimination between new primary tumors and true recurrences. This score outperformed the clinical prognosis determination in terms of metastasis-free survival.

We chose to base our study on a series of young (<50 years old) premenopausal women not only because young age is recognized as one of the most important independent prognostic factors for ipsilateral breast recurrence (34–40) but also to ensure a very high level of homogeneity. In addition, all patients had undergone breast-conserving surgery followed by whole-breast radiotherapy for their initial breast cancers, which were selected as either ductal or lobular invasive carcinomas, and were treated at the same cancer center.

Our results show that some ipsilateral breast cancers share with their primary tumors many DNA CNA breakpoints at the same locations (precision to within a SNP, as illustrated in Fig. 1). From these observations, we produced a method of determining true recurrences that relies on a number of assumptions. The first and most obvious is that the vast majority of breast cancers are of clonal origin. The second is that a tumor retains a substantial number of genomic alterations throughout its evolution. The third assumption, which is key to the method that we have developed, is that the exact locations of the breakpoints that are on the edge of a given change in DNA copy numbers are better hallmarks of a given tumor than the magnitude or width of the genomic alteration itself. For example, because the deletion that causes the loss of Phosphatase and TENsin homolog (PTEN) alters regulatory pathways that lead to precocious development and neoplasia in the mammary gland (41), it can be found in many breast cancers (42–44); however, the exact location of the breakpoints bordering this deletion can be specific to a given tumor. We provide as an

Table 3. Number of common breakpoints in natural (same patient) and artificial (two different patients) pairs of primary tumors (vertically) and ipsilateral breast cancers (horizontally)

No. of BKPs in		No. of BKPs in PT*																					
		77	11	46	16	94	8	22	4	31	55	12	11	58	646	89	69	127	49	60	57	41	72
IBC*	Pair	P1	P2	P3	P4	P5	P6	P10	P11	P12	P13	P14	P15	P16	P18	P19	P20	P21	P22	P23	P24	P25	P26
433	P1	6†	3	12‡	3	8	5§	5	1	4	5	6	1	1	7§	8	6	7	3	8	8	5	12‡
25	P2	0	0†	1	0	1	0	0	0	3‡	0	1	0	0	0	1	0	2	1	0	1	0	0
43	P3	3	2	23†‡§	5	5	2	10§	2§	4	6	5	4	3	4	11	5	7	6	4	8	4	9
26	P4	5	3	7	9†‡§	5	2	7	0	6§	4	4	3	2	0	9‡	3	4	5	3	6	3	5
128	P5	3	3	11	4	64†‡§	1	7	0	4	4	5	2	2	2	8	4	3	8	3	2	3	10
21	P6	3	3	4‡	3	3	3†	4‡	0	4‡	1	4‡	2	0	0	3	1	2	1	1	2	4‡	2
23	P10	3	2	4	3	3	1	3†	1	2	2	1	1	1	3	5‡	1	1	2	1	1	5‡	3
97	P11	5	2	19‡	6	9	1	9	2†§	6§	9	7	6§	5	7§	14	7	10	9	4	12	4	13
35	P12	6‡	3	4	5	4	2	3	0	6†‡§	2	2	3	2	0	4	3	3	3	1	4	4	4
74	P13	3	2	7	3	6	1	5	1	3	18†‡§	4	3	2	2	7	3	3	4	2	2	5	2
35	P14	1	2	7	3	7	3	5	0	3	5	10†‡§	2	1	3	6	3	4	3	2	3	5	4
49	P15	5	2	5	3	4	2	3	0	6‡§	4	1	5†	4	2	3	2	4	3	1	1	2	2
84	P16	2	2	3	2	3	0	2	0	4	2	0	3	23†‡§	1	1	1	3	2	0	3	3	4
53	P18	2	2	9‡	3	3	1	5	1	3	2	3	2	0	2†	7	5	3	2	3	2	3	5
150	P19	9§	4§	18	5	8	2	10§	2§	3	10	5	5	5	7§	42†‡§	13†	11	6	11	10	6	10
93	P20	4	1	6	1	5	0	3	1	2	4	1	2	1	5	7	12†‡	3	4	6	3	3	6
219	P21	2	1	12	3	6	1	5	2§	2	5	3	4	4	6	8	7	63†‡§	6	7	8	3	5
100	P22	5	2	17	5	8	1	10§	1	5	5	5	4	5	3	13	9	10	31†‡§	6	10	5	9
73	P23	7	1	10	3	6	1	7	2§	3	5	5	2	1	5	12	10	6	6	25†‡§	6	3	10
69	P24	6	2	11	5	3	2	6	1	4	5	3	2	3	5	9	5	5	3	7	23†‡§	1	11
42	P25	4	3	9	5	5	2	7	2§	4	5	5	2	2	2	5	4	4	6	1	2	18†‡§	3
88	P26	5	3	11	7	7	1	9	1	6§	5	3	2	4	3	17	5	2	8	5	9	3	43†‡§

* Number of BKPs per tumor. BKP = breakpoint; PT = primary tumor; IBC = ipsilateral breast cancer.

† Numbers correspond to the 22 natural pairs of PTs and their IBCs arising in the same patient; numbers in the other cells correspond to the 462 (22 × 21) artificial pairs of each PT with all other possible IBCs arising in other patients.

‡ Pairs with the most common BKPs per PT.

§ Pairs with the most common BKPs per IBC.

example (Supplementary Fig. 4, available online) the prototype case of PTEN deletion in which the breakpoints are identical between the primary tumor and ipsilateral breast cancer of pair 5 and yet differ in all the other tumors that also harbor a loss of PTEN.

Because clustering is commonly used to determine whether two tumors are clonally related and because it performs better than previously developed similarity scores (18,19), we addressed the issue of whether there was added value in looking at breakpoints rather than at CNAs by comparing clustering by CNAs and by breakpoints to determine the nature of the ipsilateral breast cancer. We concluded from the comparison of clusterings of CNAs and of breakpoints that breakpoint information is more valid than CNA information because when they were discordant, the definition by breakpoints always agreed with the clinical definition, which is routinely used in clinical practice.

A second issue was whether a method could be found to quantify the partial identity between two tumors. We chose to use a partial identity score rather than the results of clustering for a number of reasons. 1) Clustering methods have been designed for exploratory data analysis, so that using a score is more appropriate for a discrimination purpose. 2) A score induces a natural ordering of the pairs from the most dissimilar to the most similar, which is not the case for clustering. 3) The assessment of clonal relatedness by a score can be statistically motivated through the choice of a threshold, as we have demonstrated in the present work. For clustering, clonal relatedness of two tumors depends only on their being clustered apart on the dendrogram, which leads to inconsistent decisions over time. As illustrated by Fig. 3, if pair 2 had not been included in the study, the ipsilateral breast cancer from pair 6 would have been considered as a true recurrence rather than a new primary tumor. Conversely, the assessment of the partial identity score robustness was satisfactory with a narrow range of the cutoff (Supplementary Fig. 2, available online) and with the consistency of the ipsilateral breast cancer status (Supplementary Table 2, available online). Moreover, a score allows one to choose the cutoff that best distinguishes new primary tumors from true recurrences. In this study, we chose a type I error rate at 5% to favor sensitivity for diagnosing true recurrences over the specificity. Further studies will be needed to verify the biologic validity of this choice (Supplementary Fig. 3, available online).

In addition, we chose to weigh the influence of a common breakpoint between the ipsilateral breast cancer and its primary tumor by a factor that takes into account the frequency of this given breakpoint in a population of similar tumors. This weighting changed the determination of three of 22 pairs.

The clinical definition considered an ipsilateral breast cancer as a new primary tumor when the partial identity score did not in three instances. In the first because of a change in location for pairs 12 and 20, in the second because of a lesser degree of differentiation for pair 16, and in the third because of a change in histology for pair 22. The first example illustrates the possibility that a true recurrence can occur at a distance from the first cancer. The second exemplifies the possibility for a true recurrence to have many but not all of the striking alterations present in the primary tumor.

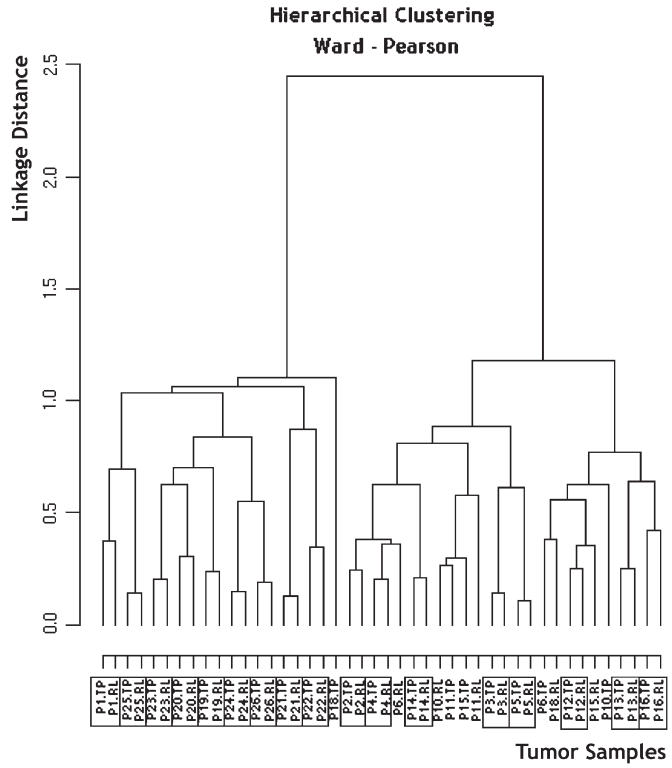


Fig. 2. Dendogram of hierarchical clustering by DNA copy number alterations (Ward–Pearson) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

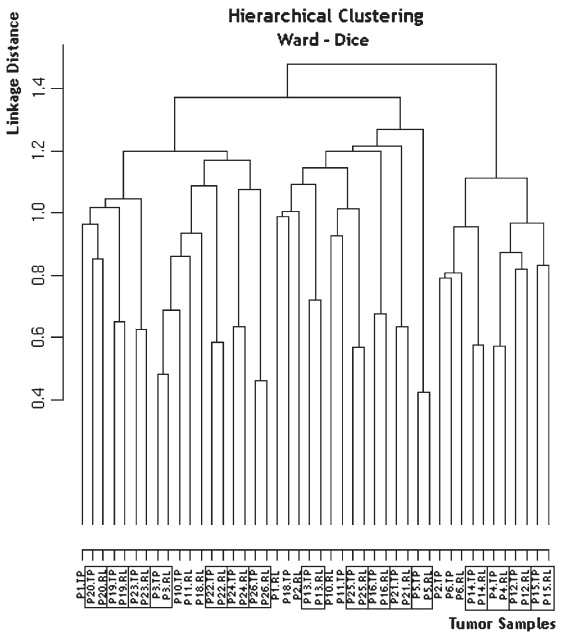


Fig. 3. Dendogram of hierarchical clustering by breakpoints (Ward–Dice) of 22 available pairs of primary tumors (TP) and their ipsilateral breast cancer (RL). **Boxes** represent natural pairs with a true recurrence, that is, a pair of tumors from one patient clustered together.

A criticism that can be made of the clinical definition is that it assumes that a true recurrence is derived from its primary tumor instead of only being related to it. A true recurrence, according to some clinical definitions (5,6,11), cannot be more differentiated than its primary tumor. Usual classifications define differentiation according to histologic grading, DNA ploidy, or the presence of ductal carcinoma in situ. They are based on the assumption that tumors accumulate genetic alterations with time (22,45,46) and that the chronologic order of these alterations reflects the development of a tumor clone. This assumption is, however, challenged by the fact that the ipsilateral breast cancers are neither more aggressive nor more undifferentiated than their primary tumors (47).

The situation with pair 22 illustrates another possible limitation of histologic determination. Here, the clinical status of the ipsilateral breast cancer was of a new primary tumor because its histologic type was a micropapillary carcinoma, whereas the initial tumor was a ductal carcinoma. However, after further histologic analysis, a minor component of micropapillary carcinoma was revealed in the initial carcinoma that otherwise would have been overlooked (Supplementary Fig. 5, available online). This finding implies that, in some instances, the current histologic taxonomy, which is based more on architectural features than on biologic ones, could become obsolete and that some ipsilateral breast cancers could qualify as true recurrences without sharing the same histologic type as their primary tumors.

We observed that patients with true recurrences had lower metastasis-free survival than patients with new primary tumors and that this difference became statistically significant when the partial identity score, instead of clinical definition, was used to define ipsilateral breast cancer types. This observation has been shared by many authors (5,6,10,12). Possible explanations are,

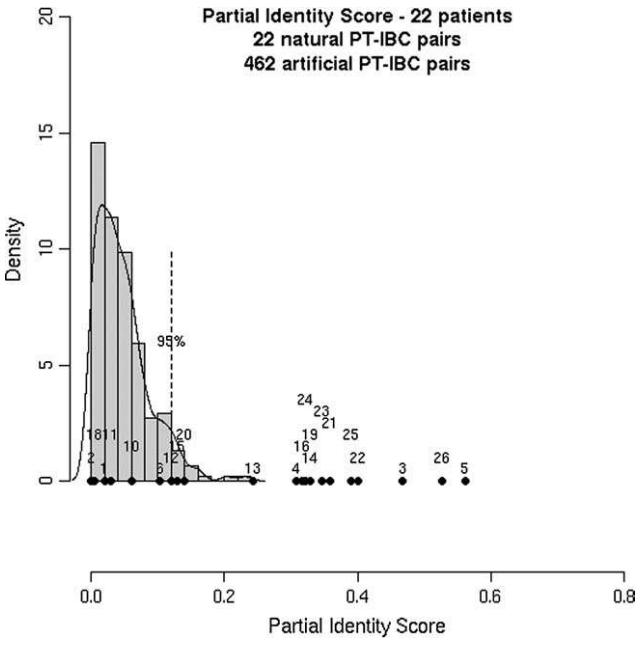


Fig. 4. Partial identity score. Histogram performed on 462 artificial pairs (two different patients) of tumors and representation of the 22 natural (same patient) pairs of primary tumors (PT)/ipsilateral breast cancer (IBC). x-axis: partial identity score (the higher the score, the more likely the IBC is a true recurrence), y-axis: number of artificial pairs in **boxes**. The **vertical dashed bar** represents the upper 5th percentile of the artificial pairs distribution and the threshold above which true recurrences were defined (rejection of the null hypothesis). Each **dot** represents one of the 22 natural pairs (its identifier is written above it).

first, that a true recurrence is the expression of clones that are resistant to adjuvant treatment and therefore could be more difficult to eradicate and, second, that it could be the tip of the iceberg, that is, distant metastases. Conversely, new primary tumors have a prognosis similar to de novo primary cancers but can also reflect a genetic predisposition to develop breast cancer, in the contralateral breast in particular. The clinical implication should therefore be to advocate the use of a systemic treatment in the case of true recurrences and the use of either chemoprevention, such as hormone therapies (48–50) or screening with magnetic resonance imaging (51–53), for patients who are diagnosed with new primary tumors. Here, using breakpoint information led to a better discrimination between new primary tumors and true recurrences in terms of metastasis-free prognosis than the clinical definition.

We also hope that a better distinction among ipsilateral breast cancers of tumors that are genetically related to their primary tumors, that is, true recurrences, will help reveal genetic differences that would provide new information on radioresistance and tumor aggressiveness. To date, little is known about the differential or similarity of the pangenomic expression or the nature of both new primary tumors and ipsilateral breast cancers. Kreike et al. (54) performed a gene expression analysis of 18 000 cDNAs in nine pairs of primary breast cancer with their ipsilateral breast recurrences among women who were younger than 51 years at the time of their initial breast-conserving therapy. Paired data analysis showed no set of genes that had consistently different levels of expression in primary tumors and local recurrences. Another route that has still scarcely been explored is the search for a biologic signature to predict the risk of local recurrence, especially after breast-conserving treatment (54–56). A better distinction between new primary tumors and true recurrences is needed to perform a supervised study based on the occurrence of true recurrences only and not of all ipsilateral breast cancers.

However, our scoring method, which is based on the DNA breakpoint partial identity, has two shortcomings. First, it suffers from the need to conserve unaltered, freshly frozen tissue samples of both the primary tumor and the ipsilateral breast recurrence. This problem should, however, be resolved in time with the possibility of performing the same genomic studies on formalin-fixed paraffin-embedded tissue samples (57–61) or when cryoconservation of either biopsies or fine-needle aspirations (because only 250 ng of DNA is needed, ie, less than 50 000 cells) become standard practice and will make it possible to perform SNP arrays on many more patients. Second, it requires selecting tumors with a cancer cellularity of more than 50%, discarding in the process a number of potentially analyzable tumors. This loss should be diminished in time with both a better selection of frozen tissue material due to the increased experience of the pathologist and the possibility of performing laser capture microdissection.

References

1. Temple WJ, Russell ML, Parsons LL, et al. Conservation surgery for breast cancer as the preferred choice: a prospective analysis. *J Clin Oncol.* 2006;24(21):3367–3373.
2. Clarke M, Collins R, Darby S, et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet.* 2005; 366(9503):2087–2106.
3. Engel J, Kerr J, Schlesinger-Raab A, Sauer H, Holzel D. Quality of life following breast-conserving therapy or mastectomy: results of a 5-year prospective study. *Breast J.* 2004;10(3):223–231.
4. Moyer A. Psychosocial outcomes of breast-conserving surgery versus mastectomy: a meta-analytic review. *Health Psychol.* 1997;16(3): 284–298.
5. Haffty BG, Carter D, Flynn SD, et al. Local recurrence versus new primary: clinical analysis of 82 breast relapses and potential applications for genetic fingerprinting. *Int J Radiat Oncol Biol Phys.* 1993;27(3): 575–583.
6. Huang E, Buchholz TA, Meric F, et al. Classifying local disease recurrences after breast conservation therapy based on location and histology: new primary tumors have more favorable outcomes than true local disease recurrences. *Cancer.* 2002;95(10):2059–2067.
7. Gage I, Recht A, Gelman R, et al. Long-term outcome following breast-conserving surgery and radiation therapy. *Int J Radiat Oncol Biol Phys.* 1995;33(2):245–251.
8. Touboul E, Buffat L, Belkacemi Y, et al. Local recurrences and distant metastases after breast-conserving surgery and radiation therapy for early breast cancer. *Int J Radiat Oncol Biol Phys.* 1999;43(1):25–38.
9. Recht A, Silen W, Schnitt SJ, et al. Time-course of local recurrence following conservative surgery and radiotherapy for early stage breast cancer. *Int J Radiat Oncol Biol Phys.* 1988;15(2):255–261.
10. Komoike Y, Akiyama F, Iino Y, et al. Analysis of ipsilateral breast tumor recurrences after breast-conserving treatment based on the classification of true recurrences and new primary tumors. *Breast Cancer.* 2005;12(2): 104–111.
11. Smith TE, Lee D, Turner BC, Carter D, Haffty BG. True recurrence vs. new primary ipsilateral breast tumor relapse: an analysis of clinical and pathologic differences and their implications in natural history, prognoses, and therapeutic management. *Int J Radiat Oncol Biol Phys.* 2000;48(5): 1281–1289.
12. Schlechter BL, Yang Q, Larson PS, et al. Quantitative DNA fingerprinting may distinguish new primary breast cancer from disease recurrence. *J Clin Oncol.* 2004;22(10):1830–1838.
13. Wang ZC, Buraimoh A, Iglehart JD, Richardson AL. Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. *Breast Cancer Res Treat.* 2006;97(3):301–309.
14. Vicini FA, Antonucci JV, Goldstein N, et al. The use of molecular assays to establish definitively the clonality of ipsilateral breast tumor recurrences and patterns of in-breast failure in patients with early-stage breast cancer treated with breast-conserving therapy. *Cancer.* 2007;109(7): 1264–1272.
15. van der Sijp JR, van Meerbeeck JP, Maat AP, et al. Determination of the molecular relationship between multiple tumors within one patient is of clinical importance. *J Clin Oncol.* 2002;20(4):1105–1114.
16. Shibata A, Tsai YC, Press MF, Henderson BE, Jones PA, Ross RK. Clonal analysis of bilateral breast cancer. *Clin Cancer Res.* 1996;2(4): 743–748.
17. Kuukasjarvi T, Karhu R, Tanner M, et al. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* 1997;57(8):1597–1604.
18. Waldman FM, DeVries S, Chew KL, Moore DH 2nd, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *J Natl Cancer Inst.* 2000;92(4):313–320.
19. Teixeira MR, Ribeiro FR, Torres L, et al. Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. *Br J Cancer.* 2004;91(4):775–782.
20. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* 2004;64(9):3060–3071.
21. Balaton AL, Coindre JM, Collin F, et al. Recommendations for the immunohistochemical evaluation of hormone receptors on paraffin sections of breast cancer. Study Group on Hormone Receptors using

Immunohistochemistry FNCLCC/AFAQAP. National Federation of Centres to Combat Cancer/French Association for Quality Assurance in Pathology. *Ann Pathol.* 1996;16:144–148.
22. Nowell PC. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23–28.
23. Sambrook J, Fritsch EF, Maniatis T. *Molecular Cloning. A Laboratory Manual.* 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
24. Hupe P, La Rosa P, Liva S, Lair S, Servant N, Barillot E. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene.* 2007;26:6641–6652.
25. Kaplan EL, Meier P. Nonparametric estimation from incomplete observation. *J Am Stat Assoc.* 1958;53:457–481.
26. Cox DR, Oakes D. *Analysis of Survival Data*. London: Chapman & Hall; 1984.
27. Hupe P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics.* 2004;20(18):3413–3422.
28. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature.* 2006;444(7118):444–454.
29. La Rosa P, Viara E, Hupe P, et al. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics.* 2006;22(17):2066–2073.
30. Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26:297–302.
31. Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* 1963;58:236–244.
32. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer.* 1994; 73(3):643–651.
33. Sobin LH, Wittekind C. TNM classification of malignant tumours. New York: Wiley-Liss; 2002.
34. Vrieling C, Collette L, Fourquet A, et al. Can patient-, treatment- and pathology-related characteristics explain the high local recurrence rate following breast-conserving therapy in young patients?. *Eur J Cancer.* 2003;39(7):932–944.
35. Fourquet A, Campana F, Zafrani B, et al. Prognostic factors of breast recurrence in the conservative management of early breast cancer: a 25-year follow-up. *Int J Radiat Oncol Biol Phys.* 1989;17(4):719–725.
36. Borger J, Kemperman H, Hart A, Peterse H, van Dongen J, Bartelink H. Risk factors in breast-conservation therapy. *J Clin Oncol.* 1994;12(4): 653–660.
37. Elkhuizen PH, van de Vijver MJ, Hermans J, Zonderland HM, van de Velde CJ, Leer JW. Local recurrence after breast-conserving therapy for invasive breast cancer: high incidence in young patients and association with poor survival. *Int J Radiat Oncol Biol Phys.* 1998;40(4): 859–867.
38. Elkhuizen PH, Voogd AC, van den Broek LC, et al. Risk factors for local recurrence after breast-conserving therapy for invasive carcinomas: a case-control study of histological factors and alterations in oncogene expression. *Int J Radiat Oncol Biol Phys.* 1999;45(1):73–83.
39. Oh JL, Bonnen M, Outlaw ED, et al. The impact of young age on locoregional recurrence after doxorubicin-based breast conservation therapy in patients 40 years old or younger: how young is "young"?. *Int J Radiat Oncol Biol Phys.* 2006;65(5):1345–1352.
40. Bollet MA, Sigal-Zafrani B, Mazeau V, et al. Age remains the first prognostic factor for loco-regional breast cancer recurrence in young (<40 years) women treated with breast conserving surgery first. *Radiother Oncol.* 2007;82(3):272–280.
41. Li G, Robinson GW, Lesche R, et al. Conditional loss of PTEN leads to precocious development and neoplasia in the mammary gland. *Development.* 2002;129(17):4159–4170.
42. Sapolsky RJ, Hsie L, Berno A, Ghandour G, Mittmann M, Fan JB. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genet Anal.* 1999;14(5–6):187–192.
43. Jonsson G, Staaf J, Olsson E, et al. High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization. *Genes Chromosomes Cancer.* 2007;46(6):543–558.
44. Perez-Tenorio G, Alkhori L, Olsson B, et al. PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer. *Clin Cancer Res.* 2007;13(12): 3577–3584.
45. Chen LC, Kurisu W, Ljung BM, Goldman ES, Moore D 2nd, Smith HS. Heterogeneity for allelic loss in human breast cancer. *J Natl Cancer Inst.* 1992;84(7):506–510.
46. Lininger RA, Fujii H, Man YG, Gabrielson E, Tavassoli FA. Comparison of loss heterozygosity in primary and recurrent ductal carcinoma in situ of the breast. *Mod Pathol.* 1998;11(12):1151–1159.
47. Sigal-Zafrani B, Bollet MA, Antoni G, et al. Are ipsilateral breast tumour invasive recurrences in young (40 years) women more aggressive than their primary tumours?. *Br J Cancer.* 2007;97(8):1046–1052.
48. Powles TJ, Ashley S, Tidy A, Smith IE, Dowsett M. Twenty-year follow-up of the Royal Marsden randomized, double-blinded tamoxifen breast cancer prevention trial. *J Natl Cancer Inst.* 2007;99(4):283–290.
49. Cuzick J, Forbes JF, Sestak I, et al. Long-term results of tamoxifen prophylaxis for breast cancer—96-month follow-up of the randomized IBIS-I trial. *J Natl Cancer Inst.* 2007;99(4):272–282.
50. Veronesi U, Maisonneuve P, Rotmensz N, et al. Tamoxifen for the prevention of breast cancer: late results of the Italian Randomized Tamoxifen Prevention Trial among women with hysterectomy. *J Natl Cancer Inst.* 2007;99(9):727–737.
51. Kriege M, Brekelmans CT, Boetes C, et al. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N Engl J Med.* 2004;351(5):427–437.
52. Kuhl CK, Schrading S, Leutner CC, et al. Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. *J Clin Oncol.* 2005;23(33): 8469–8476.
53. Lehman CD, Gatsonis C, Kuhl CK, et al. MRI evaluation of the contralateral breast in women with recently diagnosed breast cancer. *N Engl J Med.* 2007;356(13):1295–1303.
54. Kreike B, Halfwerk H, Kristel P, et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res.* 2006;12(19): 5705–5712.
55. Nuyten DS, Kreike B, Hart AA, et al. Predicting a local recurrence after breast-conserving therapy by gene expression profiling. *Breast Cancer Res.* 2006;8(5):R62.
56. Niméus E, Krogh M, Malmström P, Strand C, Fredriksson I, Karlsson P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence despite postoperative radiotherapy after breast conserving surgery. In: 29th Annual *San Antonio Breast Cancer Symposium* 2006. San Antonio, TX: 2007;103(1):115–124.
57. Isola J, DeVries S, Chu L, Ghazvini S, Waldman F. Analysis of changes in DNA sequence copy number by comparative genomic hybridization in archival paraffin-embedded tumor samples. *Am J Pathol.* 1994;145(6): 1301–1308.
58. Devries S, Nyante S, Korkola J, et al. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J Mol Diagn.* 2005;7(1):65–71.
59. Johnson NA, Hamoudi RA, Ichimura K, et al. Application of array CGH on archival formalin-fixed paraffin-embedded tissues including small numbers of microdissected cells. *Lab Invest.* 2006;86(9): 968–978.
60. Oosting J, Lips EH, van Eijk R, et al. High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. *Genome Res.* 2007;17(3):368–376.
61. Schubert EL, Hsu L, Cousens LA, et al. Single nucleotide polymorphism array analysis of flow-sorted epithelial cells from frozen versus fixed tissues for whole genome analysis of allelic loss in breast cancer. *Am J Pathol.* 2002;160(1):73–79.

Funding

Institut Curie, the "Courir pour la vie, Courir pour Curie" association, the "Odyssea" association and the PHRC 2006 (AOM 06 149).

Notes

M. A. Bollet and N. Servant contributed equally to this work. The authors thank the members of the departments of Tumor Biology (Martial Caly, Blandine Massemin, Michèle Galut), Biostatistics (Eléonore Gravier, Chantal Gautier), Translational Research (David Gentien, Cécile Reyes, Audrey Rapinat, Benoît Albaud, Vincent Lepetit), and Bioinformatics (Philippe La Rosa, Séverine Lair) who participated in this study. The authors are also indebted to Anne Vincent-Salomon, Patricia de Crémoux, Dominique Stoppa-Lyonnet, and particularly Olivier Delattre for their very valuable comments on this work. Finally, they thank all the members of the Institut Curie Breast Cancer Group.

The sponsors had no role in the study design, data collection, interpretation of the results, preparation of the manuscript, or the decision to submit the manuscript for publication.

Manuscript received June 4, 2007; revised October 16, 2007; accepted November 13, 2007.

Bibliography

[1] F. Abramovich and Y. Benjamini. Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In A. A. and O. G., editors, *Wavelets and Statistics*, volume 103, pages 5–14. Springer-Verlag, 1995.

[2] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[3] D. Albertson, B. Ylstra, R. Segraves, C. Collins, S. Dairkee, D. Kowbel, W. Kuo, J. Gray, and D. Pinkel. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genetics*, 25:144–146, 2000.

[4] C. Ambroise. *Approche probabiliste en classification automatique et contraintes de voisinage*. PhD thesis, Université Technique de Compiègne, France, 1996.

[5] C. Ambroise, G. Govaert, and M. D. Dang. Clustering of spatial data by the EM algorithm. In A. Soares, J. Gomes-Hernandez, and R. Froidevaux, editors, *Geostatistics for Environmental Applications*, pages 493–504. Kluwer Academic Publisher, 1997.

[6] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, 25:60–83, 2000.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1):289–300, 1995.

[8] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1198, 2001.

[9] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3): 491, 2006.

[10] G. Blanchard and E. Roquain. Adaptive FDR control under independence and dependence. Arxiv preprint math.ST/0707.0536v2, 2008.

[11] M. A. Bollet, N. Servant, P. Neuvial, C. Decraene, I. Lebigot, J.-P. Meyniel, Y. De Rycke, A. Savignoni, G. Rigaill, P. Hupé, A. Fourquet, B. Sigal-Zafrani, E. Barillot, and J.-P. Thiery. High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers. *J Natl Cancer Inst*, 100(1):48–58, 2008.

[12] P. Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6: 199, Aug 2005.

[13] T. Cai, J. Jin, and M. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007.

[14] A. Célisse and S. Robin. A leave-p-out based estimation of the proportion of null hypotheses. *Arxiv preprint arXiv:0804.1189*, 2008.

[15] Y. Cheng and G. M. Church. Biclustering of expression data. In *Int Conf Intell Syst Mol Biol*, volume 8, pages 93–103, 2000.

[16] Z. Chi. On the performance of FDR control: constraints and a partial solution. *Ann. Statist.*, 35(4):1409–1431, 2007.

[17] Z. Chi. Sample size and positive false discovery rate control for multiple testing. *Electronic Journal of Statistics*, 1:77–118, 2007.

[18] Z. Chi and Z. Tan. Positive false discovery proportions: intrinsic bounds and adaptive control. *Statistica Sinica*, 18(3):837–860, 2008.

[19] K. Chin, S. DeVries, J. Fridlyand, P. Spellman, R. Roydasgupta, W. Kuo, A. Lapuk, R. Neve, Z. Qian, T. Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

[20] W. S. Cleveland and E. Grosse. Computational methods for local regression. *Statistics and Computing*, 1:47–62, 1991.

[21] W. S. Cleveland, S. J. Devlin, and E. Grosse. Regression by local fitting. *Journal of Econometrics*, 37:87 –114, 1988.

[22] A.-A. Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette (reprinted 1984 as vol. 1 of Cournot's Oeuvres Complètes, ed. Bernard Bru, Paris: J. Vrin), 1843.

[23] C. Dalmasso, P. Broët, and T. Moreau. A simple procedure for estimating the false discovery rate. *Bioinformatics*, 21(5):660–668, Mar 2005. Evaluation Studies.

[24] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[25] D. Donoho and J. Jin. Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.*, 34(6):2980–3018, 2006.

[26] M. D. Donsker. An invariance principle for certain probability limit theorems. *Mem. Amer. Math. Soc*, 6:12, 1951.

[27] M. Elati, P. Neuvial, M. Bolotin-Fukuhara, E. Barillot, F. Radvanyi, and C. Rouveirol. LICORN: LearIng COoperative Regulation Networks. *Bioinformatics*, 23(18):2407–2414, 2007.

[28] H. Finner and M. Roters. On the False Discovery Rate and Expected Type I Errors. *Biometrical Journal*, 43(8):985, 2001.

[29] H. Finner and M. Roters. Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.*, 30(1):220–238, 2002.

[30] H. Finner, T. Dickhaus, and M. Roters. On the False Discovery Rate and an Asymptotically Optimal Rejection Curve. Ann. Statist. (to appear).

[31] C. Fuhrmann, O. Schmidt-Kittler, N. Stoecklein, K. Petat-Dutter, C. Vay, K. Bockler, R. Reinhardt, T. Ragg, and C. Klein. High-resolution array comparative genomic hybridization of single micrometastatic tumor cells. *Nucleic Acids Research*, 2008.

[32] A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, and P. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12): 4241–57, 2000.

[33] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar. An adaptive step-down procedure with proven fdr control. Ann. Statist. (to appear).

[34] C. R. Genovese and L. Wasserman. Operating Characteristics and Extensions of the False Discovery Rate. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):499, 2002. doi: 10.1111/1467-9868.00347.

[35] C. R. Genovese and L. Wasserman. A Stochastic Process Approach to False Discovery Control. *Ann. Statist.*, 32(3):1035–1061, 2004.

[36] C. R. Genovese and L. Wasserman. Exceedance Control of the False Discovery Proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

[37] C. R. Genovese, N. A. Lazar, and T. E. Nichols. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *Neuroimage*, 15(4):870–878, 2002.

[38] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[39] W. Gibbs. L'imbroglio génétique du cancer. *Pour la science*, (310): 80–87, 2003.

[40] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100 (1):57–70, Jan 2000.

[41] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004. ISSN 1476-4687 (Electronic).

[42] N. Hengartner and P. Stark. Finite-Sample Confidence Envelopes for Shape-Restricted Densities. *Ann. Statist.*, 23(2):525–550, 1995.

[43] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[44] P. Hupé, P. La Rosa, S. Liva, S. Lair, N. Servant, and E. Barillot. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene*, 2007.

[45] A. Idbaih, B. Boisselier, M. Sanson, E. Crinière, S. Liva, Y. Marie, C. Carpentier, S. Paris, F. Laigle-Donadey, K. Mokhtari, M. Kujas, K. Hoang-Xuan, O. Delattre, and J.-Y. Delattre. Tumor genomic profiling and tp53 germline mutation analysis of first-degree relative familial gliomas. *Cancer Genetics and Cytogenetics*, 176(2):121–126, 2007.

[46] A. Idbaih, M. Kouwenhoven, J. Jeuken, C. Carpentier, T. Gorlia, J. M. Kros, P. French, J. L. Teepen, O. Delattre, J.-Y. Delattre, M. van den Bent, and K. Hoang-Xuan. Chromosome 1p loss evaluation in anaplastic oligodendrogliomas. *Neuropathology*, 28(4):440–443, Aug 2008.

[47] Y. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.*, 1997.

[48] Y. I. Ingster. Minimax detection of a signal for l n-balls. *Math. Methods Statist.*, 7(4):401–428, 1999.

[49] Y. I. Ingster and I. A. Suslina. Nonparametric Goodness-of-Fit Testing under Gaussian Models. *Lecture Notes in Statistics*, 169, 2003.

[50] J. Jin. Detection boundary for sparse mixtures. Unpublished manuscript, 2002.

[51] J. Jin and T. Cai. Estimating the Null and the Proportion of non-Null Effects in Large-Scale Multiple Comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007.

[52] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083): 818–821, 1992. ISSN 0036-8075 (Print).

[53] A. Koren, I. Tirosh, and N. Barkai. Autocorrelation analysis reveals widespread spatial biases in microarray experiments. *BMC Genomics*, 8:164, 2007. ISSN 1471-2164 (Electronic).

[54] P. La Rosa, E. Viara, P. Hupé, G. Pierron, S. Liva, P. Neuvial, I. Brito, S. Lair, N. Servant, N. Robine, E. Manié, C. Brennetot, I. Janoueix-Lerosey, V. Raynal, N. Gruel, C. Rouveirol, N. Stransky, M.-H. Stern, O. Delattre, A. Aurias, F. Radvanyi, and E. Barillot. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*, 22(17):2066–2073, Sep 2006.

[55] M. Langaas, B. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J Bioinform Comput Biol*, 67(4):555–572, 2005.

[56] E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *Ann. Statist*, 33(3):1138–1154, 2005.

[57] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[58] W. K. Lim, K. Wang, C. Lefebvre, and A. Califano. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 23(13):282–288, Jul 2007. Comparative Study.

[59] S. Liva, P. Hupé, P. Neuvial, I. Brito, E. Viara, P. La Rosa, and E. Barillot. CAPweb: a bioinformatics CGH array Analysis Platform. *Nucleic Acids Res*, 34(Web Server issue):477–481, Jul 2006.

[60] L. A. Loeb, K. R. Loeb, and J. P. Anderson. Multiple mutations and cancer. *Proc Natl Acad Sci U S A*, 100(3):776–781, Feb 2003.

[61] N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 2006.

[62] C. J. Miller, C. R. Genovese, R. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, J. Schneider, and A. Moore. Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6):3492–3505, 2001.

[63] J. A. Morris, S. A. Gayther, I. J. Jacobs, and C. Jones. A suite of perl modules for handling microarray data. *Bioinformatics*, 24(8): 1102–1103, 2008.

[64] P. Neuvial. Asymptotic properties of false discovery rate controlling procedures under independence. In revision for Electronic Journal of Statistics, 2008.

[65] P. Neuvial. Intrinsic bounds and false discovery rate control in multiple testing problems. Submitted, 2008.

[66] P. Neuvial, P. Hupé, I. Brito, S. Liva, E. Manié, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7(1):264, May 2006.

[67] D. Pe'er, A. Regev, and A. Tanay. Minreg: inferring an active regulator set. *Bioinformatics*, 18:258–267, 2002.

[68] M. Perone Pacifico, C. R. Genovese, I. Verdinelli, and L. Wasserman. False Discovery Control for Random Fields. *Journal of the American Statistical Association*, 99(468):1002–1015, 2004.

[69] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, 1998.

[70] D. B. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[71] D. B. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[72] S. Pounds and C. Cheng. Improving false discovery rate estimation. *Bioinformatics*, 20(11):1737–1745, Jul 2004.

[73] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(1):35, 2007.

[74] G. Rigaill, P. Hupé, A. Almeida, P. La Rosa, J. Meyniel, C. Decraene, and E. Barillot. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. *Bioinformatics*, 24 (6):768, 2008.

[75] J. P. Romano and A. M. Shaikh. On stepdown control of the false discovery proportion. In J. Rojo, editor, *Optimality: The Second Erich L. Lehmann Symposium*, volume 33, Beachwood, Ohio, USA, 2006. Institute of Mathematical Statistics.

[76] J. P. Romano and A. M. Shaikh. Step-up procedures for control of generalizations of the family-wise error rate. *Ann. Statist.*, 34:1850–1873, 2006.

[77] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *Ann. Statist*, 35(4):1378–1408, 2007.

[78] M. Ruiz, K. Floor, P. Roepman, J. Rodriguez, G. Meijer, W. Mooi, E. Jassem, J. Niklinski, T. Muley, N. van Zandwijk, E. F. F. Smit, K. Beebe, L. Neckers, B. Ylstra, and G. Giaccone. Integration of gene dosage and gene expression in non-small cell lung cancer, identification of hsp90 as potential target. *PLoS ONE*, 3(3), 2008.

[79] S. K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30(1):239–257, 2002.

[80] T. Schweder and E. Spjotvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

[81] A. Shabalin, H. Tjelmeland, C. Fan, C. Perou, and A. Nobel. Merging two gene expression studies via cross platform normalization. *Bioinformatics*, 2008.

[82] J. P. Shaffer. Multiple hypothesis testing. *Annual Reviews in Psychology*, 46(1):561–584, 1995.

[83] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751, 1986.

[84] A. V. Skorokhod. Limit Theorems for Stochastic Processes. *Theory of Probability and its Applications*, 1:261, 1956.

[85] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. L. S, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29: 263–4, 2001.

[86] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874, 2001. ISSN 0027-8424 (Print).

[87] T. P. Speed, editor. *Statistical analysis of gene expression data*. CRC Press, 2003.

[88] P. Spellman, G. Sherlock, M. Zhang, V. I. V. K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–97, 1998.

[89] J. D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002.

[90] J. D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035, 2003.

[91] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, Aug 2003.

[92] J. D. Storey and R. Tibshirani. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software*, pages 272–290. Springer, New York, 2003.

[93] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of

false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004.

[94] M. Subbotin. On the law of frequency of errors. *Matematicheskii Sbornik*, 31:296–301, 1923.

[95] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[96] W. Sun and T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Ass.*, 102:901–912, 2007.

[97] J. W. H. Swanepoel. The limiting behavior of a modified maximal symmetric 2s-spacing with applications. *Ann. Statist.*, 27(1):24–35, 1999.

[98] M. R. Teixeira, F. R. Ribeiro, L. Torres, N. Pandis, J. A. Andersen, R. A. Lothe, and S. Heim. Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. *Br J Cancer*, 91(4): 775–782, 2004. ISSN 0007-0920 (Print).

[99] W. J. Temple, M. L. Russell, L. L. Parsons, S. M. Huber, C. A. Jones, J. Bankes, and M. Eliasziw. Conservation surgery for breast cancer as the preferred choice: a prospective analysis. *J Clin Oncol*, 24(21): 3367–3373, 2006. ISSN 1527-7755 (Electronic).

[100] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

[101] A. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, 2003.

[102] J. Tukey. T13 N: The Higher Criticism. *Course notes, Princeton University*, 1976.

[103] K. Unger, E. Malisch, G. Thomas, H. Braselmann, A. Walch, G. Jackl, P. Lewis, E. Lengfelder, T. Bogdanova, J. Wienberg, and Z. H. Array CGH demonstrates characteristic aberration signatures in human papillary thyroid carcinomas governed by RET/PTC. *Oncogene*, 2008.

[104] M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat Appl Genet Mol Biol*, 3(1):15, 2004. doi: 10.2202/1544-6115.1042.

[105] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[106] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[107] L. J. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[108] M. Vermeer, R. van Doorn, R. Dijkman, X. Mao, S. Whittaker, P. van Voorst Vader, M. Gerritsen, M. Geerts, S. Gellrich, O. Soderberg,

et al. Novel and highly recurrent chromosomal alterations in sezary syndrome. *Cancer Research*, 68(8):2689, 2008.

[109] W. B. Wu. On false discovery control under dependence. *Ann. Statist.*, 36(1):364–380, 2008.

[110] Z. Wu, R. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–918, 2004.

[111] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[112] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Appendix

APPENDIX A

Vignette of R package MANOR

MANOR: Micro-Array NORmalization of array-CGH data

Pierre Neuvial [1,2,3], Philippe Hupé [1,2,3,4], Isabel Brito [1,2,3], Emmanuel Barillot [1,2,3]

September 4, 2008

1. Institut Curie, 26 rue d'Ulm, Paris cedex 05, F-75248 France
2. INSERM, U900, Paris, F-75248 France
3. École des Mines de Paris, ParisTech, Fontainebleau, F-77300 France
4. UMR 144 CNRS, Paris, F-75248 France
`manor@curie.fr`

Contents

1 Overview

This document gives an overview of the *MANOR* package, which is devoted to the normalization of Array Comparative Genomic Hybridization (array-CGH) data(9; 7; 8; 4; 3). Normalization is a crucial step of microarray analysis which aims at separating biologically relevant signal from experimental artifacts. Typical input data is a file generated by an image analysis software such as Genepix or SPOT (5), containing several measurements for each biological variable of interest, i.e. several replicated *spots* for each *clone*; this spot-level data is filtered with various statistical criteria (including a spatial bias detection step which is described in (6)), and aggregated into clean clone-level data.

Using the *arrayCGH* framework developped in the package GLAD, which is available under Bioconductor. We propose the formalism of `flags` to handle clone and spot filtering: the core of the normalization process consists in applying to an *arrayCGH* object a list of flags that successively exclude from the data all irrelevant spots or clones.

We also define quality scores (`qscores`) that quantify the quality of an array after normalization: these scores can be used directly to compare the quality of different arrays after the same normalization process, or to compare the efficiency of different normalization processes on a given array or on a given batch of arrays.

This document is organized as follows: after a short description of optional items we add to *arrayCGH* objects (section 2, we introduce the classes *flag* (section 3) and *qscore* (section 4) with their attributes and dedicated methods; then we describe two useful graphical representation functions (section 6), namely `genome.plot` and `report.plot`; Afterwards we give a short description of the array-CGH datasets we provide (section 5); finally we illustrate the usage of *MANOR* by a sample R script (section 7).

2 *arrayCGH* class

For the purpose of normalization we have added several optional items to the *arrayCGH* objects defined in the R package *GLAD*, including:

cloneValues a data frame with aggregated (clone-level) information, quite similar to *profileCGH* objects of *GLAD*

id.rep the name of a variable common to `cloneValues` and `arrayValues`, that can be used as an identifier for the replicates.

3 *flag* class

We view the process of filtering microarray data, and especially array-CGH data, as a succession of steps consisting in *excluding* from the data unreliable spots or clones (according to criteria such as signal to noise ratio or replicate consistency), and *correcting* signal values from various non-biologically relevant sources of variations (such as spotting effects, spatial effects, or intensity effects).

We introduce the formalism of *flags* to deal with this filtering issue: in the two following subsections, we describe the attributes and methods devoted to *flag* objects.

3.1 Attributes

A *flag* object `f` is a list whose most important items are a function (`f$FUN`) which has to be applied to an object of class *arrayCGH*, and a character value (`f$char`) which identifies flagged spots. Optionally further arguments can be passed to `f$FUN` via `f$args`, and a label can be added via `f$label`. The examples of this subsection use the function `to.flag`, which is explained in subsection 3.2.

3.1.1 Exclusion and correction flags

As stated above, we make the distinction between flags that *exclude* spots from further analysis and flags that *correct* signal values:

exclusion flags If `f` is an exclusion flag, `f$FUN` returns a list of spots to exclude and `f$char` is a non `NULL` value that quickly identifies the flag. In the following example, we define `SNR.flag`, a *flag* objects that excludes spots whose signal to noise ratio lower than the threshold `snr.thr`.

```
> SNR.FUN <- function(arrayCGH, var.FG, var.BG, snr.thr) {
+     which(arrayCGH$arrayValues[[var.FG]] < arrayCGH$arrayValues[[var.BG]] *
+         snr.thr)
+ }
> SNR.char <- "B"
> SNR.label <- "Low signal to noise ratio"
> SNR.flag <- to.flag(SNR.FUN, SNR.char, args = alist(var.FG = "REF_F_MEAN",
+     var.BG = "REF_B_MEAN", snr.thr = 3))
```

correction flags If `f` is a correction flag, `f$FUN` returns an object of type *arrayCGH* and `f$char` is `NULL`. In the following example, `global.spatial.flag` computes a spatial trend on the array, and corrects the signal log-ratios from this spatial trend:

```
> global.spatial.FUN <- function(arrayCGH, var) {
+     if (!is.null(arrayCGH$arrayValues$Flag))
+         arrayCGH$arrayValues$LogRatio[which(arrayCGH$arrayValues$Flag !=
+             "")] <- NA
+     Trend <- arrayTrend(arrayCGH, var, span = 0.03, degree = 1,
+         iterations = 3)
+     arrayCGH$arrayValues[[var]] <- Trend$arrayValues[[var]] -
+         Trend$arrayValues$Trend
+     arrayCGH
+ }
> global.spatial.flag <- to.flag(global.spatial.FUN, args = alist(var = "LogRatio"))
```

3.1.2 Permanent and temporary flags

We introduce an additional distinction between *permanent* and *temporary* flags in order to deal with the case of spots or clone that are known to be biologically relevant, but that have not to be taken into account for the computation of a scaling normalization coefficient. For example in breast cancer, when the reference DNA comes from a male, we expect a gain of the X chromosome and a loss of the Y chromosome in the tumoral sample, and we do not want log-ratio values for X and Y chromosome to bias the estimation of a scaling normalization coefficient.

Any *flag* object therefore contains an argument called `type`, which defaults to `"perm"` (*permanent*) but can be set to `"temp"` in the case of a temporary flag. In the following example, `chromosome.flag` is a *temporary* flag that identifies clones correcponding to X and Y chromosome:

```
> chromosome.FUN <- function(arrayCGH, var) {
+     var.rep <- arrayCGH$id.rep
+     w <- which(!is.na(match(as.character(arrayCGH$cloneValues[[var]]),
+         c("X", "Y"))))
+     l <- arrayCGH$cloneValues[w, var.rep]
+     which(!is.na(match(arrayCGH$arrayValues[[var.rep]], as.character(l))))
+ }
> chromosome.char <- "X"
> chromosome.label <- "Sexual chromosome"
> chromosome.flag <- to.flag(chromosome.FUN, chromosome.char, type = "temp.flag",
+     args = alist(var = "Chromosome"), label = chromosome.label)
```

3.2 Methods

3.2.1 to.flag

The function `to.flag` is used of the creation of *flag* objects, with the specificities described in subsection 3.1.

```
> args(to.flag)
```

```
function (FUN, char = NULL, args = NULL, type = "perm.flag",
    label = NULL)
NULL
```

3.2.2 flag.arrayCGH

Function `flag.arrayCGH` simply applies function `flag$FUN` to a *flag* object for filtering, and returns:

- a filtered array with field `arrayCGH$arrayValues$Flag` filled with the value of `flag$char` for each spot to be excluded from further analysis in the case of an exclusion flag;
- an array with corrected signal value in the case of a correction flag.

```
> args(flag.arrayCGH)

function (flag, arrayCGH)
NULL
```

3.2.3 flag.summary

Function `flag.summary` computes spot-level information about normalization (including the number of flagged spots and numeric normalization parameters), and displays it in a convenient way. This function can either be applied to an object of type *arrayCGH*:

```
> args(flag.summary.arrayCGH)

function (arrayCGH, flag.list, flag.var = "Flag", nflab = "not flagged",
    ...)
NULL
```

or to plain spot-level information, by using the default method:

```
> args(flag.summary.default)

function (spot.flags, flag.list, nflab = "not flagged", ...)
NULL
```

4 *qscore* class

As we point out in the introduction of this document, evaluating the quality of an array-CGH after normalization is of major importance, since it helps answering the following questions:

- which is the best normalization process ?
- which array is of best quality ?
- what is the quality of a given array ?

To this purpose we define quality scores (*qscores*), which attributes and methods are explianed in the two following subsections.

4.1 Attributes

A *qscore* object `qs` is a list which contains a function (`qs$FUN`), a name (`qs$name`), and optionnally a label (`qs$label`) and arguments to be passed to `qs$FUN` (`qs$args`). In the following example, the quality score `pct.spot.qscore` evaluates the percentage of spots that have passed the filtering steps of normalization; it provides an evaluation of the array quality for a given normalization process. The function `to.qscore` is explained in subsection 4.2.

```
> pct.spot.FUN <- function(arrayCGH, var) {
+     100 * sum(!is.na(arrayCGH$arrayValues[[var]]))/dim(arrayCGH$arrayValues)[1]
+ }
> pct.spot.name <- "SPOT_PCT"
> pct.spot.label <- "Proportion of spots after normalization"
> pct.spot.qscore <- to.qscore(pct.spot.FUN, name = pct.spot.name,
+     args = alist(var = "LogRatioNorm"), label = pct.spot.label)
```

4.2 Methods

4.2.1 to.qscore

The function `to.qscore` is used of the creation of *qscore* objects, with the specificities described in subsection 4.1.

```
> args(to.qscore)

function (FUN, name = NULL, args = NULL, label = NULL, dec = 3)
NULL
```

4.2.2 qscore.arrayCGH

Function `qscore.arrayCGH` simply computes and returns the value of *qscore* for *arrayCGH*:

```
> args(qscore.arrayCGH)

function (qscore, arrayCGH)
NULL
```

4.2.3 qscore.summary.arrayCGH

Function `qscore.summary.arrayCGH` computes all quality scores of a list (using function `qscore.arrayCGH`), and displays the results in a convenient way.

```
> args(qscore.summary.arrayCGH)

function (arrayCGH, qscore.list)
NULL
```

5 Data

We provide examples of array-CGH data coming from two different platforms. These data illustrate the need for appropriate within-array normalization methods, and especially the need for methods that handle spatial effects.

```
> data(spatial)
```

For each array we provide raw data (generated by Genepix or SPOT (5)), as well as the corresponding *arrayCGH* object before and after normalization.

These arrays illustrate the main source of non biological variability of these data sets, namely spatial effects. We classify these effects into two non exclusive types: local bias and global gradients. In the case of *local bias*, entire areas of the array show lower or higher signal values than the rest of the array, with no biological explanation (array `edge`); to our experience, this particular type of artifact roughly affects an array out of two. In the case of *global gradients*, the array shows an obvious signal gradient from one side of the slide to the other (array `gradient`).

5.1 `edge`

Bladder cancer tumors were collected at Henri Mondor Hospital (CrÃl'teil, France) (1) and hybridized on arrays CGH composed of 2464 Bacterian Artificial Chromosomes (F. Radvanyi, D. Pinkel et al., unpublished results); each of these BAC is spotted three times on the array, and the three replicates are neighbors on the array. We give the example of an `arrayCGH` with local spatial effects (figure 1): high log-ratios cluster in the upper-right corner of the array.

5.2 `gradient`

We give the example of two arrays from a breast cancer data set from Institut Curie (O. Delattre, A. Aurias et al., unpublished results). These arrays consist of 3342 clones, organized as a 4×4 superblock that is replicated three times. This data set is affected by the two types of spatial effects: local bias areas (as for the previous data set), and spatial gradients from one side of the array to the other. The array `gradient` illustrates this second type of spatial effect.

6 Graphical representations

As for any type of data analysis, appropriate graphical representations are of major importance for data understanding. Array-CGH data are typically ratios or log-ratios, that correspond to locations on the array (spots) and to locations on the genome (clones). Therefore in the case of array-CGH data normalization, two complementary types of representations are necessary:

- a dotplot of the array, that takes into account the array design. This is a crucial tool in the case of array-CGH data normalization for two reasons: first it provides an easy way to *identify* spatial artifacts such as row, column, print-tip group effects, as well as spatial bias and spatial gradients on the array; then it performs a post-normalization *control*, to ensure that the normalization procedure reached its goals, i.e. significantly reduced the observed effects.

```
> data(spatial)
> arrayPlot(edge, "LogRatio", main = "Local spatial effects", zlim = c(-1,
+     1), mediancenter = TRUE, bar = "h")
```

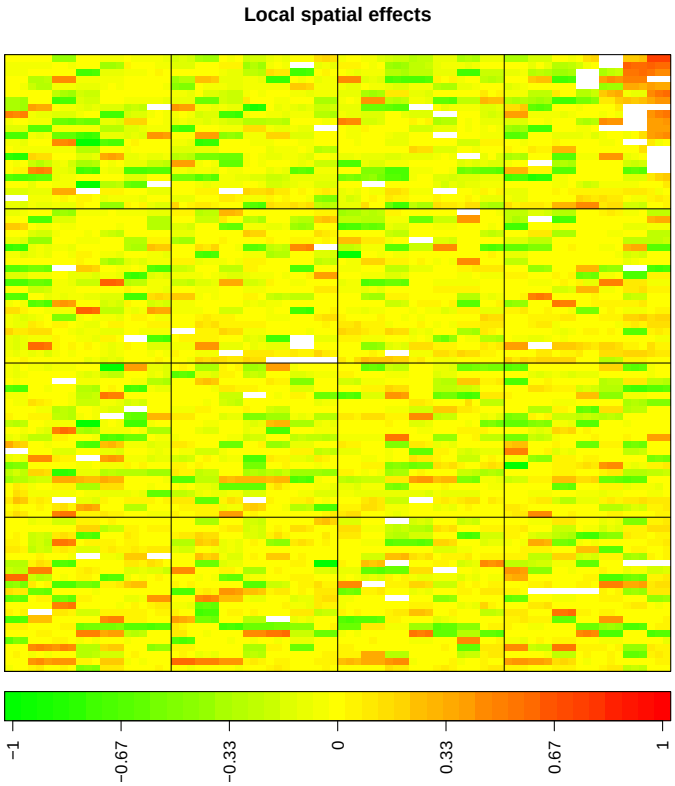


Figure 1: *array with local spatial effects.*

```
> data(spatial)
> arrayPlot(gradient, "LogRatio", main = "Spatial gradient", zlim = c(-2,
+     2), mediancenter = TRUE, bar = "h")
```

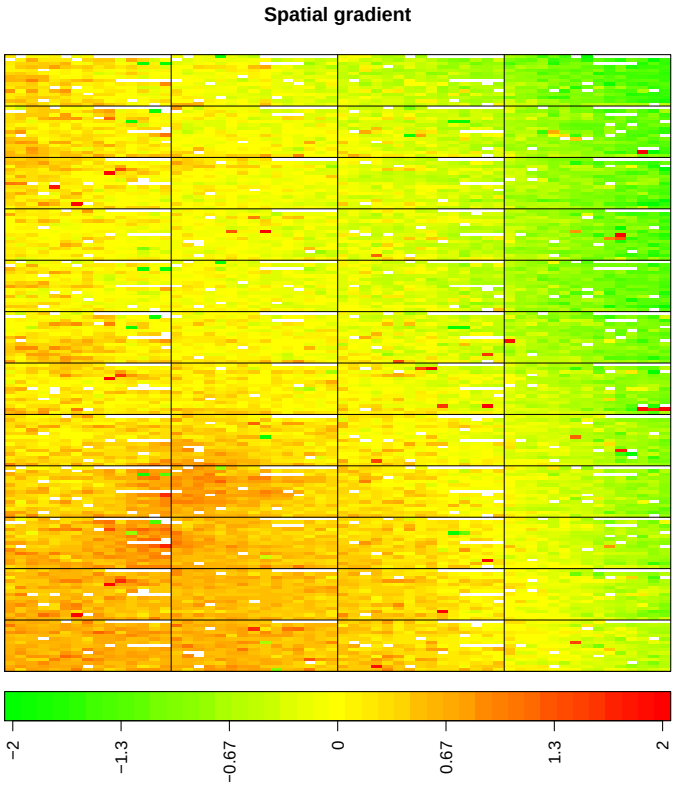


Figure 2: *Example of array with spatial gradient.*

- a plot of the signal values along the genome, which gives a visual impression of the array quality on the edge of biological relevance; comparing the signal shape before and after normalization provides a qualitative idea of the imrpovement in data quality provided by the normalization method.

The `arrayPlot` method provided by the *GLAD* package and based on `maImage` (2) addresses the first point; we add two methods to this toolbox:

- the `genome.plot` method displays a plot of any signal value (e.g. log-ratios) along the genome;

- the `report.plot` method successively calls `arrayPlot` and `genome.plot` in order to provide a simultaneous vision of the data using the two relevant metrics (array and genome), with approproate color scales.

6.1 genome.plot

This method provides a convenient way to plot a given signal along the genome; the signal values can be colored according to their level (which is the default comportment of the function) or to the level of any other variable, in the following way:

- if the variable is numeric (e.g. signal to noise ratio), the function assumes that it is a quantitative variable and adapts a color palette to its values (figure 3)

```
> data(spatial)
> genome.plot(edge.norm, chrLim = "LimitChr", cex = 1)
```

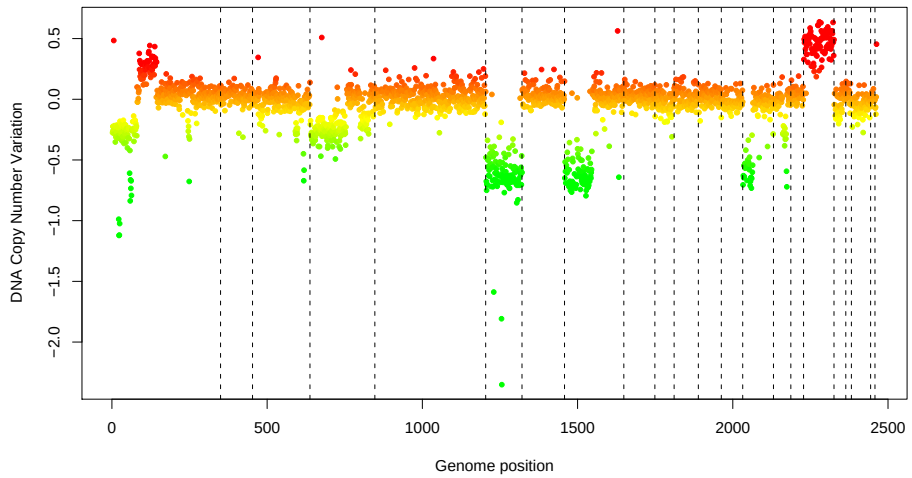


Figure 3: *Pan-genomic profile of the array. Colors are proportional to log-ratio values.*

- if the variable is not numeric (e.g. the copy number variation as estimated by *GLAD*, or a character variable making the disitnction between flagged and un-flagged clones), the

```
> data(spatial)
> edge.norm$cloneValues$ZoneGNL <- as.factor(edge.norm$cloneValues$ZoneGNL)
> genome.plot(edge.norm, col.var = "ZoneGNL", chrLim = "LimitChr",
+     cex = 1)
```

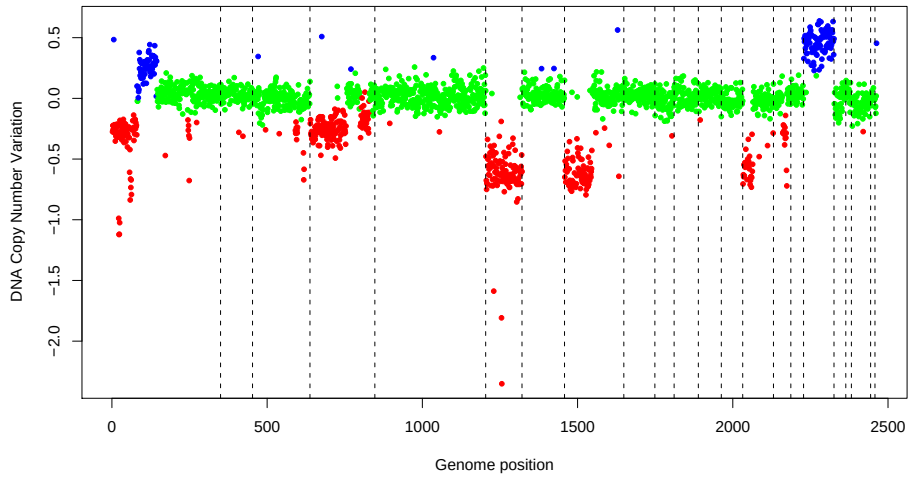


Figure 4: *Pan-genomic profile of the array. Colors correspond to the values of the variable "ZoneGNL".*

function counts the number of modalities of the variable and defines an appropriate color scale using the `rainbow` function (figure 4).

6.2 report.plot

This method successively calls `arrayPlot` and `genome.plot`; it checks for color scale consistency between plots, and can automatically set the plot layout (figure 5).

```
> data(spatial)
> report.plot(edge.norm, chrLim = "LimitChr", zlim = c(-1, 1),
+     cex = 1)
```

Figure 5: *`report.plot`: array image and pan-genomic profile after normalization.*

7 Sample `MANOR` sessions

In this section we illustrate the use of *MANOR* on two CGH arrays. Our examples contain several steps, including data preparation, flag definition, array normalization, quality criteria definition, and quality assessment of the array, and highlights of the normalization process.

7.1 array `edge`

7.1.1 Data preparation: `import`

```
> dir.in <- system.file("data", package = "MANOR")
> spot.names <- c("LogRatio", "RefFore", "RefBack", "DapiFore",
+     "DapiBack", "SpotFlag", "ScaledLogRatio")
> clone.names <- c("PosOrder", "Chromosome")
> edge <- import(paste(dir.in, "/edge.txt", sep = ""), type = "spot",
+     spot.names = spot.names, clone.names = clone.names, add.lines = TRUE)

[1] "number of lines does not match array design: adding empty lines..."
```

7.1.2 Normalization: `norm.arrayCGH`

Figure 6 shows the results of the normalization process.

```
> data(flags)
> data(spatial)
> local.spatial.flag$args <- alist(var = "ScaledLogRatio", by.var = NULL,
+     nk = 5, prop = 0.25, thr = 0.15, beta = 1, family = "gaussian")
> flag.list <- list(spatial = local.spatial.flag, spot = spot.corr.flag,
+     ref.snr = ref.snr.flag, dapi.snr = dapi.snr.flag, rep = rep.flag,
+     unique = unique.flag)
> edge.norm <- norm.arrayCGH(edge, flag.list = flag.list, FUN = median,
+     na.rm = TRUE)

[1] "spatial"
[1] "mean of unbiased zone :  -0.0231566395663957"
[1] "Spatial bias has been detected"
  zone.number           mu effectif effectif.cumul frequency.cumul biased.zone
4           5  0.467833333       66             66      0.00918964           1
3           4  0.045546490     1581           1647      0.22932331           0
5           3  0.004946157     2693           4340      0.60428850           0
1           2 -0.034216274     1868           6208      0.86438318           0
2           1 -0.079646817      974           7182      1.00000000           0
[1] "spot"
[1] "ref.snr"
[1] "dapi.snr"
[1] "rep"
[1] "unique"

> edge.norm <- sort.arrayCGH(edge.norm, position.var = "PosOrder")
```

7.1.3 Quality assessment: `qscore.summary.arrayCGH`

```
> profileCGH <- as.profileCGH(edge.norm$cloneValues)
> profileCGH <- daglad(profileCGH, smoothfunc = "lawsglad", lkern = "Exponential",
+     model = "Gaussian", qlambda = 0.999, bandwidth = 10, base = FALSE,
+     round = 2, lambdabreak = 6, lambdaclusterGen = 20, param = c(d = 6),
+     alpha = 0.001, msize = 5, method = "centroid", nmin = 1,
+     nmax = 8, amplicon = 1, deletion = -5, deltaN = 0.1, forceGL = c(-0.15,
+         0.15), nbsigma = 3, MinBkpWeight = 0.35, verbose = FALSE)

[1] "Smoothing for each Chromosome"
[1] "Optimization of the Breakpoints"
[1] "Check Breakpoints Position"

> edge.norm$cloneValues <- as.data.frame(profileCGH)
> edge.norm$cloneValues$ZoneGNL <- as.factor(edge.norm$cloneValues$ZoneGNL)
```

```
> report.plot(edge.norm, chrLim = "LimitChr", zlim = c(-1, 1),
+     cex = 1)
```

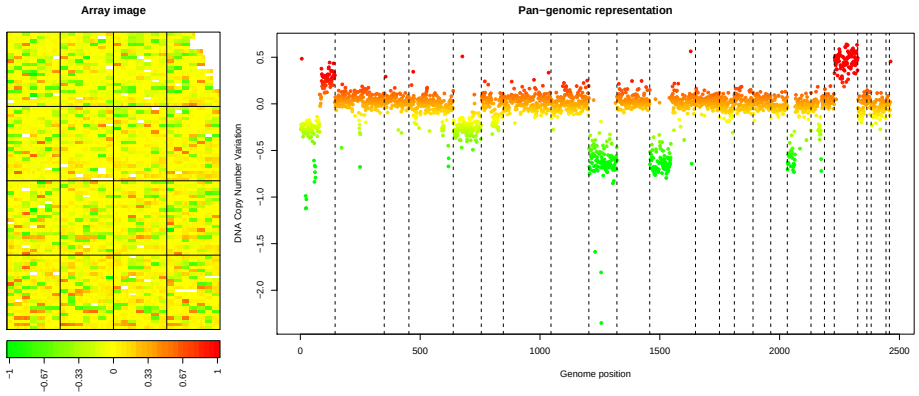


Figure 6: *array 'edge' after normalization.*

```
> data(qscores)
> qscore.list <- list(smoothness = smoothness.qscore, var.replicate = var.replicate.qscore,
+     dynamics = dynamics.qscore)
> edge.norm$quality <- qscore.summary.arrayCGH(edge.norm, qscore.list)
> edge.norm$quality
```

```
              name                                          label score
1 LOCAL_SMOOTHNESS Local signal variability along the genome 0.021
2    VAR_REPLICATE      Average variability among replicates 0.011
3  SIGNAL_DYNAMICS Dynamics of the DNA copy number variation 0.399
```

7.1.4 Highlights of the normalization process: `html.report`

Function `html.report` generates an HTML file with key features of the normalization process: array image and genomic profile before and after normalization, spot-level flag report, and value of the quality criteria.

```
> html.report(edge.norm, dir.out = ".", array.name = "an array with local bias",
+     chrLim = "LimitChr", light = FALSE, pch = 20, zlim = c(-2,
+         2), file.name = "edge")
```

The results of the previous command can be viewed in the file edge.html.

7.2 array gradient

Here we give the example of the normalization of an array with spatial gradient.

7.2.1 Data preparation: `import`

```
> spot.names <- c("Clone", "FLAG", "TEST_B_MEAN", "REF_B_MEAN",
+     "TEST_F_MEAN", "REF_F_MEAN", "ChromosomeArm")
```

```
> clone.names <- c("Clone", "Chromosome", "Position", "Validation")
> ac <- import(paste(dir.in, "/gradient.gpr", sep = ""), type = "gpr",
+     spot.names = spot.names, clone.names = clone.names, sep = "\t",
+     comment.char = "@", add.lines = TRUE)

[1] "number of lines does not match array design: adding empty lines..."
[1] "calculating array design..."

> ac$arrayValues$F1 <- log(ac$arrayValues[["TEST_F_MEAN"]], 2)
> ac$arrayValues$F2 <- log(ac$arrayValues[["REF_F_MEAN"]], 2)
> ac$arrayValues$B1 <- log(ac$arrayValues[["TEST_B_MEAN"]], 2)
> ac$arrayValues$B2 <- log(ac$arrayValues[["REF_B_MEAN"]], 2)
> Ratio <- (ac$arrayValues[["TEST_F_MEAN"]] - ac$arrayValues[["TEST_B_MEAN"]])/(ac$arrayValues
+     ac$arrayValues[["REF_B_MEAN"]])
> Ratio[(Ratio <= 0) | (abs(Ratio) == Inf)] <- NA
> ac$arrayValues$LogRatio <- log(Ratio, 2)
> gradient <- ac
```

7.2.2 Normalization: `norm.arrayCGH`

Figure 7 shows the results of the normalization process.

```
> data(spatial)
> data(flags)
> flag.list <- list(local.spatial = local.spatial.flag, spot = spot.flag,
+     SNR = SNR.flag, global.spatial = global.spatial.flag, val.mark = val.mark.flag,
+     position = position.flag, unique = unique.flag, amplicon = amplicon.flag,
+     chromosome = chromosome.flag, replicate = replicate.flag)
> gradient.norm <- norm.arrayCGH(gradient, flag.list = flag.list,
+     FUN = median, na.rm = TRUE)

[1] "local.spatial"
[1] "mean of unbiased zone :  8.4048170773639"
[1] "There is no spatial bias"
  zone.number       mu effectif effectif.cumul frequency.cumul biased.zone
1           7 8.688599      566            566      0.05641946           0
2           6 8.588816      741           1307      0.13028309           0
3           5 8.485022     1473           2780      0.27711324           0
4           4 8.458262     2436           5216      0.51993620           0
5           3 8.403100     2185           7401      0.73773923           0
6           2 8.347311     2075           9476      0.94457735           0
7           1 8.179531      556          10032      1.00000000           0
[1] "spot"
[1] "SNR"
[1] "global.spatial"
[1] "val.mark"
[1] "position"
```

```
[1] "unique"
[1] "amplicon"
[1] "chromosome"
[1] "replicate"
```

```
> gradient.norm <- sort.arrayCGH(gradient.norm)
```

```
> genome.plot(gradient.norm, chrLim = "LimitChr", cex = 1)
```

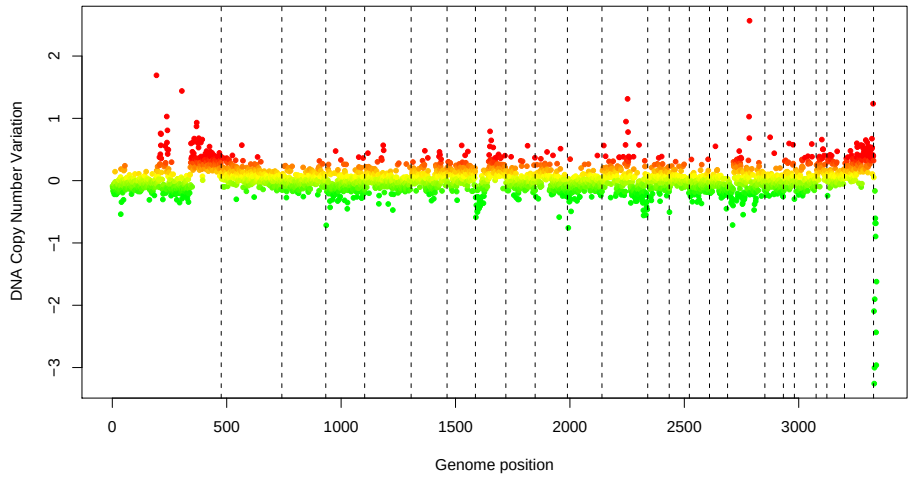


Figure 7: *array* ***gradient*** *after normalization.*

7.2.3 Quality assessment: `qscore.summary.arrayCGH`

```
> profileCGH <- as.profileCGH(gradient.norm$cloneValues)
> profileCGH <- daglad(profileCGH, smoothfunc = "lawsglad", lkern = "Exponential",
+     model = "Gaussian", qlambda = 0.999, bandwidth = 10, base = FALSE,
+     round = 2, lambdabreak = 6, lambdaclusterGen = 20, param = c(d = 6),
+     alpha = 0.001, msize = 5, method = "centroid", nmin = 1,
+     nmax = 8, amplicon = 1, deletion = -5, deltaN = 0.1, forceGL = c(-0.15,
+         0.15), nbsigma = 3, MinBkpWeight = 0.35, verbose = FALSE)
```

```
[1] "Smoothing for each Chromosome"
[1] "Optimization of the Breakpoints"
[1] "Check Breakpoints Position"
```

```
> gradient.norm$cloneValues <- as.data.frame(profileCGH)
> gradient.norm$cloneValues$ZoneGNL <- as.factor(gradient.norm$cloneValues$ZoneGNL)
> data(qscores)
> qscore.list <- list(smoothness = smoothness.qscore, var.replicate = var.replicate.qscore,
```

```
+     dynamics = dynamics.qscore)
> gradient.norm$quality <- qscore.summary.arrayCGH(gradient.norm,
+     qscore.list)
> gradient.norm$quality
```

```
             name                                      label score
1 LOCAL_SMOOTHNESS Local signal variability along the genome 0.032
2    VAR_REPLICATE      Average variability among replicates 0.050
3  SIGNAL_DYNAMICS Dynamics of the DNA copy number variation 0.294
```

7.2.4 Highlights of the normalization process: `html.report`

Function `html.report` generates an HTML file with key features of the normalization process: array image and genomic profile before and after normalization, spot-level flag report, and value of the quality criteria.

```
> html.report(gradient.norm, dir.out = ".", array.name = "an array with spatial gradient",
+     chrLim = "LimitChr", light = FALSE, pch = 20, zlim = c(-2,
+         2), file.name = "gradient")
```

The results of the previous command can be viewed in the file gradient.html.

8 Session information

The version number of R and packages loaded for generating this document are:

```
> sessionInfo()
```

```
R version 2.7.1 (2008-06-23)
i386-apple-darwin8.10.1

locale:
fr_FR.UTF-8/fr_FR.UTF-8/C/C/fr_FR.UTF-8/fr_FR.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] MANOR_1.13.1 GLAD_1.16.0

loaded via a namespace (and not attached):
[1] tools_2.7.1
```

9 Supplementary data

The package *MANOR* provides sample gpr and spot files, as examples to the `import` funciton. However, due to space limitations, only the first 100 lines these file are provided in the current distribution of *MANOR*. The full files can be downloaded from here:

- 'gpr' file: gradient.gpr
- 'spot' file: edge.txt

References

[1] C. Billerey, D. Chopin, M. H. Aubriot-Lorton, D. Ricol, S. Gil Diez de Medina, B. Van Rhijn, M. P. Bralet, M. A. Lefrere-Belda, J. B. Lahaye, C. C. Abbou, J. Bonaventure, E. S. Zafrani, T. van der Kwast, J. P. Thiery, and F. Radvanyi. Frequent FGFR3 mutations in papillary non-invasive bladder (pTa) tumors. *Am. J. Pathol.*, 158:955–1959, 2001.

[2] S. Dudoit and Y. H. Yang. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York, 2003.

[3] P. Hupé, N. Stransky, J-P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratios to gain and loss of DNA regions. *Bioinformatics*, 20:3413 – 3422, 2004.

[4] A. S. Ishkanian, C. A. Malloff, S. K. Watson, R. J. DeLeeuw, B. Chi, B. P. Coe, A. Snijders, D. G. Albertson, D. Pinkel, M. A. Marra, V. Ling, C. MacAulay, and W. L. Lam. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, 36:299–303, 2004.

[5] A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. *Genome Res.*, 12:325–332, 2002.

[6] P. Neuvial, P. Hupé, I. Brito, S. Liva, E. Manié, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7(1):264, May 2006.

[7] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, 1998.

[8] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law S, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29:263–4, 2001.

[9] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20:399–407, 1997.

APPENDIX B

Correlating DNA copy number and expression microarrays

poster presented at ISMB 2007

GTCA: Genome Transcriptome Correlation Analysis within R

Pierre Neuvial[1], Pierre Gestraud[1], Carlo Lucchesi[2], Emmanuel Barillot[1]

[1] Institut Curie, Service Bioinformatique, and [2] Institut Curie, INSERM U 830
26 rue d'Ulm, 75248 Paris Cedex 05, France
gtca@curie.fr

Cancer cells are characterised by modifications in the expression of their genes, and DNA copy number changes are one of the key genetic mechanisms responsible for such modifications. A few recent studies have characterised the overall influence of DNA copy number changes on gene expression (gene dosage effect), using parallel, high-throughput microarray measurements. We developed GTCA, an R package that implements a statistically sound methodology for Genome Transcriptome Correlation Analysis, including data pre-processing, statistical analysis, visualisation and biological interpretation. We illustrate the use of GTCA on two recent data sets (breast cancer [2], and bladder cancer [9]).

1 Extraction of biological information

The input of GTCA consists of normalised expression and copy number data, as well as the corresponding probe annotations (including start and end positions), and optionally sample annotations. We assume that smoothed DNA copy numbers (`smt`) and loss/gain/amplification calls are available (bladder and breast cancer array-CGH profiles were analysed using GLAD [6]).

Inferring missing copy numbers

We take advantage of the consistency of the copy number signal along the genome to infer missing copy number values for each genome array. Based on smoothed copy numbers and calls of the two nearest genome probes, we define the inferred `smt` as:

- a weighted mean of the two smoothed copy numbers if calls are equal
- the smoothed copy number of the most altered call if calls are different

This step ensures that all correlations are based on the *same number of observations*. Therefore there is an unique bijective transformation from correlation coefficients to p-values, which may be used in graphical representations (see vertical axes of figure 2).

Probe matching

For most array technologies genome probes and transcriptome probes have different genome locations, so that an objective probe matching algorithm is needed. Each transcriptome probe is classified into one of the 3 following situations (only based on annotation data); all genome probe matches of the class are then reported in order not to lose biological information.

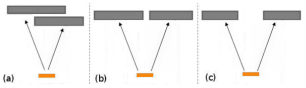


Fig. 1. *Probe matching algorithm. Gray boxes denote genome probes, orange boxes denote transcriptome probes.*

$[a]$ transcriptome probe **IN** one or more genome probes

$[b \setminus a]$ transcriptome probe **INTERSECTS** one or more genome probes

$[c \setminus (a \cup b)]$ transcriptome probe **BETWEEN** genome probes

Reducing complexity

Our probe matching algorithm allows multiple genome probe matches for each transcriptome probe to avoid loss of biological information, at the price of a certain redundancy. We therefore aggregate genome probe matches corresponding to the same transcriptome probe if they have the same smoothed copy number profiles across samples. This results in a *smaller number of correlations* to be calculated, and a *less severe multiple testing problem*. The following table illustrates the result of these first three steps.

data set	sample size	transcriptome probes	genome probes	missing `smt`	IN	INTERSECTS	BETWEEN	probe matches	distinct matches	complexity reduction
breast [2]	89	21881	1836	7%	7%	5%	88%	40667	26330	35%
bladder [9]	57	8111	2215	6%	8%	7%	85%	14884	9194	38%

2 Correlation analysis

Two standard and complementary measures of correlation

r : the Pearson correlation coefficient measures the extent to which the association between copy number and expression is linear; its sensitivity to outliers makes it suitable to detect associations within regions that are amplified in only few samples;

ρ : the Spearman correlation coefficient is the Pearson coefficient between measurements ranks; it is therefore robust to outliers and able to detect non-linear associations.

Correlation coefficients are calculated using smoothed copy number data (`smt`).

Two multiple hypothesis testing adjustment procedures

FDR: we use the Benjamini-Hochberg procedure [1] to control the False Discovery Rate, that is, the expected proportion of false positives among those loci selected.

FWER: we use the Holm procedure [4] to control the Family-Wise Error Rate, that is, the probability that one or more loci among those selected is a false positive. This procedure is more powerful that the traditional Bonferroni procedure.

Both adjustment procedures can be applied genome-wide or chromosome by chromosome, which is useful when only few chromosomes are expected to show a significant gene dosage effect.

3 Results visualisation

Chromosome plots: correlations along chromosomes

Figure 2 illustrates the strong correlation between copy number and expression of E2F3, which is a known oncogene in bladder tumours [7]. FWER or FDR-adjusted significance levels can be added on the right-side axis, for easy interpretation of correlation coefficients.

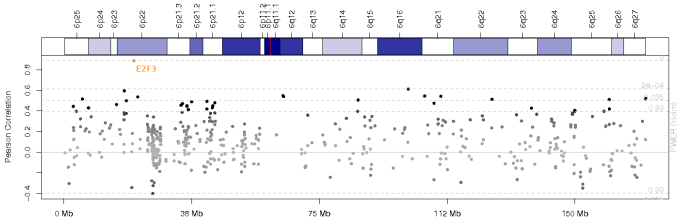


Fig. 2. *Pearson correlation coefficients along chromosome 6 (bladder cancer data). The right-side axis displays FWER-adjusted significance levels. Gene E2F3 is marked in orange.*

Chromosome plots: adjusted p-values along chromosomes

Displayed adjusted significances are negative-log-scaled for ease of visualisation and interpretation; corresponding adjusted significance levels are displayed on the right-side axis, as illustrated figure 3 for FDR adjustment. The same type of visualisation is also available for FWER adjustment.

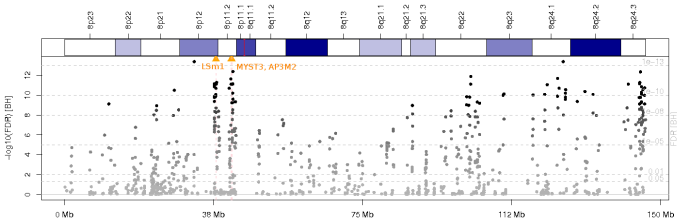


Fig. 3. *FDR-adjusted p-values of Pearson correlation coefficients along chromosome 8 (breast cancer data). Orange triangles indicate three highly significant genes ($p < 2.2\ 10^{-16}$). The right-side axis displays adjusted significance levels in linear scale.*

Gene plots: DNA copy numbers versus expression levels

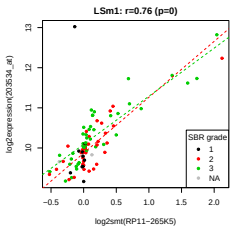


Fig. 4. *Gene plot of LSm1 (breast cancer data).*

Correlations between copy number and expression level for a particular gene of interest can be further investigated at the sample level, using a simple scatter plot of DNA copy numbers versus expression levels. Samples can be coloured according to available annotation.

On the left, the gene plot of LSm1 (a known oncogene in breast cancers [10] identified in figure 3) suggests that gene dosage effect differs between SRB grades, which is confirmed by analysis of variance ($p = 0.01$).

4 Implementation and availability

GTCA is implemented as an R package, which is freely available for academics, and will shortly be submitted to Bioconductor [3]. It has been interfaced as a plug-in for VAMP [8]. This plug-in will be available in the next release of the web platform ACTuDB [5].

References

[1] Y. Benjamini and Y. Hochberg. 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, 1995.

[2] K. Chin, S. DeVries, J. Fridlyand, P.T. Spellman, R. Roydasgupta, W.L. Kuo, A. Lapuk, R.M. Neve, Z. Qian, T. Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

[3] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[4] S Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[5] P. Hupé, P. La Rosa, S. Liva, S. Lair, N. Servant, and E. Barillot. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene*, 2007.

[6] P. Hupé, N. Stransky, J.-P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, Dec 2004.

[7] C.D. Hurst, H. Fiegler, P. Carr, S. Williams, N.P. Carter, and M.A. Knowles. High-resolution analysis of genomic copy number alterations in bladder cancer by microarray-based comparative genomic hybridization. *Oncogene*, 23:2250–2263, 2004.

[8] P. La Rosa, E. Viara, P. Hupé, G. Pierron, S. Liva, P. Neuvial, I. Brito, S. Lair, N. Servant, N. Robine, E Manié, C. Brennetot, I. Janoueix-Lerosey, V. Raynal, N. Gruel, C. Rouveirol, N. Stransky, M-H. Stern, O. Delattre, A. Aurias, F. François Radvanyi, and E. Barillot. VAMP: visualization and analysis of array-CGH, transcriptome and other molecular profiles. *Bioinformatics*, 22(17):2066–2073, Sep 2006.

[9] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S.G.D. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C Spraggon, A Graham, J Southgate, B Asselain, Y Allory, C C Abbou, D G Albertson, J-P Thiery, D K Chopin, D Pinkel, and F Radvanyi. Regional copy number–independent deregulation of transcription in cancer. *Nature Genetics*, 38:1386–1396, 2006.

[10] KL Streicher, ZQ Yang, S. Draghici, and SP Ethier. Transforming function of the LSM1 oncogene in human breast cancers with the 8p11–12 amplicon. *Oncogene*, 2006.

APPENDIX C

Visualization and analysis of molecular profiles

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 17 2006, pages 2066–2073
doi:10.1093/bioinformatics/btl359

Genome analysis

VAMP: Visualization and analysis of array-CGH, transcriptome and other molecular profiles

Philippe La Rosa[1,*], Eric Viara[1], Philippe Hupé[1,2], Gaëlle Pierron[3], Stéphane Liva[1], Pierre Neuvial[1], Isabel Brito[1], Séverine Lair[1], Nicolas Servant[1], Nicolas Robine[1,4], Elodie Manié[3], Caroline Brennetot[3], Isabelle Janoueix-Lerosey[3], Virginie Raynal[3], Nadège Gruel[3], Céline Rouveirol[2], Nicolas Stransky[2], Marc-Henri Stern[3], Olivier Delattre[3], Alain Aurias[3], François Radvanyi[2] and Emmanuel Barillot[1]

[1]Institut Curie, Service Bioinformatique, 26 rue d'Ulm, Paris, 75248 cedex 05, France, [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, Paris, 75248 cedex 05, France, [3]Institut Curie, INSERM U509, 26 rue d'Ulm, Paris, 75248 cedex 05, France and [4]Institut Curie, CNRS, Université Pierre et Marie Curie UMR 7147, 26 rue d'Ulm, Paris, 75248 cedex 05, France

Received on January 11, 2006; revised on May 31, 2006; accepted on June 25, 2006
Advance Access publication July 4, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: Microarray-based CGH (Comparative Genomic Hybridization), transcriptome arrays and other large-scale genomic technologies are now routinely used to generate a vast amount of genomic profiles. Exploratory analysis of this data is crucial in helping to understand the data and to help form biological hypotheses. This step requires visualization of the data in a meaningful way to visualize the results and to perform first level analyses.

Results: We have developed a graphical user interface for visualization and first level analysis of molecular profiles. It is currently in use at the Institut Curie for cancer research projects involving CGH arrays, transcriptome arrays, SNP (single nucleotide polymorphism) arrays, loss of heterozygosity results (LOH), and Chromatin ImmunoPrecipitation arrays (ChIP chips). The interface offers the possibility of studying these different types of information in a consistent way. Several views are proposed, such as the classical CGH karyotype view or genome-wide multi-tumor comparison. Many functionalities for analyzing CGH data are provided by the interface, including looking for recurrent regions of alterations, confrontation to transcriptome data or clinical information, and clustering. Our tool consists of PHP scripts and of an applet written in Java. It can be run on public datasets at http://bioinfo.curie.fr/vamp

Availability: The VAMP software (Visualization and Analysis of array-CGH,transcriptome and other Molecular Profiles) is available upon request. It can be tested on public datasets at http://bioinfo.curie.fr/vamp. The documentation is available at http://bioinfo.curie.fr/vamp/doc

Contact: vamp@curie.fr

1 INTRODUCTION

Array Comparative Genome Hybridization (array-CGH) is a recently developed technology based on DNA microarrays (Pinkel *et al*., 1998; Snijders *et al*., 2001; Solinas-Toldo *et al*., 1997; Ishkanian *et al*., 2004) that can be used to investigate DNA copy number differences between two samples. A CGH array-generally consists of spotted clones of genomic sequences (e.g. bacterial artificial chromosomes) that cover part or all of the genome. Both DNA samples are labeled with distinct fluorescent dyes and undergo competitive hybridization onto the CGH array. The array is then scanned with a scanner or a CCD camera, and the acquired image is analyzed (gridding, spot addressing, spot segmentation, spot quantification, outlier detection), normalized (to remove as much as possible any systematic spatial or intensity biases, e.g. Neuvial *et al*., (2005), duplicate statistical analysis is then carried out (each clone is generally spotted in several copies), and adequate statistical algorithms detect any loss or gain regions (Hupé *et al*., 2004; Olshen *et al*., 2004; Fridlyand *et al*., 2004; Jong *et al*., 2003; Picard *et al*., 2005; Eilers and de Menezes, 2005; Bilke *et al*., 2005). CGH arrays are often used in cancer research because chromosome aberrations are thought to be causal in tumor progression (Albertson *et al*., 2003; Pinkel and Albertson, 2005). Here, normal DNA is used as reference and the test sample would be tumoral biopsy DNA. The normal sample has two copies of each genomic region, whereas tumor DNA may show losses or gains in certain DNA regions. Measurement of the signal intensities of the reference and tumor samples for each clone makes it possible to determine the lost or gained regions in the tumor sample. Further analyses can include the determination of recurrent loss or gain of DNA regions, clustering of samples and determination of candidate oncogenes and candidate tumor suppressor genes within the altered regions (based on their annotations or on their transcription level). It is also possible to link array-CGH results to the clinical phenotype or to biological parameters through, for example, supervised classification or correlation analysis. The visualization of the data is a crucial step in the analysis procedure and is essential for hypothesis formulation and model-free reasoning. We have developed, in the framework of large-scale array-CGH projects, a graphical user interface that allows several visualization modes of the CGH profiles and offers several data analysis tools. The software also displays a large variety of genomic profiles, such as transcriptome,

*To whom correspondence should be addressed.

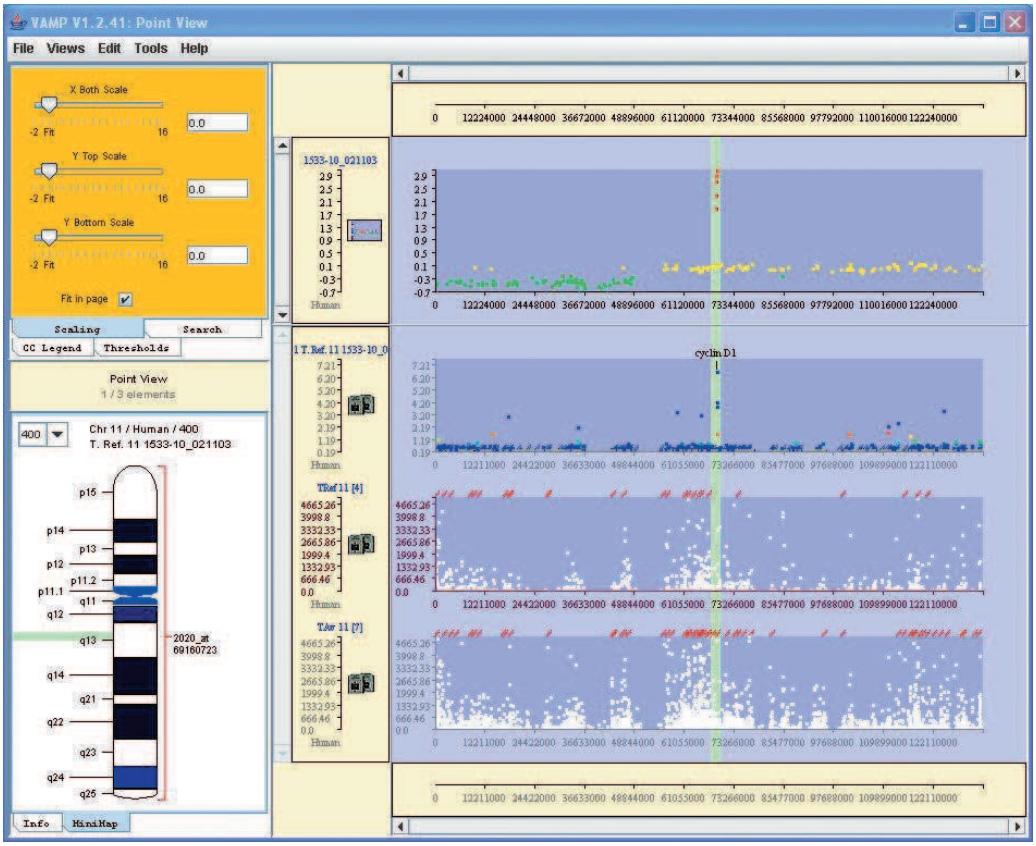


Fig. 1. Array-CGH (top profile) versus transcriptome ratio (second profile in descending order), computed for Affymetrix U95 array of a bladder tumor sample and of a reference sample. This confrontation pinpoints the probable implication of the oncogene cyclin D1 in this tumor. The third and fourth profiles in descending order correspond to a reference profile (average normal bladder tissue profile) and the profile of the tumor under study, respectively. The second profile is the ratio of the fourth to the reference profile.

Loss Of Heterozygosity (LOH), Vogelstein *et al*. (1989), Single nucleotide polymorphism (SNP) arrays (Bignell *et al*., 2004; Huang *et al*., 2004) and ChIP chip [Chromatin ImmunoPrecipitation coupled with microarrays, Buck and Lieb (2004)] profiles and allows addition of new tools for data treatment or analysis. We have called the software VAMP for 'Visualization and Analysis of Molecular Profiles'. In this article we first detail how data are visually presented in VAMP, and then we explain how the user interacts with the software and which functionalities are offered for data analysis. Finally, we describe the software architecture of VAMP.

2 RESULTS

2.1 Data representation

VAMP was designed to graphically represent any genomic profile along the genome axis. We started the development of VAMP for array-CGH data, but we have extended it to accept, on the same window, any kind of profile. We currently use the software for expression arrays, SNP arrays, LOH results and ChIP chip profiling, in addition to array-CGH. VAMP is currently used for three species (human, mouse and yeast) but the addition of a new species is straightforward. It is possible to visualize simultaneously, on the same window, different types of profiles for a given species, e.g. array-CGH and mRNA expression profiles of a tumor (Fig. 1). All profiles in a window are drawn on the *x*-axis with the same scale (the genome sequence), which allows an easy comparison of profiles.

A typical VAMP window is divided into three areas (Fig. 2): the main frame consists of the graphical display of the profiles; the top left frame controls zoom, search and drawing options; the bottom left frame offers the choice between textual information (Fig. 3) on the object under the mouse pointer, or context information, called MiniMap (Fig. 2).

2.1.1 Main frame VAMP currently offers several types of visualization that can be displayed in the main frame: (1) List View, (2) Profile View (Fig. 2) (3) Karyotype View (Fig. 3), (4) Dot Plot View (Fig. 4). These views all allow simultaneous visualization of several profiles (the only limitation is the memory size of the computer running VAMP, or more precisely, the memory allocated to the Java virtual machine: for example with an 800 Mb Java virtual machine memory, 700 microarrays (each with 3500 probes) can be loaded simultaneously).

- List View: the List View lists the names of all the arrays currently loaded and can be used for selecting or keeping track of the data under study.

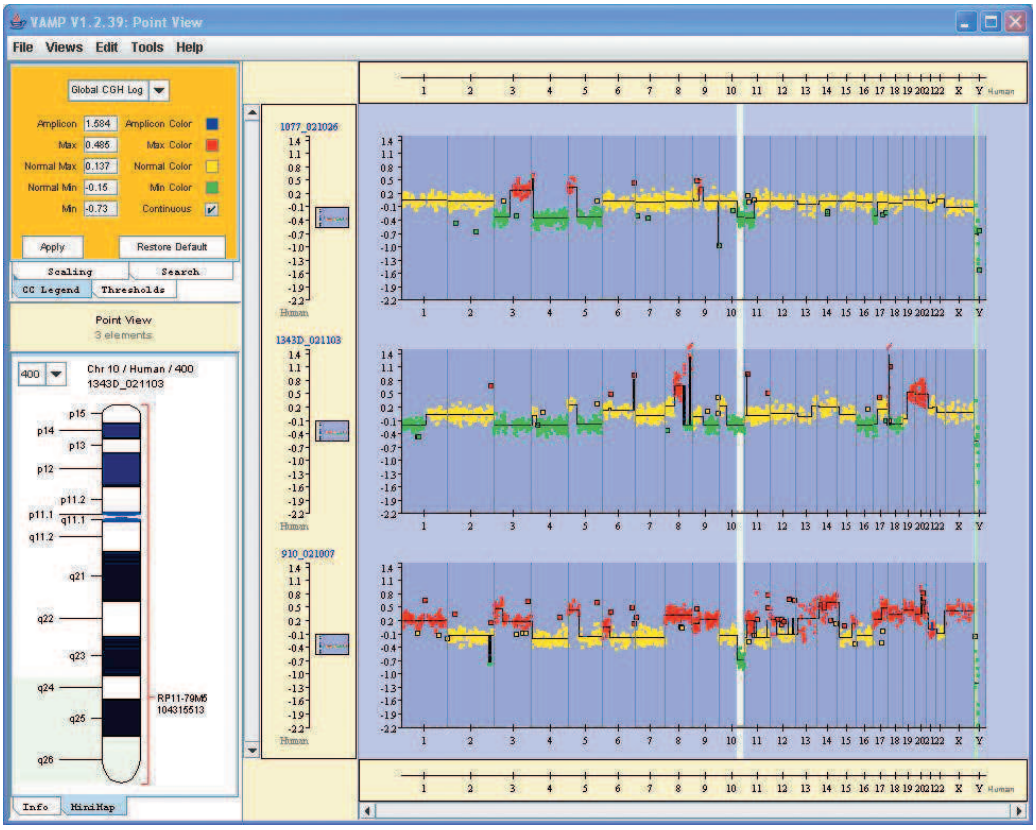


Fig. 2. Genomic View, main frame: profiles along all the concatenated chromosomes; top left: zoom control, search and drawing options; bottom left: textual information on the object under the mouse pointer or (in this figure) chromosome context information (MiniMap). The regions spanning the three tumors highlighted in green are those that are lost in all tumors (short arm of chromosome 10, and Y chromosome); these are called minimal regions.

- Profile View: the Profile View (Fig. 2) can display the profiles as points, barplots or curves. It can be split into two frames, as in Figure 1. The upper frame can, for example, contain a profile for reference when browsing a collection of profiles in the lower frame. The two frames have separate control of Y-scale and Y-scrolling, but have the same X-scale and X-scrolling. The Profile View can also display symbols for chromosome telomeres and centromeres, and can show the results of CGH ratio statistical analysis (e.g. breakpoints, or smoothed signal values, see Fig. 2).
- Karyotype View: the Karyotype View (Fig. 3) displays profiles having the well-known classical CGH rendering: vertical representations of chromosomes with cytogenetic banding and contiguous representation of sample profiles.
- Dot Plot View: the Dot Plot View does not consider the microarray probe positions on the genome, but only their ranks. It displays a collection of samples as a heat map based on the level of signal for each probe (Fig. 4).

By default, points or barplots are colored according to the signal intensity (generally using ratios of the two channels or log-ratios) using a continuous scale from red to yellow to green. All the previously mentioned views for the CGH data can be colored as a function of the array-CGH data analysis. Typically, gained DNA regions are displayed in red, lost regions in green, amplicons in blue and normal in yellow.

Whatever view is chosen, the profiles can be represented in Genomic mode or Chromosome mode. The Genomic mode simply depicts the profiles along all the concatenated chromosomes. It is the most usual representation, and allows comparison of profiles from different samples or comparison of different types of profiles from a given sample. The Chromosome mode is similar to the Genomic mode except that it only displays one particular chromosome. It is also possible to merge several chromosomes and to represent those chromosomes useful for the study.

- New Views: our object-oriented architecture easily allows us to add new types of views that can be associated with particular actions or data processing. For example, the Minimal Region functionality is associated with a particular type of view. Therefore, when profiles are pasted in the window, the Minimal Region View automatically displays the array-CGH profiles with the DNA regions recurrently lost or gained in the samples (Fig. 2).

2.1.2 Top left frame This frame controls zoom, search and drawing options. Zooming is independent on *X* and *Y* axes, and all profiles in the same window have the same zoom control, except

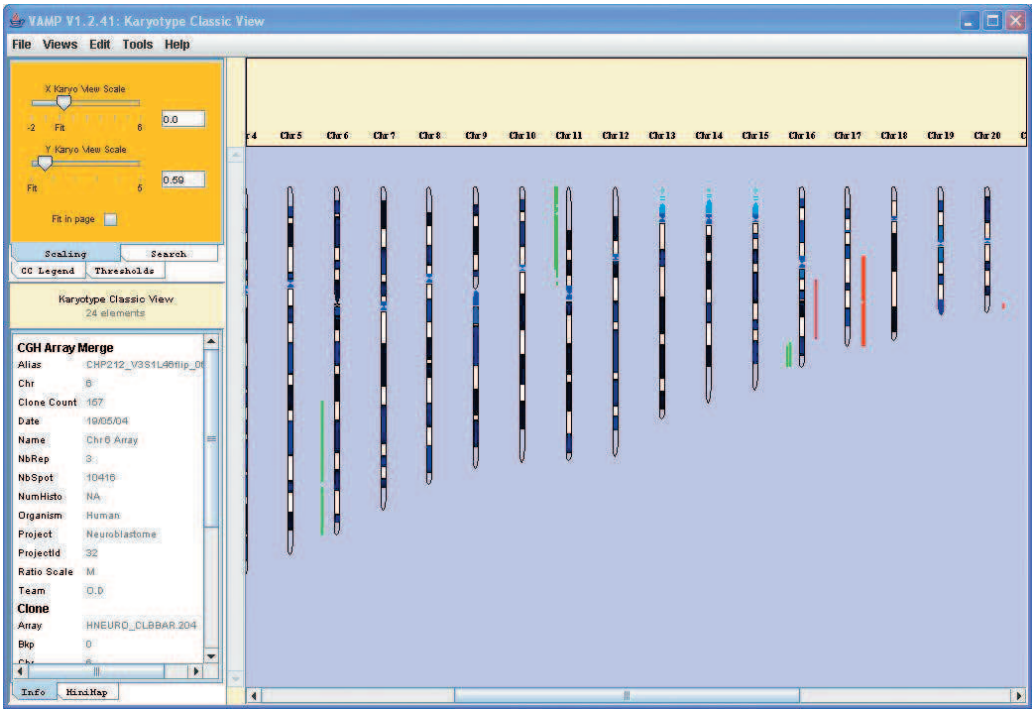


Fig. 3. Karyotype View, classic rendering of CGH data, loss regions in green, gain in red.

for *Y* zooming of the reference profile. The search can be carried out on any property attached to the arrays or the clones/probes held in an XML (eXtended Markup Language) data file or in the database (see Fig. 6 and the Software architecture presentation below). For XML data files, the list of properties is not limited,but is established at run time, leading to a very flexible search option. Drawing options include color-coding for signal values, and the threshold values to be applied; they can be either global to the application or restricted to one profile (local). User preferences can be saved on your computer in a XML configuration file.

2.1.3 Bottom left frame (Object information and context frame) The bottom left frame can either display textual information on the object under the mouse pointer (Fig. 3) or context information, called MiniMap (Fig. 2). The textual information consists of mandatory fields (object genomic position, signal value, project name, organism and data type) and any other type of complementary information stored in the XML data file. For example, in array-CGH profiles we currently display general information about the clone under the mouse pointer (name, chromosome, number of valid replicates, rank and position on the sequence, signal ratio and standard deviation, size of the clone, CGH status—gain/lost/normal) as well as information about the array (name, number of spots, number of clones, number of replicates, chromosomes covered, ratios or log-ratios) and information about the sample (sample id, project name, date). MiniMap is a special view type that gives some context on what the user is examining in the main frame: (1) a cytogenetic representation of the chromosome under the mouse pointing, with (2) a rule delimiting the region of the chromosome displayed on the main frame and (3) the name and position of the object (array-CGH clone, transcriptome microarray probe, etc.) under the mouse pointer. In this view, the display can be automatically updated when the user moves the mouse.

2.2 User interaction

All user actions are accessible either through a Menu on the menubar, or through pointing to or clicking objects. When using VAMP, the session can be saved in local XML files. Reloading the file later on allows the continuation of the analysis within the context of the previous work, or allows the exchange of results and data with colleagues. All user preferences can also be stored in local XML files. Drag and drop capability is offered for any profile, from one window to any other window, the rendering being automatically adapted (e.g. from a dot plot view to a karyotype view). An advanced printing function is offered, either in visible mode (only the profiles that are visible on the screen are printed), or in global mode (all profiles in the view are printed). A template is offered for defining the output of the printing (this can, for example, include several frames in an arbitrary composition, to which text or images can be added). It can be used for defining and printing standardized outputs. The user can also interactively monitor the print preferences.

2.3 Data analyses

VAMP allows addition of any new piece of software for data analysis and visualization of the results. Several functionalities have already been implemented either as plug-ins or within the VAMP Java source code. VAMP was initially developed for the analysis of CGH-arrays of tumoral samples. As VAMP is actually an interface, it is assumed that the microarray data have already been normalized, and also, for CGH data, that breakpoints have been established and

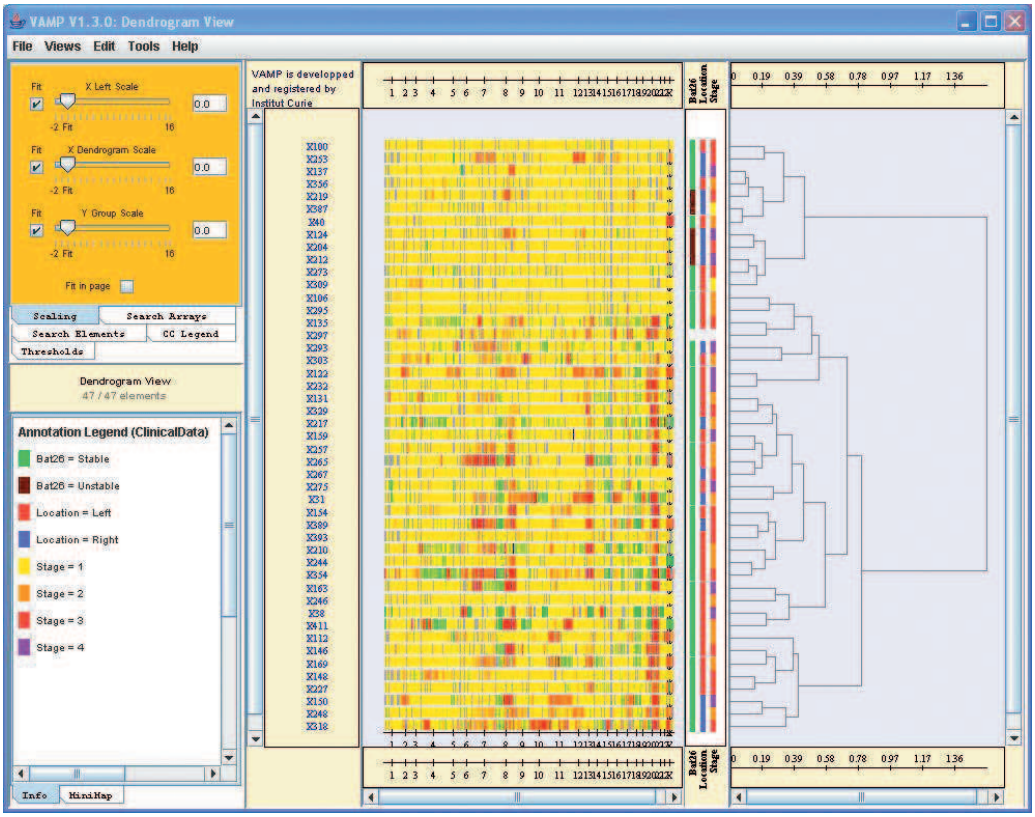


Fig. 4. VAMP interface, dotplot view of array-CGH profiles (middle panel), and dendrogram resulting from a hierarchical clustering (right panel). In between, color-coded clinical information about the samples, with a legend (bottom left). Data from Nakao *et al.* (2004).

regions of DNA loss or gain inferred. VAMP can then display in the profile frame (Fig. 2) the breakpoint positions, the status of each region (by default, green for loss, yellow for normal, red for gain, blue for amplicons), and the estimation of the signal value in each region, which is computed, for example, using smoothing techniques (Hupé *et al.*, 2004). VAMP also allows the defining of the gain and loss regions by simply applying a threshold to the signal ratios. Examples of data analyses available within VAMP are given below and are described in more detail in the software documentation (http://bioinfo.curie.fr/vamp/doc).

Finding common alterations among a collection of CGH- array profiles. CGH array analysis principally consists in finding common regions of alterations, i.e. regions that are lost in many tumors. It is essential in these studies to distinguish between recurrent and random alterations. Recurrent alterations pinpoint regions involved in tumoral progression, whereas random alterations are simply the consequence of the general instability that affects the genome of a tumor. Among the recurrent alterations we distinguish the minimal regions and the recurrent regions. Minimal regions are extracted by intersecting the profiles of many tumors and looking for a sufficient number of alterations in the tumors (this parameter is set by the user) over the smallest possible region of the profile (Fig. 2). Tumoral progression obeys a selection principle, and it would be expected that the genes that need to be altered for a cell to become tumoral must be located in the smallest possible intersection of all alterations of a region. Recurrent regions are defined differently: in a given tumor, an alteration is bounded by two extremities, which can be a breakpoint or a chromosome end; when a sufficient number of tumors have the same extremities, these extremities define a recurrent region. We have implemented a linear algorithm that detects such minimal and recurrent regions, which is described in (Rouveirol *et al.*, 2006). Gained regions appear in red in the main frame, and lost regions appear in green (Fig. 2). Amplicons (defined as gained regions with signal-ratio above a threshold typically equal to two) are colored in blue. The tumors that support a region of alteration may be optionally shadowed in the region, and for each region the user can sort these tumors.

Clustering profiles. Clustering is a general technique for unsupervised data classification widely used in microarray data analysis. A VAMP function offers the possibility to perform a hierarchical clustering (Kaufman and Rousseuw, 1990) on the profiles in the dot plot view. This can cluster genes and tumors from transcriptome arrays, or tumors from a CGH profile. In a CGH profile, the clustering uses the smoothed values of the CGH profile as variables and the Euclidean distance and Ward method for group distance computation. VAMP displays the results as a cluster view including a heat map and the trees resulting from the clustering algorithm (Fig. 4).

Comparing profiles. The Menu proposes several different data manipulation procedures for the profiles such as loading any type of

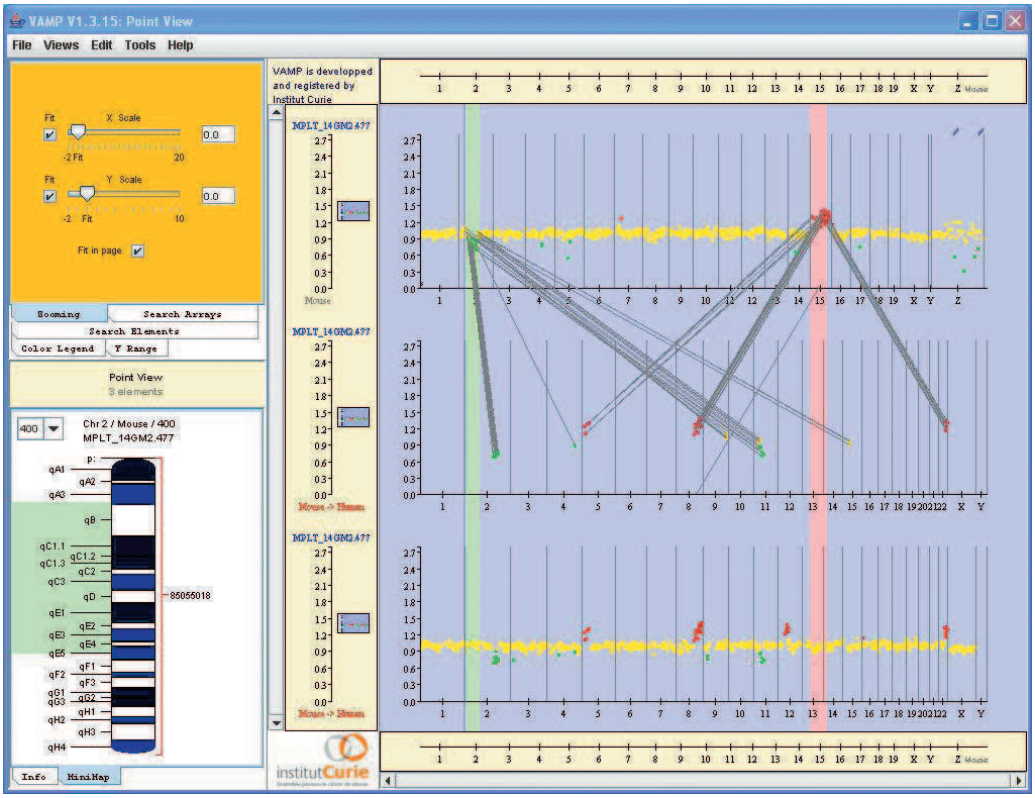


Fig. 5. Array-CGH profile for a mouse tumor (top) and its syntenic projection, i.e. a humanized array-CGH profile after mapping each mouse clone onto the human genome (bottom) and projection for two regions (middle profile) with resulting synteny relationships. Mapping is done from each clone of the mouse profile onto the location of the most similar sequence of the human genome. Mouse clones with ambiguous syntenic locations have not been mapped onto the human genome.

profile (CGH, expression, LOH, ChIP chip—an icon at the left of each profile shows the type of loaded profile) for a given sample (e.g. a typical application of VAMP is the simultaneous visualization of the DNA alterations and gene under- and over-expression in a region, Fig. 1); defining a profile as a reference and calculating the ratio of a profile to the reference (useful for one-color microarrays such as Affymetrix); averaging profiles; drawing marks (vertical bars) or regions (such as the green regions in Fig. 2) across all profiles (and simultaneously on the MiniMap); and many others.

Confrontation with sample annotation. Clinical data, or any other sample annotations, present in the XML files can be used for filtering tumors or for sorting them. This data can be visualized as color-coded bars in an annotation frame on the left of the profiles, and can be easily compared with a clustering result (Fig. 4).

Synteny analysis. VAMP can display the syntenic projection of a profile onto the genome of another species, in which that genome serves as a reference; a typical application is the projection of a mouse array-CGH profile onto the human genome (Fig. 5). In our case if an unambiguous syntenic locus was found, the mapping was done from each clone of the mouse profile onto the location of the most similar sequence of the human genome. The synteny relationships can be shown, for a selection of regions of the genome, as links from each clone of the profile to the location of the most similar sequence of the reference genome.

Other functions. The right mouse button brings up a menu with several actions associated to the clone/probe currently under the mouse pointer. These include: centering the profile around the current position; drawing of a vertical bar through all the profiles (to define a locus or a region); and linking to external web pages from NCBI clone or MapViewer (http://www.ncbi.nlm.nih.gov/mapview and Wheeler *et al*., 2005), UCSC Genome Browser (http://genome.ucsc.edu and Kent *et al*., 2002), Ensembl Contig View or CytoView (http://www.ensembl.org and Hubbard *et al*., 2005), *Saccharomyces* Genome Database (http://www.yeastgenome.org). New links are defined in a XML configuration file and adding them is straightforward. Most data and results (profiles, minimal regions, etc.) can be exported and saved in full text, csv (comma separated values) or HTML format. We refer the reader to the user manual for a description of the other functions.

2.4 Software architecture and requirements

The software architecture is shown in Figure 6. The core of the interface consists of a Java applet, and was developed using the Swing library. It runs on any operating system supporting Java 1.4.2

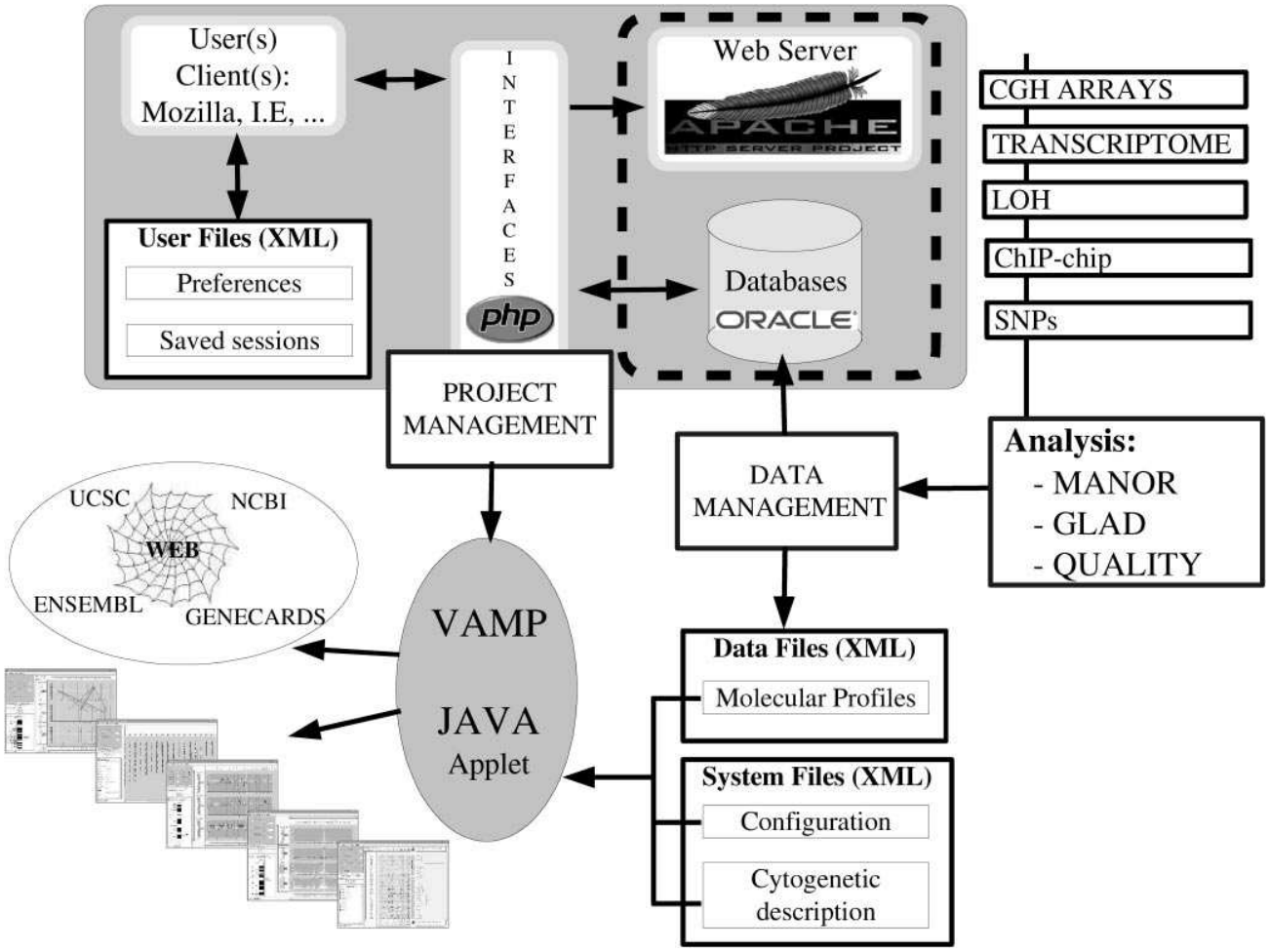


Fig. 6. Software architecture of a microarray environment based on VAMP. VAMP can also be used as a local application.

(we recommend computers with a minimum of 1 Gb memory, although 256 Mb is enough for small projects). The data used by the program are of several types:

- The genome profile information, which are retrieved either from a relational database management server (currently Oracle™) or from XML data files. These include the signal value for each clone/probe and its genomic location.
- The system files (also in XML), which includes the cytogenetic description of the genome under study and the configuration parameters (environment variables for file and URL management). Cytogenetic banding files for human ISCN 400, 550 and 850 descriptions, as well as mouse and yeast genome descriptions are also available. The user files, which consist of the user visualization preferences and saved sessions.

VAMP can be used either as a local application, with all data and configuration files directly accessible to the client, or as an applet, with all data and configuration files installed on a server. In this mode, only the user configuration file is stored locally on the client machine.

VAMP can be easily installed on any platform running Java 1.4.2. All that is needed is to convert the microarray data into XML files, with a specific syntax described in a DTD (XML Document Type Definition). The use of a database management server is not mandatory, although it is recommended for large-scale projects. Arbitrary complementary profile information can be added to the XML files, and this information can be displayed by the interface.

3 DISCUSSION

We have developed a graphical user interface for the visualization and analysis of any type of genomic profile, with an emphasis on array-CGH. VAMP is currently used in cancer genomic projects on human and mouse samples and in studying the proteins involved in the reparation, recombination and replication of DNA in yeast. It is used in Institut Curie and many labs in Europe and the United States. Several publications describing data analysis with VAMP are coming soon. Janoueix-Lerosey *et al*., (2005) describe the use of VAMP for replication timing data analysis (http://microarrays.curie.fr/publications/U509/reptiming). In Institut Curie, ∼3600 microarray profiles have been interfaced with VAMP to date.

VAMP aids greatly in finding genes of clinical and biological importance from CGH, transcriptome, LOH, ChIP chip profiles and SNP arrays. VAMP improves upon existing solutions such as SeeGH (Chi *et al*., 2004), CGHPRO (Chen *et al*., 2005), CGH-Analyzer (Margolin *et al*., 2005) or general purpose spreadsheet software, because it offers many different modes of visualization, allows the display of several samples and of several types of profiles simultaneously, and offers many data analysis functions. VAMP can be compared with other general-purpose genomic browsers such MapView (NCBI), Genome Browser of UCSC or Ensembl. VAMP is well suited to handle sample profiles and to analyse this type of data, which the other genomic browsers are not designed to do. Therefore, in cancer research it addresses a real need and is a useful tool for biologists and clinicians. Our software is fully portable and only requires a computer running Java 1.4.2 and data in XML format.

VAMP can be run on public datasets at http://bioinfo.curie.fr/vamp. The array-CGH data from Snijiders *et al*. (2001, 2005), Pollack *et al*. (2002), Veltman *et al*. (2003), Nakao *et al*. (2004), Douglas *et al*. (2004), de Leeuw *et al*. (2004), Gysin *et al*. (2005), Patil *et al*. (2005) and Bredel *et al*. (2005) are currently browsable. Expression profiles are also available for the samples from Pollack *et al*. (2002).

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie, the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, the CNRG and the Ligue contre le Cancer.

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. *et al*. (2003) Chromosome aberrations in solid tumors. *Nat. Genet*., **34**, 369–76.

Bignell,G.R. *et al*. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res*., **14**, 287–295.

Bilke,S. *et al*. (2005) Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics*, **21**, 1138–1145.

Bredel,M. *et al*. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*., **65**, 4088–4096.

Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Chen,W. *et al*. (2005) CGHPRO—a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, **6**, 85.

Chi,B. *et al*. (2004) SeeGH—a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, **5**, 13.

de Leeuw,R.J. *et al*. (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum. Mol. Genet*., **13**, 1827–1837.

Douglas,E.J. *et al*. (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. *Cancer Res*., **64**, 4817–4825.

Eilers,P.H.C. and de Menezes,R.X. (2005) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Fridlyand,J. *et al*. (2004) Application of hidden markov models to the analysis of the array CGH data. *J. Multivari. Anal*. (Special Issue on Multivariate Methods in Genomic Data Analysis), **90**, 132–153.

Gysin,S. *et al*. (2005) Analysis of genomic DNA alterations and mRNA expression patterns in a panel of human pancreatic cancer cell lines. *Genes Chromosomes Cancer*, **44**, 37–51.

Huang,J. *et al*. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics*, **1**, 287–299.

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, 447–453.

Hupé,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Ishkanian,A.S. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet*., **36**, 299–303.

Janoueix-Lerosey,I. *et al*. (2005) Preferential occurrence of chromosome breakpoints within early replicating regions in neuroblastoma. *Cell Cycle*, **4**, 1842–1846.

Jong,K. *et al*. (2003) Chromosomal breakpoint detection in human cancer. In Raidl,G.R., Cagnoni,S., Cardalda,J.J.R., Corne,D.W., Gottlieb,J., Guillot,A., Hart,E., Johnson,C.G., Marchiori,E., Meyer,J.-A. and Middendorf,M. (eds), *Applications of Evolutionary Computing, EvoWorkshops2003: EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, EvoSTIM*, vol. 2611 of *LNCS*. Springer-Verlag, University of Essex, England, UK.

Kaufman,L. and Rousseuw,P. (1990) *Finding Groups in Data—An Introduction to Cluster Analysis*, Wiley Series in Probability and Mathematical Sciences. John Wiley & Sons.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Margolin,A. *et al*. (2005) CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*, **21**, 3308–3311.

Nakao,K. *et al*. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.

Neuvial,P., Hupé,P., Brito,I., Liva,S., Manié,E., Brennetot,C., Radvanyi,F., Aurias,A. and Barillot,E. (2005) Spatial normalization of array-CGH data. *BMC Bioinformatics*, **7**, 264.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Patil,M.A. *et al*. (2005) Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis*, **26**, 2050–2057.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37** (Suppl.1), 11–17.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet*., **20**, 207–211.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Rouveirol,C. *et al*. (2005) Computation of recurrent minimal genomic alterations from CGH data. *Bioinformatics*, **22**, 849–856.

Snijders,A.M. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet*., **29**, 263–4.

Snijders,A.M. *et al*. (2005) Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, **24**, 4232–4242.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Veltman,J.A. *et al*. (2003) Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*., **63**, 2872–2880.

Vogelstein,B. *et al*. (1989) Allelotype of colorectal carcinomas. *Science*, **244**, 207–11.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, 39–45.

APPENDIX D

CGH-array analysis web platform

Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W477–W481***
doi:10.1093/nar/gkl215

CAPweb: a bioinformatics CGH array Analysis Platform

Stéphane Liva[1,*], Philippe Hupé[1,2], Pierre Neuvial[1], Isabel Brito[1], Eric Viara[1], Philippe La Rosa[1] and Emmanuel Barillot[1]

[1]Institut Curie, Service Bioinformatique and [2]Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris Cedex 05, France

Received February 14, 2006; Revised and Accepted March 24, 2006

ABSTRACT

Assessing variations in DNA copy number is crucial for understanding constitutional or somatic diseases, particularly cancers. The recently developed array-CGH (comparative genomic hybridization) technology allows this to be investigated at the genomic level. We report the availability of a web tool for analysing array-CGH data. CAPweb (CGH array Analysis Platform on the Web) is intended as a user-friendly tool enabling biologists to completely analyse CGH arrays from the raw data to the visualization and biological interpretation. The user typically performs the following bioinformatics steps of a CGH array project within CAPweb: the secure upload of the results of CGH array image analysis and of the array annotation (genomic position of the probes); first level analysis of each array, including automatic normalization of the data (for correcting experimental biases), breakpoint detection and status assignment (gain, loss or normal); validation or deletion of the analysis based on a summary report and quality criteria; visualization and biological analysis of the genomic profiles and results through a user-friendly interface. CAPweb is accessible at http://bioinfo.curie.fr/CAPweb.

INTRODUCTION

In recent years, array-CGH (comparative genomic hybridization) has become the technology of choice for large scale investigations of DNA copy number changes between two genomes. Today, CGH arrays allow the ratio of DNA copy number between a test and a reference sample to be simultaneously assessed in 2000 to 30 000 positions in the genome, giving a resolution of between 1.5 Mb to 100 kb (1,2). Its main applications are the study of diseases in which the DNA copy number varies in certain locations of the genomes, due to either constitutional mutations (hereditary or *de novo*), such as human genetic diseases (3) or somatic changes, such as in cancers (4). The identification of regions of altered DNA gives valuable information about the genes involved in the disease, and many projects have been launched worldwide to determine the genome structure of tumour cells (4). Array-CGH is also an important source of information for studying genome evolution, for example in bacteria (5) or mammals (6). We have developed a Web tool, called CAPweb (CAP: CGH array Analysis Platform), for bioinformatics analysis of CGH arrays. This tool combines the following tasks: (i) data management, (ii) array normalization, (iii) automatic breakpoint detection and assessment of gain and loss regions, (iv) quality control and (v) a graphical user interface for browsing and analysing the genomic profiles.

Several tools have recently been developed for analysing CGH array data, such as CGH-Explorer (7), ArrayCyGHt (8), CGHPRO (9), WebArray (10) or ArrayCGHbase (11), although the only web-accessible servers are ArrayCyGHt, WebArray and CAPweb. Among these three, only CAPweb allows project management and the upload of raw data files without pre-processing. It also offers unique features for the analysis and visualization of array-CGH data. CAPweb accepts raw data from the main microarray image analysis software. As far as we are aware, CAPweb is the only platform dedicated to biologists that allows the complete analysis of raw CGH arrays from the raw data to visualization and biological interpretation.

DESCRIPTION

The CAPweb server allows the user to store, analyse and manage his or her data. We will now describe its operation (Figure 1). A tutorial is accessible at http://bioinfo.curie.fr/tutorial/CAPweb/capweb_tutorial.html.

*To whom correspondence should be addressed. Tel: +33 0 1 4234 65 31; Fax: +33 0 1 42 34 65 28; Email: capweb@curie.fr

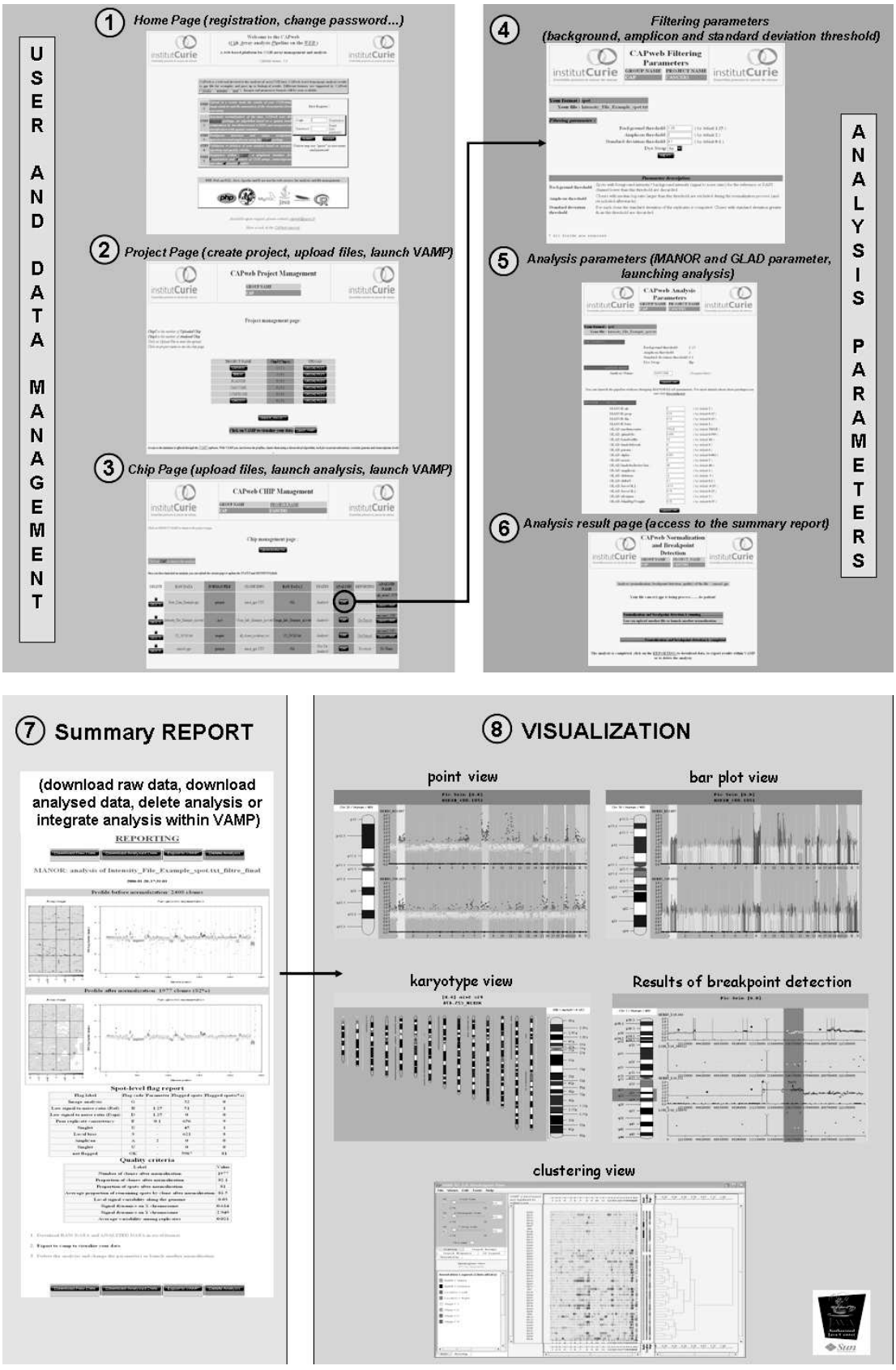


Figure 1. Different views of CAPweb Interface showing how the CGH array analysis proceeds, see text for details.

User registration, data upload and management

The first step of the analysis is user registration [Figure 1(1)], which ensures the confidentiality of the submitted data. The user is sent a login/password by email and can then create one or more projects to upload data files [Figure 1(2)]. Several input formats from microarray image analysis software are currently supported: Genepix (http://www.moleculardevices.com/pages/instruments/gn_genepix4000.html), Imagene (http://www.biodiscovery.com/index/imagene), Spot (12) and MAIA (13). CAPweb requires only two types of file: (i) a raw intensity file (one file for Genepix and MAIA, two files for Imagene and Spot) and (ii) a genomic position file mapping each spot to a name and its position on the genome under CSV (semi colon separator) format.

For each project, the 'Array Management' page [Figure 1(3)] lists all the arrays, their analysis status and the summary report file, and allows new analyses to be launched.

The array files are permanently stored on the server: the user can only browse the arrays of his or her projects, and only the user is allowed to delete them.

CGH array analysis

From the 'Array management' page, the user can launch the array analyses. The analyses are run in the background, allowing the user to use CAPweb for other analyses.

Data Normalization (MANOR). As in all microarray analyses, CGH array data must be normalized to correct for experimental artefacts while preserving the true biological signal. For this goal, CAPweb uses the Bioconductor package MANOR, which includes spot and clone filtering steps that discards spots having too low a signal-to-noise ratio or clones with a poor replicate consistency, and, most importantly, it includes a spatial normalization step. This step aims to correct for spatial effects on the arrays. We identified these as the predominant experimental artefact in the array-CGH data we have studied. The corresponding algorithm is based on a spatial trend estimation and a signal segmentation method with a spatial constraint, as described in P. Neuvial *et al.* (manuscript submitted).

Breakpoint detection and assessment of gain and loss region (GLAD). This step aims to identify chromosomal regions having an identical DNA copy number, which are delimited by breakpoints. CAPweb uses the Bioconductor package GLAD, which implements an algorithm described in (14). This method first uses the spatial structure of array-CGH data to adaptively calculate a smoothed signal value for each clone. These smoothed signal values are then used to detect breakpoints and outliers, and then genomic regions having the same underlying copy number are clustered together.

Quality control. Various statistical criteria can help the user assess the quality of the array. These include intra-replicate variability, genomic neighbour variability, the percentage of spots filtered out after image analysis and the amplitude of signal gap between regions having a different DNA copy number. These quality criteria are reported in an HTML summary report file, which also displays key features of the normalization process: array image and genomic profile before and after normalization, and a summary of the normalization. This file [Figure 1(7)] allows the user to compare the quality of the data before and after analysis. Based on this information, the user may choose to keep or discard the analysis.

This data analysis step can be run without an extensive knowledge of the underlying statistical algorithms by using default parameters. Default parameters have been calibrated by comparing quality criteria for various parameter value in two datasets: one from UCSF (218 arrays, Spot format, as a collaboration with Dan Pinkel), and one from Institut Curie/INSERM U509 (181 arrays, Genepix format). This part is described in detail elsewhere (P. Neuvial *et al.* manuscript submitted). However, CAPweb allows the user to choose the value of several parameters for filtering, spatial normalization and breakpoint detection. The summary report also helps in comparing the results of analyses carried out with different parameter values [Figure 1 (4–6)].

Visualization (VAMP) and biological analysis

Once the first level of array analysis has finished, the user can visualize and further analyse the data through a graphical user interface: VAMP—visualization and analysis of array-CGH, transcriptome and other molecular profiles (P. La Rosa *et al.* manuscript submitted) [Figure 1 (8)]. Several visualization types are proposed, such as the classical CGH karyotype view or the genome-wide multi-tumour comparison view. These allow the user to easily compare different arrays. Additional information concerning each clone or DNA region can be interactively retrieved from different public databases through external links. Other functions for analysing CGH data are provided within the interface, such as looking for minimal or recurrent regions of alterations (15), clustering, etc.

VAMP allows the user to display genomic profiles at various resolutions [from the whole genome to small regions (clone level)]. All the analyses results (breakpoint detection, assignment of gain/lost region, quality criteria, etc.) can also be displayed within VAMP. VAMP has many other functions for navigation, querying and analysis that we have not explained here; we refer the reader to the documentation and demo for further details (http://bioinfo.curie.fr/vamp/doc).

Note that the user can analyse at least 200 arrays with 1GB of memory.

IMPLEMENTATION

The CAPweb server is based on freely available components (Figure 2). The database for user management and array management was built on mySQL. PHP scripts ensure registration and project management. Perl scripts control the launching of statistical analyses written in R. A Java applet and XML files are used for the visualization. CAPweb integrates the MANOR and GLAD R packages and the VAMP software, all of which were developed at the Institut Curie.

The security in CAPweb is based on mysql authentication and cookie session. Uploaded data are considered strictly confidential. The CAPweb server is also available upon request for local installation on Unix/Linux/MacOS X operating systems.

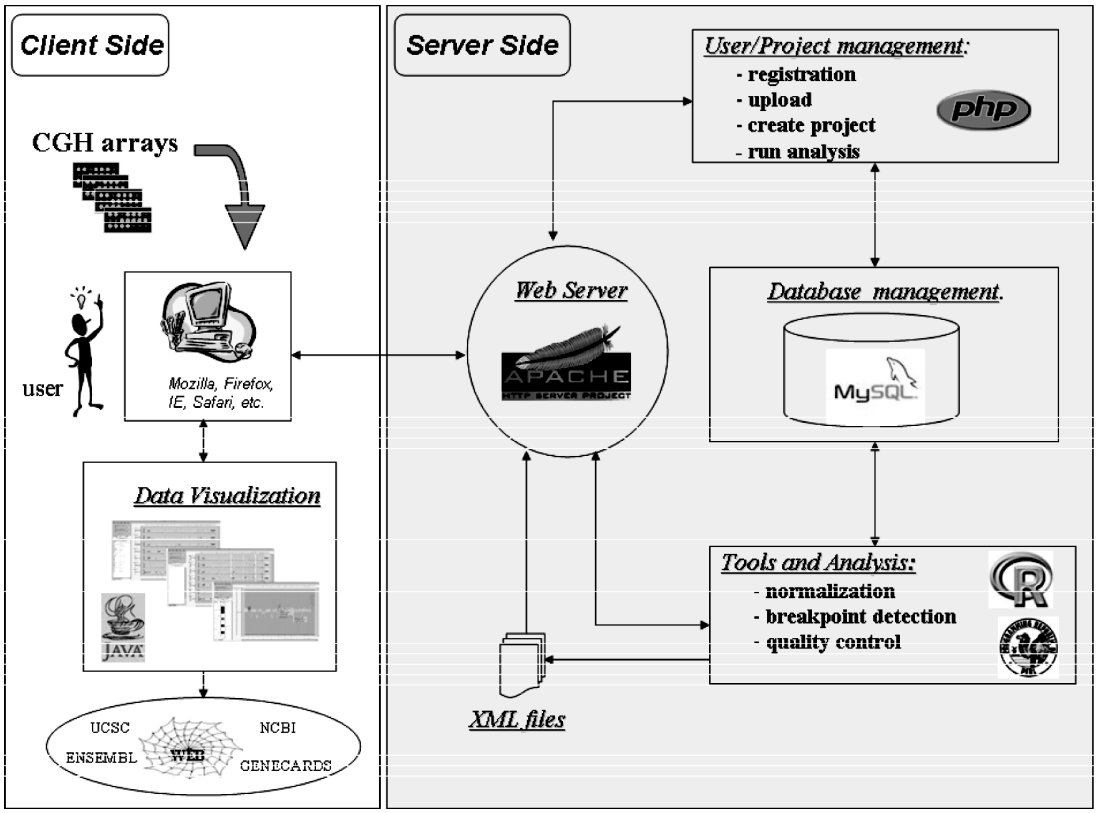


Figure 2. CAPweb software architecture, see text for details.

CONCLUSION

Array-CGH is a popular technology that is now used in many projects ranging from the characterization of tumours to the study of genome evolution. As with any large scale technology, its exploitation relies heavily on the availability of bioinformatics tools for managing and analysing the data. Many bioinformatics algorithms and interfaces have been developed but biologists have lacked a web-based platform for integrating these tools in a user-friendly manner. CAPweb offers this service and combines array normalization, quality control, breakpoint detection and the biological interpretation of the results. It also helps with data management. Currently, the public CAPweb server at the Institut Curie contains 800 arrays.

In this paper we have presented CAPweb 1.0 version. A new version is currently being developed, which will allow the user to analyse high density oligonucleotide arrays, such as Affymetrix GeneChip® Arrays or Nimblegen™ Arrays, to integrate any clinical information, and to add gene expression profiles so that copy number profiles can be compared and correlated to them.

ACKNOWLEDGEMENTS

This work was supported by the Institut Curie, the Centre National de la Recherche Scientifique, the Cancéropole Ile-de-France, the Région Ile-de-France and the association 'Courir pour la vie, Courir pour Curie'. The authors thank all our colleagues who have tested CAPweb and suggested improvements: G. Pierron, C. Brennetot, A. Idbaih, E. Manié (Institut Curie) and S. Law (UCSF). Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet.*, **29**, 263–264.
2. Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet.*, **36**, 299–303.
3. Lockwood,W.W., Chari,R., Chi,B. and Lam,W.L. (2006) Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet.*, **14**, 139–148.
4. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nature Genet.*, **37**, 11–17.
5. Fukiya,S., Mizoguchi,H., Tobe,T. and Mori,H. (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative hybridization microarray. *J. Bacteriol.*, **186**, 3911–3921.
6. Wilson,G.M., Flibotte,S., Missirlis,P.I., Marra,M.A., Jones,S., Thornton,K., Clark,A.G. and Holt,R.A. (2006) Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Res*, **16**, 173–181.
7. Lingjaerde,O.C., Baumbush,L.O., Liestol,K., Glad,I.K. and Borresen-Dale,A.L. (2005) CGH-explorer, a program for analysis of array-CGH data. *Bioinformatics*, **6**, 821–822.

8. Kim,S.Y., Nam,S.W., Lee,S.H., Park,W.S., Yoo,N.J., Lee,J.Y. and Chung,Y.J. (2005) ArrayCyGHt, a web application for analysis and visualization of array-CGH data. *Bioinformatics*, **21**, 2554–2555.
9. Chen,W., Erdogan,F., Ropers,H., Lenzner,S. and Ullmann,R. (2005) CGHPRO, a comprehensive data analysis tool for array CGH. *BMC Bioinformatics*, **6**, 85.
10. Xia,X., McClelland,M. and Wang,Y. (2005) WebArray, an online platform for microarray data analysis. *BMC Bionformatics*, **6**, 306.
11. Menten,B., Pattyn,F., De Preter,K., Robbrecht,P., Michels,E., Buysse,K., Mortier,G., De Paepe,A., van Vooren,S., Vermeesh,J. *et al*. (2005) ArrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics*, **6**, 124.
12. Jain,A.N., Tokuyasu,T.A., Snidjers,A.M., Segraves,R., Albertson,D.G. and Pinkel,D. (2002) Fully automatic quantification of microarray image data. *Genome Res*., **12**, 325–332.
13. Novikov,E. and Barillot,E. (2005) A robust algorithm for ratio estimation in two-color microarray experiments. *J. Bioinform. Comput. Biol*., **6**, 1411–1428.
14. Hupé,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss DNA regions. *Bioinformatics*, **20**, 3413–3422.
15. Rouveirol,C., Stransky,N., Hupé,P., La Rosa,P., Viara,E., Barillot,E. and Radvanyi,F. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

APPENDIX E

A particular CGH-array experiment

Figure 1 is the output of a CGH-array experiment that I have performed under the supervision of Gaëlle Pierron and Élodie Manié, on a neuroblastoma cell line that exhibits a characteristic amplification of the MYCN oncogene.

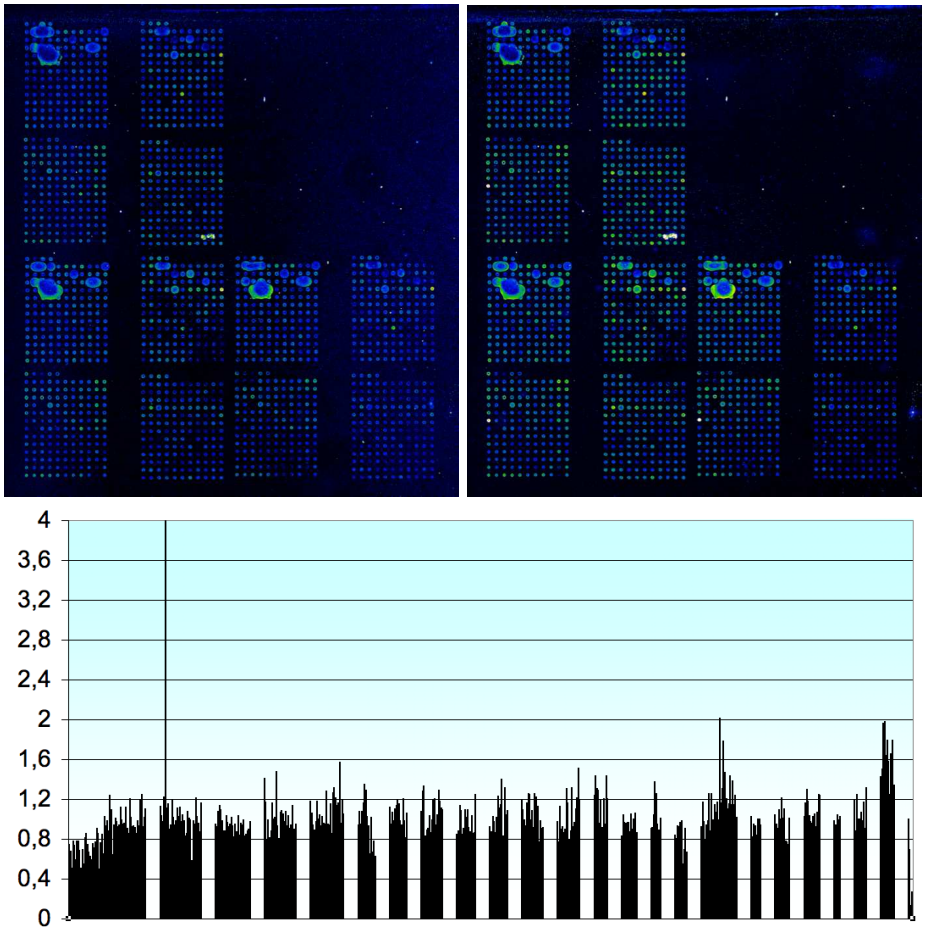


FIGURE 1. *Results of a CGH-array experiment on a neuroblastoma cell line. Top: Cy3 and Cy5 images; bottom: genomic profile (copy number ratios). The peak on chromosome 2 corresponds to the amplification of MYCN.*