Analyse statistique de données biologiques à haut débit

Julie Aubert

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2017SACLS048

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Établissements d'accueil : AgroParisTech
Institut national de la recherche agronomique

Laboratoires d'accueil : Mathématiques et informatique appliquées, UMR 518 AgroParisTech-INRA
Mathématiques et informatique appliquées du génome à l'environnement, UR 1404 INRA

Spécialité de doctorat : Mathématiques aux interfaces

Julie AUBERT

Analyse statistique de données biologiques à haut débit

Date de soutenance : 7 février 2017

Rapportrices :
FLORENCE FORBES (INRIA)
SUSAN HOLMES (Standford University)

Jury de soutenance :

FLORENCE FORBES	(INRIA) Rapportrice
CHRISTINE KERIBIN	(Université Paris Sud) Examinatrice
STÉPHANE LECROM	(UPMC) Examinateur
FRANCK PICARD	(CNRS) Président
STÉPHANE ROBIN	(INRA) Codirecteur de thèse
SOPHIE SCHBATH	(INRA) Directrice de thèse

Fondation mathématique FMJH Jacques Hadamard
INRA SCIENCE & IMPACT
AgroParisTech
MaIAGE Mathématique et Informatique Appliquées du Génome à l'Environnement

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Analyse statistique de données biologiques à haut débit

Mots clefs : Modèle à variables latentes, analyse statistique, données de comptage, binomiale négative, données omiques

Résumé : Les progrès technologiques des vingt dernières années ont permis l'avènement d'une biologie à haut-débit reposant sur l'obtention de données à grande échelle de façon automatique. Les statisticiens ont un rôle important à jouer dans la modélisation et l'analyse de ces données nombreuses, bruitées, parfois hétérogènes et recueillies à différentes échelles. Ce rôle peut être de plusieurs natures. Le statisticien peut proposer de nouveaux concepts ou méthodes inspirées par les questions posées par cette biologie. Il peut proposer une modélisation fine des phénomènes observés à l'aide de ces technologies. Et lorsque des méthodes existent et nécessitent seulement une adaptation, le rôle du statisticien peut être celui d'un expert, qui connaît les méthodes, leurs limites et avantages. Le travail présenté dans cette thèse se situe à l'interface entre mathématiques appliquées et biologie, et relève plutôt des deuxième et troisième type de rôles mentionnés.

Dans une première partie, j'introduis différentes méthodes développées pour l'analyse de données biologiques à haut débit, basées sur des modèles à variables latentes. Ces modèles permettent d'expliquer un phénomène observé à l'aide de variables cachées. Le modèle à variables latentes le plus simple est le modèle de mélange. Les deux premières méthodes présentées en sont des exemples : la première dans un contexte de tests multiples et la deuxième dans le cadre de la définition d'un seuil d'hybridation pour des données issues de puces à ADN. Je présente également un modèle de chaînes de Markov cachées couplées pour la détection de variations du nombre de copies en génomique prenant en compte de la dépendance entre les individus, due par exemple à une proximité génétique. Pour ce modèle, nous proposons une inférence approchée fondée sur une approximation variationnelle, l'inférence exacte ne pouvant pas être envisagée dès lors que le nombre d'individus augmente. Nous définissons également un modèle à blocs latents modélisant une structure sous-jacente par bloc de lignes et colonnes adaptées à des données de comptage issue de l'écologie microbienne. Les données issues de méta-codebarres ou de métagénomique correspondent à l'abondance de chaque unité d'intérêt (par exemple micro-organisme) d'une communauté microbienne au sein d'environnement (rhizosphère de plante, tube digestif humain, océan par exemple). Ces données ont la particularité de présenter une dispersion plus forte qu'attendue sous les modèles les plus classiques (on parle de sur-dispersion). La classification croisée est une façon d'étudier les interactions entre la structure des communautés microbiennes et les échantillons biologiques dont elles sont issues. Nous avons proposé de modéliser ce phénomène à l'aide d'une distribution Gamma-Poisson et développé une autre approximation variationnelle pour ce modèle particulier ainsi qu'un critère de sélection de modèle. La flexibilité et la performance du modèle sont illustrées sur trois jeux de données réelles.

Une deuxième partie est consacrée à des travaux dédiés à l'analyse de données de transcriptomique issues des technologies de puce à ADN et de séquençage de l'ARN. La première section concerne la normalisation des données (détection et correction de biais techniques) et présente deux nouvelles méthodes que j'ai proposées avec mes co-auteurs et une comparaison de méthodes à laquelle j'ai contribuée. La deuxième section dédiée à la planification expérimentale présente une méthode pour analyser les dispositifs dit en dye-switch.

Dans une dernière partie, je montre à travers deux exemples de collaboration, issues respectivement d'une analyse de gènes différentiellement exprimés à partir de données issues de puces à ADN, et d'une analyse du traductome chez l'oursin à partir de données de séquençage de l'ARN, la façon dont les compétences statistiques sont mobilisées et la plus-value apportée par les statistiques aux projets de génomique.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Title : Statistical analysis of high-throughput biology data

Keywords : latent variables models, statistical analysis, count data, negative binomial distribution, omics data

Abstract : The technological progress of the last twenty years allowed the emergence of an high-throuput biology basing on large-scale data obtained in a automatic way. The statisticians have an important role to be played in the modelling and the analysis of these numerous, noisy, sometimes heterogeneous and collected at various scales. This role can be from several nature. The statistician can propose new concepts, or new methods inspired by questions asked by this biology. He can propose a fine modelling of the phenomena observed by means of these technologies. And when methods exist and require only an adaptation, the role of the statistician can be the one of an expert, who knows the methods, their limits and the advantages.

In a first part, I introduce different methods developed with my co-authors for the analysis of high-throughput biological data, based on latent variables models. These models make it possible to explain a observed phenomenon using hidden or latent variables. The simplest latent variable model is the mixture model. The first two presented methods constitutes two examples : the first in a context of multiple tests and the second in the framework of the definition of a hybridization threshold for data derived from microarrays. I also present a model of coupled hidden Markov chains for the detection of variations in the number of copies in genomics taking into account the dependence between individuals, due for example to a genetic proximity. For this model we propose an approximate inference based on a variational approximation, the exact inference not being able to be considered as the number of individuals increases. We also define a latent-block model modeling an underlying structure per block of rows and columns adapted to count data from microbial ecology. Metabarcoding and metagenomic data correspond to the abundance of each microorganism in a microbial community within the environment (plant rhizosphere, human digestive tract, ocean, for example). These data have the particularity of presenting a dispersion stronger than expected under the most conventional models (we speak of over-dispersion). Biclustering is a way to study the interactions between the structure of microbial communities and the biological samples from which they are derived. We proposed to model this phenomenon using a Poisson-Gamma distribution and developed another variational approximation for this particular latent block model as well as a model selection criterion. The model's flexibility and performance are illustrated on three real datasets.

A second part is devoted to work dedicated to the analysis of transcriptomic data derived from DNA microarrays and RNA sequencing. The first section is devoted to the normalization of data (detection and correction of technical biases) and presents two new methods that I proposed with my co-authors and a comparison of methods to which I contributed. The second section devoted to experimental design presents a method for analyzing so-called dye-switch design.

In the last part, I present two examples of collaboration, derived respectively from an analysis of genes differentially expressed from microrrays data, and an analysis of translatome in sea urchins from RNA-sequencing data, how statistical skills are mobilized, and the added value that statistics bring to genomics projects.

Remerciements

Avant tout chose, je tiens à remercier mes deux 'papas' scientifiques, à savoir Jean-Jacques Daudin et Stéphane Robin qui m'ont accueillie dans l'équipe Statistique et Génome lors de mon recrutement à l'INRA. Merci à toi Stéphane, qui m'a incitée à me lancer dans une thèse. Cette idée à laquelle je n'étais pas forcément favorable il y a maintenant 13 ans, a fait son chemin et je me suis lancée dans l'aventure il y a quelques années. Il me fallait donc un sujet. Merci Jean-Jacques pour m'avoir permis au travers d'un projet MEM (MétaOmiques des Ecosystèmes Microbiens) de rencontrer Christophe Mougel. Cette rencontre m'a donné l'occasion de trouver à la fois un sujet de recherche intéressant et un collaborateur formidable. Pas de thèse, sans directeur. Merci à Sophie Schbath et Stéphane Robin, qui ont accepté sans réserve d'encadrer ce travail de thèse.

Je remercie mes responsables d'équipe (Stéphane puis Céline) et de laboratoire (Stéphane puis Liliane) qui m'ont fait confiance et m'ont toujours accordé une grande liberté dans mes choix de travail et collaborateurs et dans mes activités d'animation diverses et variées.

Je remercie Susan Holmes et Florence Forbes pour avoir accepté de rapporter cette thèse et pour avoir pris le temps de relire mon manuscrit. Merci pour vos conseils et suggestions.
Merci à Christine Kéribin, Stéphane Le Crom, Franck Picard pour avoir accepté de faire partie de mon jury de thèse, pour votre relecture attentive et vos discussions diverses et variées.

Merci à tous mes collaborateurs statisticiens, passés ou encore présents au sein des unités Mathématiques et Informatique Appliquées (MIA-Paris), et Mathématiques et Informatique Appliquées du Génome aux Organismes (MaIAGE) : Jean-Jacques, Stéphane, Marie-Laure, Tristan, Franck, Avner, Emilie, Xiao, Trung, Aurélie, Frédéric, Sophie, Mahendra.
Merci aux membres des réseaux SSB et StatOmique pour tous nos échanges réguliers. Merci en particulier à Christelle Hennequet-Antier, ma binôme animatrice (StatOmique) et formatrice avec qui c'est toujours aussi agréable de travailler et d'échanger. Merci à

Marie-Agnès Dillies d'avoir rejoint notre duo pour animer Statomique.

Merci aux membres du pôle planification expérimentale et RNA-Seq du PEPI Ingénierie BioInformatique et Statistique à haut débit (IBIS) de l'INRA et du bureau du PEPI IBIS pour nos échanges interdisciplinaires. Merci en particulier à Valentin Loux qui a été co-animateur avec moi de ce PEPI pendant 5 ans.

Merci à tous mes collaborateurs biologistes pour proposer de jolis problèmes et surtout pour échanger sur les besoins en statistique, l'intérêt d'une méthode ou la pertinence de résultats. La liste serait trop longue et je ne peux tous vous citer. Je remercie en particulier Oliver Sandra, Pascal Bonnarme et son équipe, Stéphane Nicolas, Julia Moralès pour m'avoir respectivement initiée au monde de la biologie de la reproduction, des fromages, du maïs ou encore des oursins. Un grand merci à Jean-Marie Beckerich pour notre travail sur la résilience dans un écosystème fromager et pour sa confiance. Merci également à Corinne Vacher et Christophe Mougel, Anouk Zancarini et Christine Le Signor : j'espère que nous continuerons à travailler ensemble.

Merci à tous ceux qui m'ont soutenue d'une manière ou d'une autre (une astuce, un bout de code, un mot, un geste, une présence), ou qui ont témoigné de l'intérêt pour mon travail. Merci à tous mes co-bureaux successifs (Camille, Colette, Nathalie, Marie, Aurore, Jean-Benoist, Xiao, Marie-Laure, Julien). Merci à l'ensemble des membres du département MMIP de l'AgroParisTech ou de l'unité MaIAGE, permanents et non-permanents, qui font que venir à Paris ou à Jouy est toujours un plaisir (parfois même culinaire, n'est-ce pas Cyprien ?). Sans oublier, les gestionnaires, secrétaires de nos laboratoires, qui font que souvent notre vie est plus facile. Merci à Eric Quémard pour l'impression des exemplaires du manuscrit de thèse pour la soutenance, qui a été laborieuse (qui connaît l'histoire de la fameuse page qui ne voulait pas s'imprimer ?).

Merci à mes amis, ma famille et belle-famille, en particulier mes parents, qui ont cru en moi et ont toujours été fiers de moi, quoi que je fasse. Un grand grand merci à mon mari, Sébastien, pour son soutien inconditionnel, sa présence au quotidien et bien sûr, pour m'avoir donné deux beaux enfants formidables, très compréhensifs quand maman a du travail à terminer ou doit partir en déplacement pour le boulot.

Et merci à toi lecteur pour l'intérêt que tu portes à ce travail.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Les progrès technologiques des vingt dernières années ont permis l'avénement d'une biologie à haut-débit reposant sur l'obtention de données à grande échelle de façon automatique. Les statisticiens ont un rôle important à jouer dans la modélisation et l'analyse de ces données nombreuses, bruitées, parfois hétérogènes et recueillies à différentes échelles. Ce rôle peut être de plusieurs natures.
Le statisticien peut proposer de nouveaux concepts, ou nouvelles méthodes inspirées par les questions posées par cette biologie. La génomique à haut-débit à l'aide de la technologie des puces à ADN, par exemple, a conduit à de grandes avancées dans l'analyse de données statistiques en grande dimension (tests multiples, régressions pénalisées et régularisées). Une expérience de transcriptomique typique vise à comparer entre deux conditions le niveau d'expression de milliers de gènes simultanément. Dans ce cas, le nombre de variables (ici les gènes) est grand, mais le nombre d'observations pour chacune des variables est relativement faible (une dizaine maximum). Ces anciennes techniques ont été remplacées par les nouvelles technologies de séquençage qui apportent elles aussi leur lot de nouveautés et défis, notamment du fait de la nature discrète des données produites.

Le statisticien peut également proposer une modélisation fine des phénomènes observés à l'aide de ces technologies. En effet, les questionnements des biologistes évoluent rapidement avec les évolutions technologiques. Aujourd'hui les chercheurs en biologie ont à leur disposition de nombreux outils pour essayer de comprendre le lien entre phénotype, génotype et environnement.

Les modèles à variables latentes permettent d'expliquer un phénomène observé à l'aide de variables cachées ou latentes. Ils constituent un outil très puissant dans de nombreux domaines d'application, et pas seulement en biologie, grâce à leur flexibilité permettant un ajustement plus fin à la réalité. Le modèle de mélange, qui entre dans cette catégorie, est par exemple couramment utilisé à des fins de classification non supervisée. La classification consiste à classer des individus ou objets dans des classes homogènes mais distinctes les unes des autres (gène exprimé versus gène non exprimé ; individu malade versus individu sain), et à découvrir les classes elles-mêmes. Il existe bien entendu des algorithmes heuristiques basés sur des métriques, comme les k-means, mais ces derniers dépendent très fortement de la métrique envisagée (comment différencier les objets entre eux ?) et du critère choisi pour regrouper et séparer les classes. Les modèles à variables latentes ont l'avantage de se situer dans un cadre probabiliste permettant une justification et une interprétation plus aisée de la classification proposée.

Et lorsque des méthodes existent et nécessitent seulement une adaptation, le rôle du statisticien peut être celui d'un expert, qui connaît les méthodes, leurs limites et avantages. Le travail présenté dans ce manuscrit se situe à l'interface entre mathématiques appliquées et biologie, et relève plutôt des deuxième et troisième type de rôles mentionnés.

Le manuscrit se découpe en 6 chapitres. Les deux premiers chapitres introductifs présentent respectivement les notions de biologie et de statistiques nécessaires à la compréhension des travaux présentés par la suite.

Les chapitres 3 et 4 présentent le développement de méthodes basées sur des modèles à variables latentes. Le modèle à variables latentes le plus simple est le modèle de mélange. Dans ce modèle, les observations sont supposées indépendantes, chacune appartenant à une classe non observée.
Les travaux présentés dans le chapitre 3 en sont des exemples : le premier dans un contexte de tests multiples et le deuxième dans le cadre de la définition d'un seuil d'hybridation pour des données issues de puces à ADN. Dans ce même chapitre, je présente un modèle de chaînes de Markov cachées avec prise en compte de la dépendance entre les individus, pour lequel nous avons proposé une inférence approchée, l'inférence exacte ne pouvant pas être envisagée dès lors que le nombre d'individus augmente. Le chapitre 4 qui a constitué le coeur de mon travail de thèse présente un modèle à blocs latents modélisant une structure sous-jacente par bloc de lignes et colonnes adaptées à des données de comptage issue de l'écologie microbienne. Ces données ont la particularité d'être surdispersées. Nous avons proposé de modéliser ce phénomène à l'aide d'une distribution Poisson-Gamma et développé une approche variationnelle pour ce modèle à blocs latents particulier ainsi qu'un critère de sélection de modèle.

Les travaux présentés dans le chapitre 5 sont dédiés à l'analyse de données de transcriptomique issues des technologies de puce à ADN et de séquençage de l'ARN. La première section est dédiée à la normalisation des données (détection et correction de biais techniques) et présente deux nouvelles méthodes que j'ai proposées avec mes co-auteurs et une comparaison de méthodes à laquelle j'ai contribuée. La deuxième section dédiée à la planification expérimentale présente une méthode pour analyser les dispositifs dit en dye-switch.

De par mon positionnement au sein de l'unité mixte INRA/AgroParisTech Mathématiques Informatique Appliquées, j'ai participé à de nombreux projets de génomique et notamment analysé un grand nombre d'expériences de transcriptome avec des puces à ADN. L'objectif du chapitre 6 est de présenter à travers deux exemples de collaboration la façon dont les compétences statistiques sont mobilisées et la plus-value apportée par les statistiques aux projets de génomique. Le premier exemple concerne la planification expérimentale, la normalisation et la recherche de gènes différentiellement exprimés à partir de données issues de puces à ADN. Le deuxième concerne la recherche de gènes différentiellement exprimés dans le cadre d'analyse du traductome chez l'oursin à partir de données de séquençage d'ARN. Les articles publiés qui auront été présentés dans le corps du manuscrit sont disponibles en annexes.

Chapitre 1

Contexte biologique

Sommaire

Ce chapitre présente les notions de biologie et les technologies associées nécessaires à la compréhension des méthodes développées dans les chapitres suivants. Le lecteur intéressé par une présentation plus détaillée des concepts biologiques peut aller lire l'ouvrage de Watson et al. [2009].

1.1 Les gènes

1.1.1 Génotype et phénotype

Un organisme vivant peut se caractériser par son **phénotype**, ensemble de ses caractères observables et par son information génétique ou **génotype**. Le génotype permet le développement et le fonctionnement de chaque être vivant et est contenu dans chacune des cellules sous forme d'ADN ou acide désoxyribonucléique. Le phénotype s'observe souvent à un niveau macroscopique, mais s'exprime à un niveau cellulaire et peut se repérer à un niveau moléculaire. Le phénotype moléculaire correspond à la présence de protéines spécifiques (phénotype observable au niveau moléculaire). Il dépend en général de l'action de plusieurs gènes et de l'environnement qui agit sur l'expression de ces gènes et sur l'activité des protéines codées par ces gènes. Et une des grandes questions en biologie est de comprendre les relations entre génotype, phénotype et environnement.

1.1.2 ADN et gène

L'**ADN** est une longue molécule formée d'une succession de nucléotides au nombre de 4 (adénine, cytosine, guanine et thymine) aussi appelés bases. Il se présente sous la forme d'une double hélice qui est l'assemblage de deux brins d'ADN complémentaires. Chaque brin d'ADN correspond à une succession ordonnée de nucléotides nommée séquence. Dans la cellule, l'ADN est associé à des protéines. Chaque molécule d'ADN accompagnée de ses protéines correspond à un chromosome. Le chromosome confère une organisation générale particulière à chaque molécule d'ADN. Chaque **gène** occupe une position spécifique sur la molécule d'ADN d'un chromosome et le gène constitue l'unité physique fondamentale de l'hérédité dont l'existence peut être confirmée par des variants alléliques [Duggan et al., 1999]. Les gènes peuvent être exprimés à travers la transcription en acide ribonucléique (ARN). On appelle région codante d'un gène, la région qui code pour une protéine.

1.2 De la génomique à la métagénomique

1.2.1 La génomique

La **génomique** consiste à étudier le fonctionnement d'un organisme, d'un organe à l'échelle du génome et non d'un seul gène. Une première étape est l'annotation d'un génome afin d'identifier les séquences informatives et la seconde consiste à rechercher la fonction et l'expression des séquences géniques (génomique fonctionnelle). Un génome permet de caractériser un organisme non seulement en tant qu'espèce mais également en tant qu'individu. Tous les individus d'une même espèce partagent en effet des portions de génome identiques, mais d'autres régions sont spécifiques à chaque individu.

1.2.2 Les SNPs (Single Nucleotide Polymorphism)

Les *SNP*s correspondent à un polymorphisme de l'ADN dans lequel il existe une variation d'une paire de bases au sein d'un segment d'un chromosome entre deux individus de la même espèce. Les SNPs sont très courants et représentent la grande partie de la variabilité génétique humaine. Même lorsqu'ils sont situés dans des gènes, ils peuvent être sans effet. En effet, du fait de la redondance du code génétique, ils peuvent ne pas modifier la séquence protéique produite. Cependant dans le cas contraire, ils peuvent être marqueurs de certaines pathologies. C'est pourquoi leur étude est d'un grand intérêt.

1.2.3 Les variants structuraux

Une variation dans la structure d'un chromosome d'un organisme est ce qu'on appelle est un variant structural. Elle est plus grande qu'un SNP, typiquement de 1 kilobase (kb) à 3 mégabases (Mb). Elle peut être de différente nature : variation en nombre de copies (insertions, délétions, duplications), inversions ou translocations. Certains variants structuraux peuvent être associés à des maladies génétiques ou pas. Les variations en nombre de copies (CNV pour *Copy Number Variation*) peuvent être ou non dans des régions codantes.

1.2.4 La métagénomique

La **métagénomique** définit l'identité d'un groupe d'individus, généralement d'espèces différentes, présents dans un environnement complexe, tel que le sol, l'océan, l'intestin d'un homme par exemple. Elle consiste à étudier le matériel génétique directement à partir d'un échantillon environnemental. L'étude se situe à l'échelle de l'ensemble des organismes présents dans cet échantillon et non plus à l'échelle d'un seul organisme sans qu'il soit nécessaire d'isoler individuellement ni de cultiver ces organismes.
On distingue la métagénomique 'ciblée' de la métagénomique dite 'shotgun'. La première

permet d'obtenir un profil de distribution taxonomique via une amplification PCR et séquençage d'un gène marqueur conservé tel que l'ARN ribosomique 16S pour les bactéries ou archées, ou le 18S pour les eucaryotes. La deuxième ne cible pas de séquence génomique particulière et permet donc d'avoir tous les gènes. A noter que l'utilisation du terme métagénomique ciblée ou basée sur des amplicons peut être considérée comme abusive. Creer et al. [2016] conseille de préférer les termes de méta-codebarres (metabarcoding en Anglais) ou de séquençage d'amplicons de gènes marqueurs et définit la métagénomique comme le seul séquençage 'shotgun' de l'ADN total issu d'échantillons environnementaux, permettant ainsi le séquençage d'organismes non cultivables. Lors d'un séquençage 'shotgun', l'ADN est fragmenté en petits segments, qui sont ensuite individuellement séquençés avant d'être réassemblés en des séquences plus longues. J'utiliserai la terminologie proposée par Creer et al. [2016] dans la suite du manuscrit.
Le **microbiome** est défini par McMurdie and Holmes [2014] comme l'écosystème composé des microorganismes vivant dans un environnement donné.

1.3 L'expression des gènes

Le génome comprend des gènes ou des fragments de séquences qui peuvent ou non être transcrits en ARN. Les ARN peuvent être de différentes natures : non codants, de transfert, ribosomique, codants. Les ARN codants sont aussi appelés ARN messagers et sont traduits en protéines par des ribosomes. Les ARN, les non-codants en particulier, sont impliqués dans la régulation de l'expression de gènes ou régions génomiques dont la fonction reste souvent inconnue.

La transcription et l'expression des gènes varient en fonction des conditions physiologiques de l'individu et de l'environnement dans lequel il se trouve.

1.3.1 La transcription

La **transcriptomique** s'intéresse à l'étude de l'expression des gènes du génome d'un organisme à un instant donné dans une condition donnée. Le processus de transcription est celui par lequel l'information contenue dans la double hélice de l'ADN est copiée en ARN. Le terme transcriptome correspond à l'ensemble des ARN issus de la transcription du génome. Par analogie avec la métagénomique, la métatranscriptomique consiste à étudier l'ensemble des gènes exprimés par un ensemble d'organismes d'un échantillon à un moment donné dans un milieu donné.

1.3.2 La traduction

La **traduction** est considérée comme active lorsque plusieurs ribosomes sont fixés sur le même ARN messager. Cet ensemble est appelé polysome. C'est grâce aux polysomes

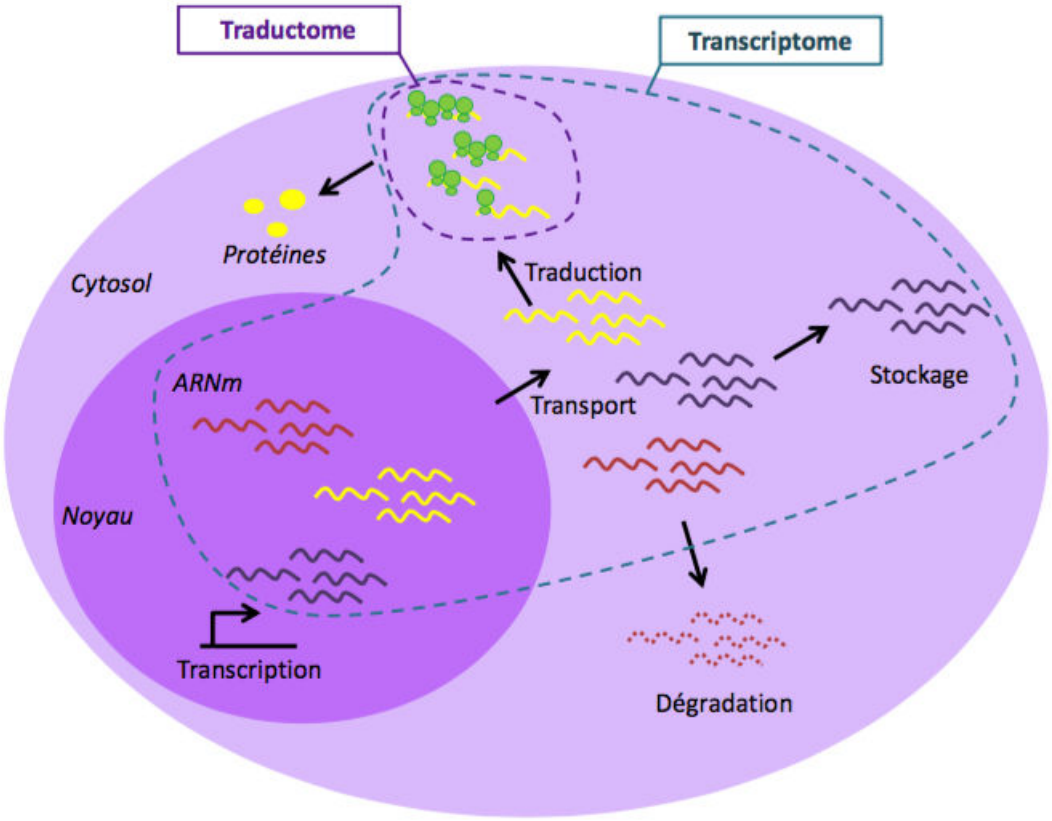


FIGURE 1.1 – Transcriptome versus traductome. L'ensemble des messagers de la cellule appartient au transcriptome cellulaire, tandis que la petite part de messagers recrutés dans les polysomes et traduits appartient au traductome cellulaire. Chassé [2015].

que s'effectue la synthèse des protéines.
On appelle traductome l'ensemble des ARN en cours de traduction. Le protéome quant à lui correspond à l'ensemble des protéines synthétisées. La figure 1.1 représente schématiquement un traductome et un transcriptome.
La caractérisation et la quantification des ARN d'un (méta-)transcriptome ou d'un traductome passent par tout un ensemble d'étapes de biologie moléculaire et de bioinformatique et se fait grâce à des séquenceurs.

1.3.3 La régulation traductionnelle

L'expression de la plupart des gènes est régulée au niveau de la transcription. Cependant il est parfois plus efficace (ajustement plus rapide et précis à des conditions variées) pour la cellule d'opérer une régulation au niveau de la synthèse des protéines. Pour que la synthèse des protéines se passe correctement, trois étapes sont nécessaires :

1. l'initation de la traduction : recrutement du ribosome sur l'ARN messager, insertion de l'ARN de transfert chargé dans le site P du ribosome et positionnement précis du ribosome sur le codon d'initiation ;
2. l'élongation de la traduction, qui correspond à la synthèse du polypeptide ;

3. et la terminaison de la traduction en réponse à des codons-stop.

Dans le cas de la traduction, la régulation se produit le plus souvent au niveau de l'initiation.

Le **traductome** correspond à l'image à un instant donné de l'état de traduction de chaque ARNm au sein de la cellule : ensemble des ARN messagers traduits. Sa mesure donne accès au nombre de ribosomes fixés sur chaque transcrit, descripteur de l'efficacité de la traduction. Son étude est possible par le couplage de méthodes de biologie moléculaire (telle que le polysome ou ribosome profiling) et de techniques de séquençage à haut débit. Une présentation des techniques d'analyse du traductome est présentée dans Chassé et al. [2016].

1.4 Les technologies à haut-débit

Nous présentons, dans cette section, deux techniques permettant la détection et la quantification de profils d'expression à partir d'une expérience typique de transcriptomique à haut-débit afin d'appréhender les caractéristiques des données produites.

1.4.1 Techniques par hybridation

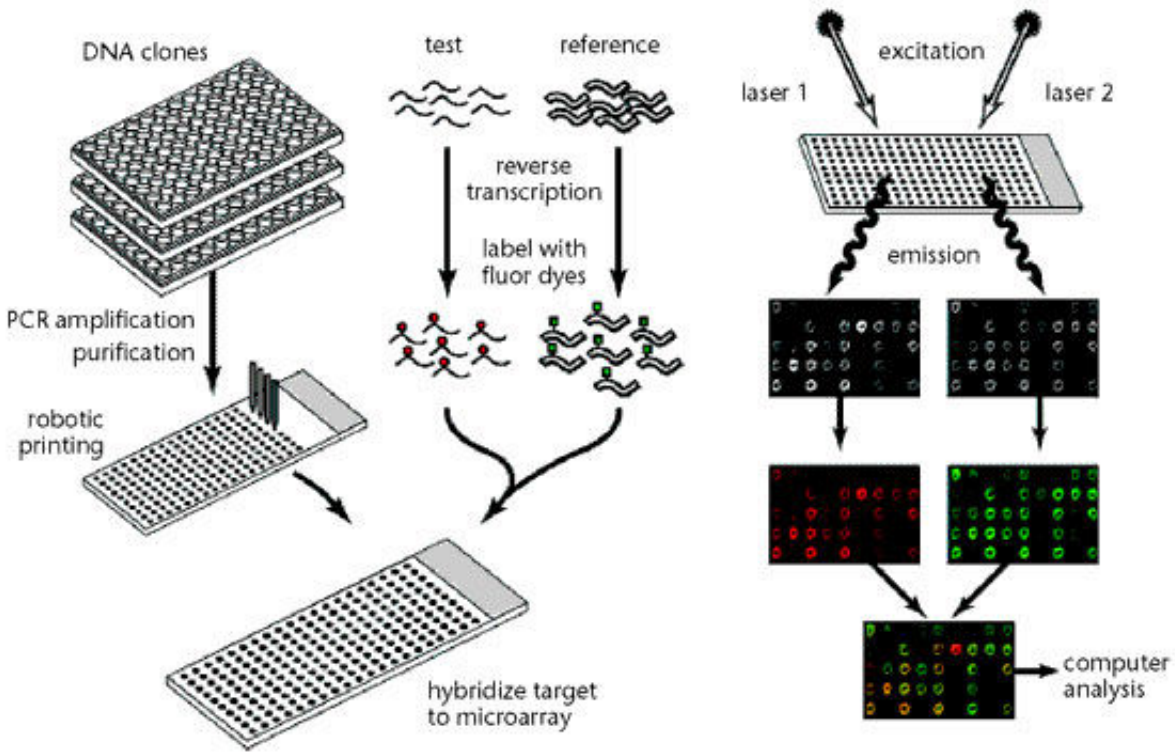


FIGURE 1.2 – Une expérience d'hybridation sur puces à ADN deux couleurs, figure tirée de Duggan et al. [1999]

La figure 1.2 décrit le principe de base d'une expérience par hybridation. La technologie utilisée est celle des puces à ADN. Ces puces sont des supports miniatures (type lames de verre) contenant des molécules d'ADN appelées sondes correspondant à l'organisme d'intérêt et ordonnées sur ce support. L'ARN messager de l'échantillon biologique d'intérêt est extrait, stabilisé par rétro-transcription en ARN complémentaire, marqué à l'aide d'un fluorochrome (ou *dye* en anglais) et déposé sur la puce. S'ensuivent l'étape d'hybridation à proprement parlée et des étapes de nettoyage, séchage pour ne garder que les cibles fixées aux sondes. Le niveau d'hybridation est mesuré pour chaque échantillon biologique à l'aide d'un scanner et d'un logiciel d'analyse d'image après excitation des fluorochromes aux niveaux de longueur d'onde adéquats. On parle de puces à un canal quand un seul fluorochrome est utilisé, à deux canaux quand deux fluorochromes sont utilisés etc. Le signal ainsi obtenu est de nature continu et proportionnel au niveau de transcription des sondes.

1.4.2 Séquençage à haut débit

Les premières techniques de séquençage de l'ADN datent des années 1970s et ont évolué très rapidement. Toutes consistent à définir l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné. La nature des données produites et donc la façon dont elles vont être analysées dépend de la technologie utilisée. Nous renvoyons le lecteur vers Goodwin et al. [2016] pour une revue récente des technologies de séquençage. Les technologies de séquençage à haut débit sont apparues dans les années 2000 et consistent à produire en un *run* de séquençage des milliers de séquences à moindre coût. Alors que le premier séquençage du génome humain terminé en 2003 a duré près de 15 ans et coûté 2 milliards d'euros, en 2007 le génome d'un seul individu (James Watson) était séquencé en 2 mois pour 1,5 million de dollars et en 2013 la société Illumina proposait de séquencer pour 1 000 dollars n'importe quel génome humain en quelques heures. Aujourd'hui, il est possible de caractériser les milliers d'espèces présentes dans un échantillon environnemental particulier. Les données en sortie de séquenceurs correspondent à des lectures (ou *reads*). Une lecture correspond à un ensemble de bases adjacentes séquençées à partir d'une extrémité ou des deux extrémités d'un fragment d'ADN. Lors d'un séquençage de génome ou métagénome, on cherche ensuite à retrouver la ou les séquences originelles dont sont issues ces lectures, soit en alignant ces lectures sur un génome de référence (ou *mapping*) soit en reconstruisant directement ce génome (assemblage *de novo*). Avec ces technologies, il est aussi possible de séquencer des fragments d'ARN transcrits : c'est ce qu'on nomme 'RNA-seq' [Wang et al., 2009]. L'expression des gènes est quantifiée par le nombre de transcrits produits par ces derniers. En séquençant les fragments d'ARN transcrits, et en alignant les lectures issues de ce séquençage sur le génome, on obtient un nombre de lectures alignées sur chacun des gènes, représentant le niveau d'expression de ces derniers. Les données obtenues par ces technologies sont de nature discrète.

Les méthodes d'analyse développées pour les techniques par hybridation ne sont donc plus applicables et de nouvelles méthodes sont nécessaires.

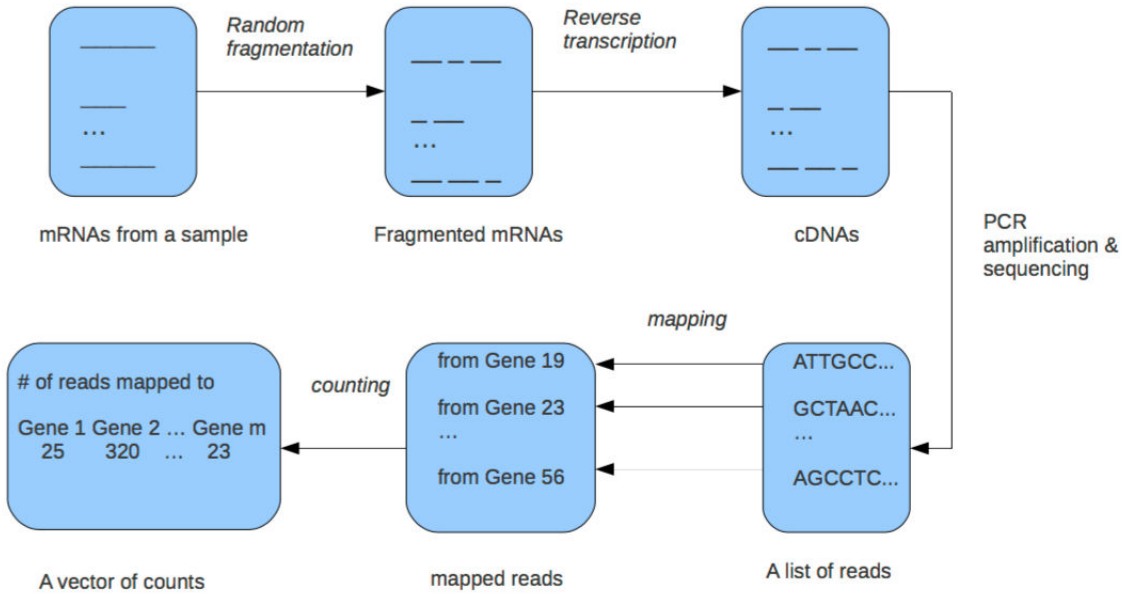


FIGURE 1.3 – Une expérience de séquençage d'ARN du point de vue du statisticien, figure tirée de Li et al. [2012]. Les ARNm sont fragmentés de façon aléatoire. Ces fragments sont rétro-transcrits dans une banque d'ADN complémentaire, elle-même ensuite amplifiée par PCR et séquencée, produisant ainsi une liste de lectures. Ces lectures sont alignées sur un transcriptome connu qui consiste en m gènes. Le nombre de lectures alignées sur chaque gène donne une mesure de l'expression de ce gène. En résumé, le séquençage d'un échantillon aboutit à un vecteur de comptages de longueur m.

La figure 1.3 décrit le processus de production de données issues du séquençage de l'ARN d'un échantillon biologique. Les coûts diminuant, le nombre de séquences obtenues pour un seul échantillon peut être bien plus grand que nécessaire. Dans ce cas, il peut être intéressant de séquencer plusieurs échantillons biologiques en même temps afin de réduire le coût de l'expérience. Ce processus de multiplexage décrit dans la figure 1.4 est possible techniquement grâce à l'utilisation de *barcodes*. Un barcode est une courte séquence d'ADN ajoutée aux lectures d'un échantillon donné pendant l'amplification (étapes A et B). Les séquences ont alors besoin d'être 'démultiplexées', réattribuées à leur échantillon d'origine au cours de traitements ultérieurs (étape C) afin d'être alignées sur la séquence de référence (étape D).

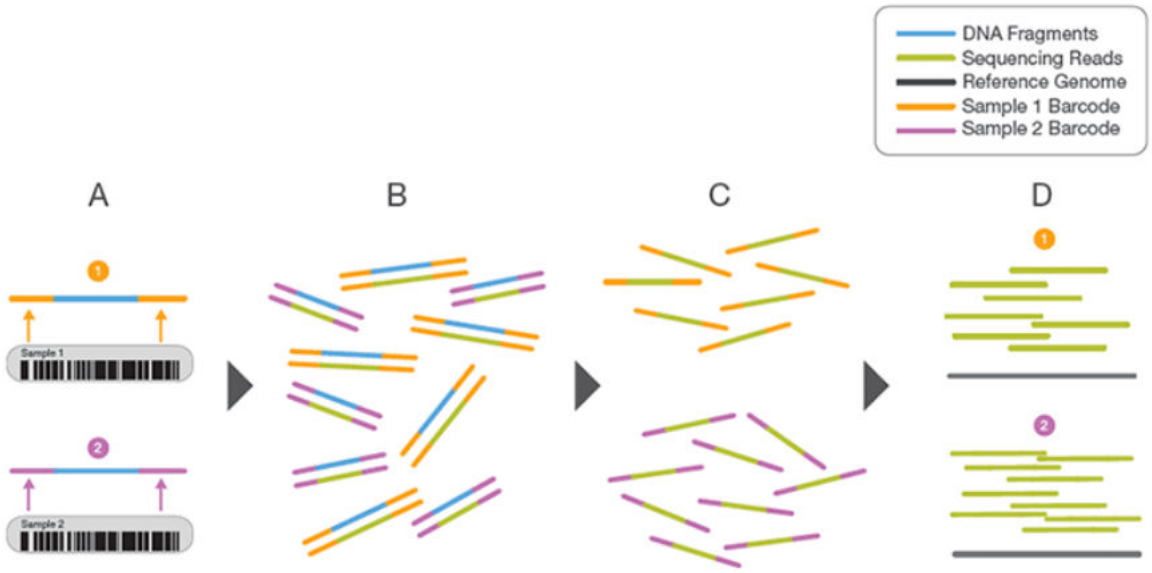


FIGURE 1.4 – Représentation conceptuelle du multiplexage, figure tirée de http://www.illumina.com/technology/next-generation-sequencing/

1.5 Références

H. Chassé. *Régulations traductionnelles de l'embryon précoce d'oursin*. PhD thesis, Ecole Doctorale 515 - Complexité du Vivant, 2015. iii, 5

H. Chassé, S. Boulben, V. Costache, P. Cormier, and J. Morales. Analysis of translation using polysome profiling. *Nucleic Acid Research*, 2016. 6

S. Creer, K. Deiner, S. Frey, D. Porazinska, P. Taberlet, W. K. Thomas, C. Potter, and H. M. Bik. The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, 7(9) :1008–1018, 2016. doi : 10.1111/2041-210X.12574. 4

D. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. Trent. Expression profiling using cdna microarrays. *Nature Genetics*, 21 :10–14, 1999. iii, 2, 6

S. Goodwin, J. McPherson, and R. McCombie. Coming of age : ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17 :333–351, 2016. 7

J. Li, D. Witten, I. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3) :523–538, 2012. iii, 8

P. McMurdie and S. Holmes. Waste not, want not : Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4) :e1003531, 2014. doi : 10.1371/journal.pcbi.1003531. 4

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq : a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10 :57–63, 2009. 7

J. Watson, T. Baker, S. Bell, A. Gann, M. Levine, and R. Losick. *Biologie moléculaire du gène*. Pearson, 2009. 2

Chapitre 2

Préambule statistique

Sommaire

Ce chapitre introduit les concepts statistiques principaux qui seront utiles dans les chapitres suivants. Les sections de ce chapitre sont indépendantes les unes des autres. La première section présente le cadre des modèles de mélange, qui constitue le cadre théorique des méthodes proposées dans le chapitre 3. La deuxième aborde la question de la sélection de modèles. Nous présentons ensuite une vue générale de l'analyse de données de comptage utile pour les chapitres 3 à 6. Ce chapitre se termine par une présentation de la problématique des tests multiples, suivie dans la dernière section d'une présentation de la régression locale mentionnée dans le chapitre 6.

Notations et conventions Ci-dessous quelques notations que nous utiliserons dans la suite du manuscrit :
Y : variables observées
Z : variables non observées (cachées, latentes)
θ : vecteur des paramètres
$p_\theta(.)$ ou $f(.;\theta)$: fonction de densité de probabilité de paramètres θ
$\mathbb{E}_\theta$: moments sous p_θ. Parfois $\mathbb{E}_\theta$ sera remplacé par $\mathbb{E}_p$ ou $\mathbb{E}$
$\sum_i = \sum_{i=1}^n$ etc.
$\sum_{i,j,k,g} = \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^{\mathrm{K}} \sum_{g=1}^{\mathrm{G}}$

Pour les distributions classiques, nous adoptons les notations suivantes :
$\mathscr{U}_{[a,b]}$: distribution uniforme sur l'intervalle $[a,b]$
$\mathscr{M}(n,\pi)$: distribution multinomiale issue de n tirages de vecteur de probabilités $\pi = (\pi_1, \cdots, \pi_{\mathrm{K}})$ avec $\sum_{k=1}^{\mathrm{K}} \pi_k = 1$
$\mathscr{P}(\gamma)$: distribution de Poisson de paramètre γ
$\mathscr{G}(a,b)$: distribution Gamma de paramètre d'échelle a et de paramètre d'intensité b
$\mathscr{NB}(\mu,\phi)$: distribution binomiale négative de moyenne μ et de variance $(\mu + \phi\mu^2)$.

2.1 Modèle à variables latentes

Pour une présentation plus complète des modèles à variables latentes avec des applications en génomique, nous renvoyons le lecteur vers les notes de cours de Stéphane Robin (https://www6.inra.fr/mia-paris/Media/Fichier/PagePerso/StephaneRobin/hiddenstruct-genome-14a).

2.1.1 Variables latentes indépendantes : modèle de mélange

Les modèles de mélange supposent que les observations se répartissent en K classes latentes (ou cachées) telles que la distribution des variables observées Y_i dépend de la

valeur de la variable latente non observée Z_i.

Modèle

Le modèle le plus simple est le suivant :

1. les variables latentes (Z_i) sont indépendantes et identiquement distribuées avec $P(Z_i = k) = \pi_k$,
2. les variables observées (Y_i) sont indépendantes conditionnellement aux variables latentes (Z_i),
3. Y_i sachant $Z_i = k$ suit une distribution paramétrique $F(\gamma_k)$, appelée loi d'émission, de densité de probabilité $f(.;\gamma_k)$.

La fonction de densité d'une observation est parfois notée

$$f(y;\theta) = \sum_{k=1}^{K} \pi_k f_k(y;\gamma_k),$$

où $\pi = (\pi_k)_k$ sont les proportions du mélange avec $\forall k, \pi_k \in [0\,1]$ et $\sum_{k=1}^{K} \pi_k = 1$, f_k est la densité de probabilité de la composante k paramétrée par γ_k et $\theta := (\pi, \gamma) = ((\pi_k), (\gamma_k))$ est le vecteur des paramètres du modèle.

L'objectif de l'inférence de ce type de modèle est de fournir un estimateur des paramètres θ.

Ce modèle revient à écrire :

$$p_\theta(\mathbf{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k)^{Z_{ik}},$$

$$p_\theta(\mathbf{Y}|\mathbf{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(Y_i;\gamma_k)^{Z_{ik}}$$

avec $Z_{ik} = \mathbb{1}\{Z_i = k\}$.

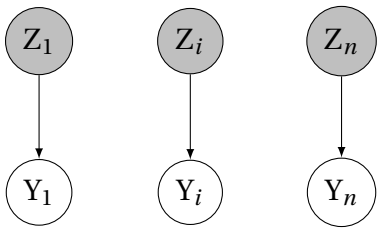


FIGURE 2.1 – Représentation graphique du modèle de mélange correspondant. Légende : les variables observées (cercles sans remplissage), les variables latentes (cercles grisés).

Inférence

La méthode la plus classiquement utilisée pour l'inférence de ce type de modèle repose sur une approche de maximum de vraisemblance. La maximisation directe par rapport à θ de la log-vraisemblance observée $\log p_\theta(\mathbf{Y})$ est difficile et souvent sans solution analytique.

L'algorithme *Expectation Maximization* ou algorithme EM [Dempster et al., 1977; McLachlan and Peel, 2000] permet de maximiser la log-vraisemblance des données $\log p_\theta(\mathbf{Y})$, sans jamais avoir à la calculer et est basé sur une décomposition de la vraisemblance des données dites incomplètes :

$$\log p_\theta(\mathbf{Y}) \quad = \quad \mathbb{E}_\theta\left[\log p_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Y}\right] - \mathbb{E}_\theta\left[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right]. \tag{2.1}$$

Les données observées $(\mathrm{Y}_i)_i$ sont considérées comme incomplètes tant que les variables latentes $(\mathrm{Z}_i)_i$ ne sont pas observées.

Cette décomposition permet de relier la log-vraisemblance incomplète souvent difficilement calculable à la log-vraisemblance complète plus sympathique.

La log-vraisemblance de notre modèle s'écrit

$$\log p_\theta(\mathrm{Y}) \quad = \quad \sum_{i=1}^{n} \log\left[\sum_{k=1}^{\mathrm{K}} \pi_k f(\mathrm{Y}_i;\gamma_k)\right] \tag{2.2}$$

et la log-vraisemblance complète

$$\log p_\theta(\mathrm{Y},\mathrm{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} \mathrm{Z}_{ik}\left[\log\pi_k + \log f(\mathrm{Y}_i;\gamma_k)\right].$$

L'algorithme EM consiste à itérer jusqu'à convergence une étape E d'estimation des moments de la distribution conditionnelle des variables cachées sachant les observations $\mathbb{E}_\theta\left[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right]$ à une étape M de maximisation de l'espérance conditionnelle de la log-vraisemblance complète $\arg\max_\theta \sum_Z p_\theta(\mathrm{Z}|\mathrm{Y})\log p_\theta(\mathrm{Y},\mathrm{Z})$.

La convergence du modèle se vérifie à l'aide d'un critère d'arrêt portant soit sur les valeurs des paramètres soit sur la log-vraisemblance.
On peut montrer que la vraisemblance observée augmente à chaque étape mais il n'y a aucune garantie sur la convergence systématique de l'algorithme vers l'estimateur du maximum de vraisemblance de θ.

Classification

La classification des observations dans des groupes est souvent d'un intérêt majeur lors de l'utilisation des modèles de mélange. L'algorithme EM ne fournit pas à propre-

ment parlé de classification formelle des observations dans des groupes. Cependant les $\tau_{ik} = P(Z_i = k|Y)$ fournissent une mesure de la confiance avec laquelle on peut classer une observation dans un groupe. L'incertitude de classification d'un individu est donnée par l'entropie conditionnelle de Z_i :

$$\mathcal{H}\left[p_\theta(Z_i|Y)\right] = \mathcal{H}\left[p_\theta(Z_i|Y_i)\right] = -\sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}.$$

La règle du Maximum A Posteriori (MAP) peut être utilisée quand il est nécessaire qu'une observation soit affectée à un groupe. Elle consiste à affecter l'observation dans le groupe pour lequel sa probabilité a posteriori est la plus grande.

$$\hat{Z} = \arg\max_z p_\theta(Z = z|Y).$$

Dans le cas du modèle de mélange, $\hat{Z} = (\hat{Z}_i)_i$ avec $\hat{Z}_i = \arg\max_k \tau_{ik}$.

2.1.2 Variables latentes dépendantes : exemple des chaînes de Markov cachées

Un modèle de chaîne de Markov cachée est un modèle de chaîne de Markov MC dont la séquence des états n'est pas directement observée. Chaque état émet des 'observations' qui, elles, sont par définition observables. Il est défini par un état initial de loi ν, des états possibles et des probabilités π de passer d'un état à un autre appelées *transitions*. L'état initial définit les probabilités de commencer dans chacun des états. Les transitions s'opérent entre chaque unité de temps. L'état caché au temps i est donc dépendant de l'état caché au temps $i-1$.

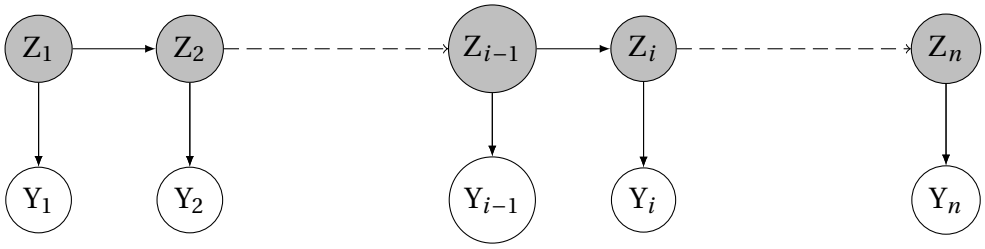


FIGURE 2.2 – Représentation graphique du modèle de Markov caché. Légende : les variables observées (cercles sans remplissage), les variables latentes (cercles grisés)

Une représentation graphique de ce modèle est visible sur la figure 2.2.

Modèle

Un modèle de chaîne de Markov cachée peut s'écrire de la façon suivante :

— $(Z_i)_i \sim MC(\nu, \pi)$,

— $(Y_i)_i$ indépendantes par rapport aux $(Z_i)_i$,

— $Y_i|(Z_i = k) \sim F_k = F(\gamma_k)$.

La chaîne de Markov $MC(\nu, \pi)$ est définie sur l'espace d'états $[1, K]$ avec K le nombre d'états cachés. Les paramètres du modèle sont : $\theta = (\nu, \pi, \gamma)$.

Inférence par maximum de vraisemblance

La décomposition de la vraisemblance (2.1) tient toujours. Dans le cas des chaînes de Markov cachées, la log-vraisemblance complète s'écrit :

$$\begin{aligned} \log p_\theta(Y, Z) &= \log[p_\theta(Z)\, p_\theta(Y|Z)] \\ &= \sum_k Z_{1k} \log \nu_k + \sum_{i,k} Z_{ik} \log f(Y_i; \gamma_k) + \sum_{i \geq 2} \sum_{k,\ell} Z_{i-1,k} Z_{i\ell} \log \pi_{k\ell}. \end{aligned}$$

Remarquons que seul le dernier terme de la log-vraisemblance complète diffère par rapport au modèle de mélange présenté dans la section précédente.

Dans l'étape E de l'algorithme EM, nous avons besoin de calculer $\mathbb{E}(Z_{ik}|Y)$. Contrairement au cas du modèle de mélange précédent, du fait de la structure de dépendance, la distribution conditionnelle n'est plus factorisable : $\mathbb{E}[Z_{ik}|Y] = p(Z_i = k|Y) \neq p(Z_i = k|Y_i)$. Cependant la structure de dépendance conditionnelle reste simple. Et les termes $\tau_{ik} = \mathbb{E}[Z_{ik}|Y]$ et $\mathbb{E}[Z_{i-1,k} Z_{i\ell}|Y]$ peuvent se calculer en itérant des étapes dites Forward et Backward [Baum et al., 1970; Devijver, 1985]. Les τ_{ik} traduisent l'indépendance conditionnelle entre le passé et le futur du processus à chaque temps i. On note $Y_a^n = \{Y_a, \cdots, Y_b\}$, avec $a \leq b$.

$$\begin{aligned} \tau_{ik} = p(Z_i = k|Y) &= \frac{p(Z_i = k, Y_1^n)}{p(Y_1^n)} \\ &= \frac{p(Y_{i+1}^n|Z_i = k)\, p(Z_i = k, Y_1^i)}{p(Y_1^i)\, p(Y_{i+1}^n|Y_1^i)} \\ &= \frac{p(Y_{i+1}^n|Z_i = k)}{p(Y_{i+1}^n|Y_1^i)} \times p(Z_i = k|Y_1^i). \end{aligned}$$

Le premier terme est celui calculé dans l'étape Backward et le second dans l'étape Forward.

— **Forward :**

Pour $i = 1$,

$$F_{1\ell} = p(Z_i = \ell|Y_1) = \frac{\nu_\ell f_\ell(Y_1; \gamma_\ell)}{\sum_\ell \nu_\ell f_\ell(Y_1; \gamma_\ell)},$$

et $\forall\, i \neq 1$,

$$F_{i\ell} \propto \sum_k \pi_{k\ell} F_{i-1,k} f_\ell(Y_i; \gamma_\ell).$$

— **Backward :**

Pour $i = n$,

$$\tau_{nk} = F_{nk} = p(Z_n = k|Y_1^n)$$

et $\forall i \neq n$,

$$\tau_{ik} = F_{ik} \sum_{\ell} \pi_{k\ell} \tau_{i+1,\ell} G_{n+1,\ell},$$

avec $G_{i+1,\ell} = p(Z_{i+1,\ell}|Y_1^i) = \sum_k \pi_{k\ell} F_{ik}$.

Les formules de l'étape Forward montre que $p(Z_i = k|Y_1^i)$ dépend seulement de Y_i et de $F_{i-1,k}$ ce qui signifie que conditionnellement à Y_1^i et Z_{i-1}, Z_i est indépendant de Z_1^{i-2}. La distribution conditionnelle des états cachés Z conditionnellement aux observations Y est alors encore une chaîne de Markov.

L'étape M de maximisation des paramètres est similaire aux modèles de mélange.

Classification : algorithme de Viterbi

Les variables latentes sont souvent très intéressantes en pratique du fait de leur signification. Dans le cas des chaînes de Markov cachées, il est intéressant de déterminer la suite des états cachés la plus probable. Lorsque les variables latentes ne sont pas indépendantes, le MAP (Maximum A Posteriori) joint ne correspond pas au MAP marginal. Et la suite peut être trouvée grâce à l'algorithme de Viterbi [Viterbi, 1967] qui alterne les deux étapes ci-dessous :

— Forward : $V_{1k} = \nu_k f_k(Y_1)$ et pour tout $i \geq 2$:

$$V_{i\ell} = \max_k V_{i-1,k} \pi_{k\ell} f_\ell(Y_{1i}),$$
$$S_{i-1}(\ell) = \arg\max_k V_{i-1,k} \pi_{k\ell} f_l(Y_{1i}),$$

— Backward : $\hat{Z}_n = \arg\max_k V_{nk}$ et pour tout $i > n$:

$$\hat{Z} = S_i(\hat{Z}_{i+1})$$

.

2.1.3 Structure de dépendance plus complexe : approche variationnelle

L'interprétation variationnelle de l'algorithme EM est présentée dans Neal and Hinton [1999] et Bishop [2006]. Ormerod and Wand [2010] proposent une présentation des méthodes variationnelles pour les statisticiens. Et une présentation plus complète est disponible dans Ghahramani and Beal [2000]. Les approches variationnelles peuvent être utiles à la fois en inférence bayésienne et en inférence fréquentiste, quand la spécification de la vraisemblance nécessite le conditionnement par rapport à un vecteur de variables la-

tentes.

Interprétation variationnelle de l'algorithme EM

Calcul variationnel Le terme de variationnel vient du calcul des variations (ou calcul variationnel) qui regroupe un ensemble de méthodes permettant de minimiser une fonctionnelle. Une fonctionnelle peut être définie comme un opérateur $\mathscr{F}[q]$ qui prend en entrée une fonction q et renvoit un nombre. Le calcul variationnel consiste à trouver la valeur de q qui maximise ou minimise $\mathscr{F}[q]$. La fonction q qui minimise la fonctionnelle $\mathscr{F}[q] = \int \mathrm{L}(x, q(x))\,dx$ satisfait l'équation différentielle d'Euler-Lagrange $\frac{\partial}{\partial q(x)}\mathrm{L}(x, q(x)) = 0$.

Définition d'une borne inférieure de la log-vraisemblance En utilisant l'inégalité de Jensen, on peut introduire dans l'expression de la log-vraisemblance (2.2) une fonction des variables cachées $q(\mathrm{Z})$ et décomposer la log-vraisemblance de la façon suivante :

$$\begin{aligned}
\log p_\theta(\mathrm{Y}) &= \log \sum_{\mathrm{Z}} p_\theta(\mathrm{Y},\mathrm{Z}) \\
&= \log \sum_{\mathrm{Z}} \left(\frac{p_\theta(\mathrm{Y},\mathrm{Z})}{q(\mathrm{Z})} \right) q(\mathrm{Z}) \\
&\geq \sum_{\mathrm{Z}} q(\mathrm{Z}) \log \frac{p_\theta(\mathrm{Y},\mathrm{Z})}{q(\mathrm{Z})} \\
&= \sum_{\mathrm{Z}} q(\mathrm{Z}) \log p_\theta(\mathrm{Y},\mathrm{Z}) - \sum_{\mathrm{Z}} q(\mathrm{Z}) \log q(\mathrm{Z}) \\
&= \mathscr{J}(q,\theta).
\end{aligned} \tag{2.3}$$

$\mathscr{J}(q,\theta)$ est une fonctionnelle qui constitue une borne inférieure de la log-vraisemblance des données. Cette fonctionnelle est parfois appelée *énergie libre*. Le terme $-\sum_{\mathrm{Z}} q(\mathrm{Z}) \log q(\mathrm{Z})$ que l'on note aussi $\mathscr{H}(q(\mathrm{Z}))$correspond à l'entropie de Shannon de la fonction q.
L'algorithme EM peut alors être vu comme un algorithme itérant deux étapes de maximisation de $\mathscr{J}(q,\theta)$ la première par rapport à la fonction q, la deuxième par rapport aux vecteurs des paramètres θ. A noter que l'entropie de $q(\mathrm{Z})$ ne dépend pas de θ et que l'étape M consiste juste à maximiser l'espérance conditionnelle de la log-vraisemblance complète.

Divergence de Kullback-Leibler

La divergence de Kullback-Leibler ou entropie relative est une mesure de dissimilarité entre deux distributions de probabilité p et q. Elle se définit par

$$\mathscr{D}_{\mathrm{KL}}(p(\theta)||q(\theta) = \int_\Theta \log\left(\frac{p(\theta)}{q(\theta)}\right) p(\theta)\,d\theta.$$

$\mathscr{D}_{\mathrm{KL}}(p(\theta)||q(\theta)$ peut aussi s'écrire comme $\mathbb{E}_{p(\theta)}\left[\log\left(\frac{p(\theta)}{q(\theta)}\right)\right]$. En cas de distributions p et q discrètes, l'intégrale se simplifie en somme. Elle a comme propriété d'être toujours positive. Elle s'annule si et seulement si $p = q$.
La différence entre la borne inférieure $\mathscr{J}(q,\theta)$ (2.3) et la log-vraisemblance $\log p_\theta(\mathrm{Y})$ s'écrit :

$$\begin{aligned}\log p_\theta(\mathrm{Y}) - \mathscr{J}(q,\theta) &= \log \sum_{\mathrm{Z}} p_\theta(\mathrm{Y},\mathrm{Z}) - \mathscr{J}(q,\theta) \\ &= \log \sum_{\mathrm{Z}} p_\theta(\mathrm{Y},\mathrm{Z}) - \sum_{\mathrm{Z}} q(\mathrm{Z}) \log \frac{q(\mathrm{Z})}{p_\theta(\mathrm{Y},\mathrm{Z})}\end{aligned} \tag{2.4}$$

Le premier terme correspond à log-vraisemblance jointe des données et le deuxième à la divergence de Kullback-Leibler entre la fonction $q(\mathrm{Z})$ et la fonction $p_\theta(\mathrm{Y},\mathrm{Z})$. Elle s'annule pour $q(\mathrm{Z}) = p_\theta(\mathrm{Y},\mathrm{Z})$ ce qui correspond exactement à l'étape E de l'algorithme EM.

Inférence approchée

Quand l'inférence exacte n'est pas faisable du fait d'une dépendance introduite par des variables latentes par exemple, ou par des distributions compliquées, des techniques approchées d'inférence peuvent être utilisées. Les approches variationnelles utilisent l'interprétation variationnelle de l'algorithme EM et contraignent q à avoir une forme permettant les calculs, c'est-à-dire à être factorisable. Il suffit alors de maximiser la borne inférieure $\mathscr{J}(q,\theta)$ (2.3) sous cette contrainte.

Restriction à la classe des fonctions produits Cette restriction donne des solutions explicites pour chaque composante du produit en fonction des autres, ce qui permet une résolution du problème d'optimisation à l'aide d'un algorithme itératif. Cette classe de fonction donne naissance aux approximations de type 'champs moyen'.

Algorithme VEM (Variational EM) L'algorithme VEM consiste à itérer une étape E variationnelle et une étape M dont les objectifs sont précisés ci-dessous :

- **Etape VE :** Maximiser $\mathscr{J}(q,\theta)$ par rapport à q avec θ fixé sous la contrainte de factorisation de q. Cela revient à minimiser $\mathrm{KL}(q||p_\theta)$.
- **Etape M :** Maximiser $\mathscr{J}(q,\theta)$ par rapport à θ avec q fixé.

La borne inférieure augmente à chaque itération et l'algorithme aboutit à un maximum (souvent local).

2.2 Sélection de modèle

Dans le cadre des modèles de mélange, une des questions difficiles est le choix du nombre de composantes. Ce nombre de classes K peut être dicté par la question posée. Cependant il est le plus généralement inconnu et choisir le nombre de classes fait partie

de la question. Les critères les plus usuels sont des critères de vraisemblance pénalisée de la forme

$$\log p_{\hat{\theta}_K}(Y) - \beta \text{pen}(K),$$

où β est une constante positive et pen(K) est une fonction de pénalité. L'objectif d'un critère de sélection de modèle est de permettre de choisir un modèle assurant une bonne adéquation aux données (contrôlée par le terme de log-vraisemblance) tout en étant parcimonieux (rôle de la pénalité dépendant de la dimension du modèle). Sélectionner un modèle dans une collection de modèles $(M_i)_i$ revient à sélectionner le nombre de classes ou composantes à partir des données via la minimisation d'un critère pénalisé. Lebarbier and Mary-Huard [2006] présentent de façon détaillée les fondements et interprétation du critère BIC, un des critères les plus couramment utilisés.

2.2.1 Le critère BIC

Le critère BIC (Bayesian Information Criterion) [Schwarz, 1978] est une approximation du calcul de la vraisemblance des données conditionnellement au modèle fixé. Il s'agit d'une approche bayésienne de sélection de modèle, c'est-à-dire que les paramètres et les modèles sont considérés comme des variables aléatoires munies d'une distribution a priori. Le choix se fait en fonction de la probabilité a posteriori d'un modèle donné M_i. Le critère BIC sélectionne le modèle qui maximise cette probabilité :

$$M_{BIC} = \arg\max_{M_i} P(M_i|Y).$$

Le critère est basé sur une distribution a priori pour le nombre de classes, une distribution conditionnelle des paramètres sachant le nombre de classes et de la distribution conditionnelle des observations sachant les paramètres.

Par la formule de Bayes, on a $M_i : P(M_i|Y) \propto P(M_i)P(Y|M_i)$. Supposons que l'a priori sur les modèles soit uniforme, il reste à calculer $P(Y|M_i)$. Le calcul exact est rarement possible. $P(Y|M_i)$ peut être approché via l'utilisation d'une approximation de Laplace (voir Lebarbier and Mary-Huard [2006] pour une démonstration détaillée).

Proposition 2.2.1 *Approximation de Laplace*
Soit une fonction $L : \mathbf{R}^d \to \mathbf{R}$ *telle que* L *est deux fois différentiable sur* $\mathbf{R}^d$ *et atteint un unique maximum sur* $\mathbf{R}^d$ *en* u^*, *alors*

$$\int_{\mathbf{R}^d} e^{nL(u)} = e^{nL(u^*)} \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} |-L''(u^*)|^{-\frac{1}{2}} + \mathcal{O}^{n-1}$$

Il en découle la proposition suivante :

Proposition 2.2.2 *Sous des conditions de régularité,*

$$\log p(\mathrm{Y}|\mathrm{K}) = \log p_{\hat{\theta}_\mathrm{K}}(\mathrm{Y}) - \frac{d_\mathrm{K}}{2}\log n + \mathcal{O}(1)$$

où d_K correspond au nombre de paramètres indépendants du modèle à K *composantes.*

Le critère BIC sélectionne le modèle à $\hat{\mathrm{K}}_{\mathrm{BIC}}$ composantes tel que :

$$\hat{\mathrm{K}}_{\mathrm{BIC}} = \arg\max_{\mathrm{K}} \log p_{\hat{\theta}_\mathrm{K}}(\mathrm{Y}) - \frac{d_\mathrm{K}}{2}\log n.$$

2.2.2 Le critère ICL

Le critère ICL proposé par Biernacki et al. [2000] permet de tenir compte de l'objectif de classification. Ce critère pénalise non plus la log-vraisemblance des données observées, mais la log-vraisemblance des données complétée par les variables latentes. Il se définit comme :

$$\mathrm{ICL}(m,\mathrm{K}) = \log p_{\hat{\theta}_\mathrm{K}}(\mathrm{Y},\mathrm{Z})] - \frac{d_\mathrm{K}}{2}\log n.$$

Z étant inconnu, les auteurs proposent de remplacer Z par $\hat{\mathrm{Z}} = \mathrm{MAP}(\hat{\theta})$. Z peut aussi être remplacée par son espérance conditionnelle.
Le critère ICL peut également s'écrire comme le critère BIC pénalisé par un terme d'entropie tenant compte de l'incertitude de classification. Ce terme d'entropie $-\mathbb{E}[\log p(\mathrm{Z}|\mathrm{Y})]$ est petit quand l'observation est classée avec une grande confiance. Le modèle sélectionné est finalement le modèle à $\hat{\mathrm{K}}_{\mathrm{ICL}}$ composantes tel que :

$$\hat{\mathrm{K}}_{\mathrm{ICL}} = \arg\max_{\mathrm{K}} \left(\mathbb{E}_{\hat{\theta}_\mathrm{K}}[\log p_{\hat{\theta}_\mathrm{K}}(\mathrm{Y},\mathrm{Z})|\mathrm{Y}] - \frac{d_\mathrm{K}}{2}\log n \right).$$

2.3 Données de comptage

2.3.1 Introduction

Une partie de cette section s'inspire du livre de Cameron and Trivedi [2013] auquel le lecteur intéressé par plus de détails peut se référer.

Le bruit observé dans les données issues de séquençage à haut débit (chapitre 1 section 1.4) peuvent provenir de trois sources différentes : le bruit inévitable, inhérent au processus de comptage (dominant pour les gènes faiblement exprimés), le bruit technique, issu de la préparation de l'échantillon, que l'on espère négligeable et le bruit biologique. Les répétitions techniques [Marioni et al., 2008] semblent satisfaire l'hypothèse d'équidispersion du modèle de Poisson (2.3.2), ce qui n'est pas le cas de répétitions biologiques [Robinson and Smyth, 2008].

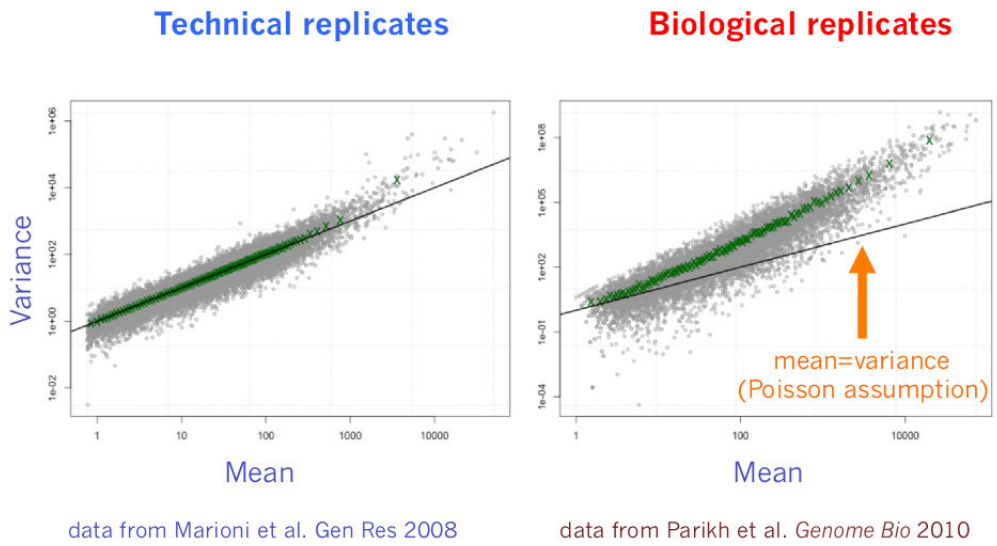


FIGURE 2.3 – Relation moyenne-variance pour deux répétitions techniques (à gauche) ; pour deux répétitions biologiques (à droite). Figure adaptée d'une présentation de D. Robinson.

Surdispersion On parle de surdispersion quand la variance excède l'espérance et de sous-dispersion dans le cas contraire. Le phénomène de surdispersion est assez courant notamment dans le domaine de l'assurance pour le nombre de sinistres attribués à une police d'assurance [de Jong and Heller, 2008] ou dans celui de la biologie. C'est le cas par exemple en écologie pour l'étude du nombre d'individus dans une niche particulière : nombre de mites rouges sur les feuilles de pommiers [Bliss and Fisher, 1953], en transcriptomique : nombre de courtes séquences alignées sur une région génomique d'intérêt [Robinson and Smyth, 2008]. Ce phénomène peut être causé par l'oubli d'une variable explicative importante dans le modèle ou par une corrélation entre les traitements étudiés. Il entraîne une surestimation de la précision des estimateurs des paramètres du modèle et peut avoir pour conséquence un nombre accru de faux positifs dans le cas d'analyse différentielle, la sélection de modèles trop complexes dans le cadre de la sélection de modèles [Richards, 2008].

2.3.2 Rappel sur les lois

Loi de Poisson

Une variable aléatoire discrète Y est distribuée selon une loi de Poisson de paramètre γ, $\gamma > 0$ si elle vérifie :

$$f(y) := \mathrm{P}(\mathrm{Y} = y) = \frac{e^{-\gamma}\gamma^{y}}{y!} \tag{2.5}$$

pour tout $y \in \mathbb{N}^{+}$. La loi de Poisson a un seul paramètre γ et a pour propriété fondamentale la propriété d'équidispersion :

$$\mathbb{E}(\mathrm{Y}) = \mathbb{V}(\mathrm{Y}) = \gamma. \tag{2.6}$$

Fonction et loi Gamma

La fonction Gamma La fonction Gamma est une extension de la fonction factorielle aux nombres réels et complexes. La fonction Gamma, notée $\Gamma(a)$ est définie par $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$, avec $a \geq 0$. Si a est un entier, alors $\Gamma(a) = (a-1)!$.
Elle admet la propriété suivante $\frac{\Gamma(y+a)}{\Gamma(a)} = \prod_{j=0}^{y-1}(j+a)$ si y est un entier.

Proposition 2.3.1 *Formule asymptotique de Stirling pour la fonction Gamma*

$$\Gamma(a) \sim a^{a-\frac{1}{2}}e^{-a}\sqrt{2\pi}, \quad |\arg(a)| < \pi.$$

La dérivée du logarithme de la fonction *gamma* est appelée la fonction *digamma* :

$$\frac{\partial \log\Gamma(a)}{\partial a} = \psi(a).$$

La fonction *digamma* suit la récurrence suivante :

$$\begin{aligned} \psi(a+1) &= \psi(a) + \frac{1}{a} \\ \psi^{(j)}(a+1) &= (-1)^j j!(a)^{-j-1} + \psi^{(j)}(a) \end{aligned}$$

où $\psi^{(j)}(a)$ est la dérivée d'ordre j de ψ.

La loi Gamma La densité d'une loi $\mathcal{G}(a,b)$ s'écrit

$$f(y) = \frac{e^{-by}y^{a-1}b^a}{\Gamma(a)} \tag{2.7}$$

Si $Y \sim \mathcal{G}(a,b)$, alors $\mathbb{E}(Y) = \frac{a}{b}$, et $\mathbb{V}(Y) = \frac{a}{b^2}$. Les paramètres a et b sont respectivement appelés paramètre de forme et paramètre d'échelle.
A noter que $\mathbb{E}(\log(Y)) = \psi(a) - \log b$. L'entropie d'une $\mathcal{G}(a,b)$ s'écrit

$$a - \log b + \log\Gamma(a) + (1-a)\psi(a).$$

Il existe une autre paramétrisation de la loi Gamma qui consiste à prendre $\frac{1}{b}$ au lieu de b.

Loi mélange de Poisson

Une approche classique pour prendre en compte le phénomène de surdispersion consiste à utiliser un modèle de mélange de Poisson continu [McLachlan and Peel, 2000]. On introduit un terme d'hétérogénéité u positif dans le paramètre de la loi de Poisson qui devient γu où γ est un paramètre inconnu et u est la réalisation d'une variable aléatoire U suivant une certaine loi G. La distribution de Y se modélise alors comme

$$f(y) = \int_0^\infty \frac{e^{-(\gamma u)}(\gamma u)^y}{y!} dG(u). \tag{2.8}$$

En posant $\mathbb{E}(U) = 1$, la moyenne reste γ tandis que la variance devient plus grande que γ, ce qui permet d'obtenir des données surdispersées :

$$\begin{aligned} \mathbb{V}(Y) &= \mathbb{V}_u(\mathbb{E}(Y|U)) + \mathbb{E}_u(\mathbb{V}(Y|U)) \\ &= \mathbb{V}_u(\gamma U) + \mathbb{E}_u(\gamma U) \\ &= \gamma^2 \mathbb{V}_u(U) + \gamma. \end{aligned} \tag{2.9}$$

Le théorème dit "Two crossings theorem" [Shaked, 1980] montre que pour Y, variable aléatoire, continue ou discrète de loi $f(y|\gamma, u)$ dans la famille exponentielle, avec $\mathbb{E}(U) = 1$, alors la différence $g(y|\gamma) = E_u[f(y|\gamma, u)] - f(y|\gamma, u)$ change de signe exactement deux fois selon le profil $\{+;-;+\}$.
Cela revient à dire que, pour une même moyenne, toute distribution marginale doit croiser deux fois la distribution conditionnelle originelle, une fois en amont et une fois en aval ce qui entraîne qu'une loi mélange de Poisson a une queue plus lourde qu'une loi Poisson de même moyenne.
La figure 2.4 illustre ce théorème avec $g(.)$ densité d'une loi Gamma-Poisson et $f(.)$ densité d'une loi de Poisson de même moyenne, et le fait que les lois de mélange permettent de modéliser plus facilement des excès de zéro, une queue plus épaisse à droite et donc le phénomène de surdispersion.

Distribution Gamma-Poisson

Un choix classique pour $G(u)$ dans (2.8) est la loi $\mathcal{G}(a, a)$ définie par (2.7).

Si Y est distribuée selon une loi de Gamma-Poisson définie comme ci-dessus alors $f(y;\gamma,u) = P(Y = y)$ s'écrit :

$$f(y;\gamma,u) = \int_0^\infty \frac{e^{-\gamma u}(\gamma u)^y}{y!} g(u)\,du \tag{2.10}$$

$$= \frac{\gamma^y}{\Gamma(y+1)} \frac{a^a}{\Gamma(a)} \int_0^\infty e^{-(a+\gamma)u} u^{(y+a-1)}\,du \tag{2.11}$$

$$\tag{2.12}$$

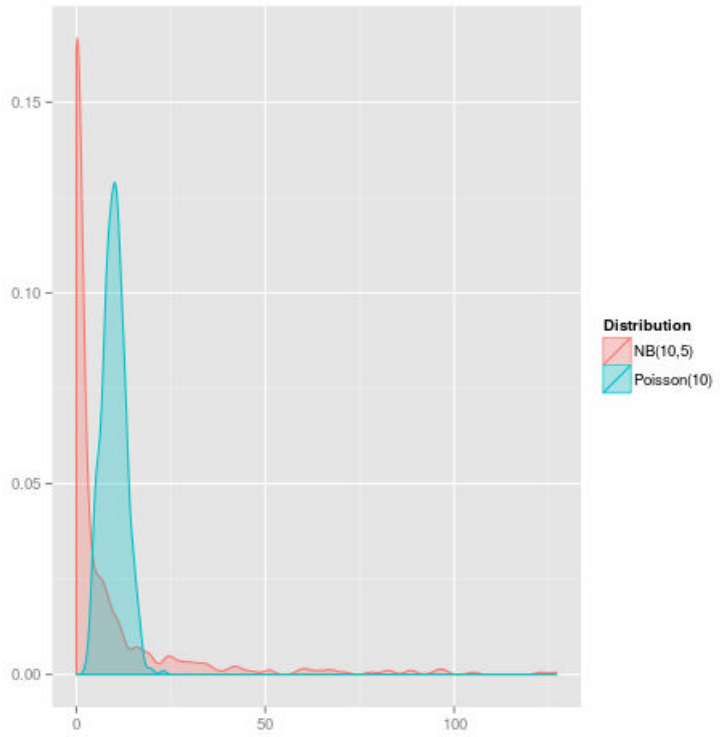


FIGURE 2.4 – 'Two crossing theorem' : densités d'une loi de Poisson $\mathscr{P}(10)$ et d'une loi binomiale négative $\mathscr{NB}(10, 0.2)$

En effectuant un changement de paramètres on retrouve une loi binomiale négative.

$$\begin{aligned} f(y;\gamma,a) &= \frac{\gamma^y}{\Gamma(y+1)}\frac{a^a}{\Gamma(a)}\frac{\Gamma(y+a)}{(a+\gamma)^{y+a}} \\ &= \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}\left(\frac{\gamma}{(a+\gamma)}\right)^y\left(\frac{a}{a+\gamma}\right)^a \\ &= \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}\left(1-\frac{a}{a+\gamma}\right)^y\left(1-\frac{\gamma}{a+\gamma}\right)^a \end{aligned}$$

On retrouve alors la formule de la densité d'une loi binomiale négative (2.13) de moyenne γ et de variance $\gamma(1+\phi\gamma)$ avec $\phi = \frac{1}{a}$. Quand $\phi = 0$, on retrouve la distribution de Poisson.

L'interprétation de la loi binomiale négative découlant d'une loi Gamma-Poisson est un résultat ancien de Greenwood and Yule [1920]. Cameron and Trivedi [1986] ont proposé une paramétrisation plus précise permettant d'aboutir à différentes fonctions de variance dont la fonction de variance $\mathbb{V}(Y) = \gamma + \phi\gamma^2$, dite NB2.

Loi binomiale négative

Hilbe [2011] présente et discute de la plupart des différentes paramétrisations possibles. Nous utiliserons tout au long de ce manuscrit la paramétrisation précédemment présentée qui suppose $\mathbb{V}(Y) = \gamma + \phi\gamma^p$ où ϕ correspond à un paramètre de surdispersion, et $p = 2$.

Si $Y \sim \mathcal{NB}(\gamma, \phi)$ alors

$$f(y;\gamma,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(y+1)\Gamma(\phi^{-1})}\left(\frac{\phi^{-1}}{\phi^{-1}+\gamma}\right)^{\phi^{-1}}\left(\frac{\gamma}{\phi^{-1}+\gamma}\right)^{y}, \tag{2.13}$$

avec $\phi > 0$ et $y \in \mathbb{N}$.
La loi binomiale négative fait partie des Familles Exponentielles Naturelles à deux paramètres dont un de dispersion, qui sont aussi appelées Familles Exponentielles de Dispersion [Jø rgensen, 1997].

A noter que la transformation $\tilde{Y} = \sqrt{\phi}\sinh^{-1}\sqrt{\frac{Y}{\phi}}$ normalise et stabilise la variance des données de telle sorte que $\tilde{Y}$ suit approximativement une loi normale centrée-réduite [Johnson et al., 2005].

Estimation des paramètres A ϕ connu, la loi binomiale négative a la forme d'un modèle linéaire généralisé avec pour fonction de lien canonique

$$\eta(\gamma_i) = \log\frac{\gamma_i}{\gamma_i + \phi^{-1}}$$

et fonction de variance

$$\mathbb{V}(y_i) = \gamma_i + \phi\gamma_i^2.$$

Mais souvent ϕ est inconnu et a besoin d'être estimé. Il existe une dizaine d'estimateurs de ϕ [AlKhasawneh, 2010; Krishna Saha, 2005]. Les plus courants sont celui des moments :

$$\hat{\phi} = \frac{\bar{y}^2}{S_n^2 - \bar{y}}$$

et celui du maximum de vraisemblance solution de :

$$\frac{\partial\mathscr{L}}{\partial\phi} = \sum_i^n\sum_{v=0}^{y_i-1}\frac{1}{\phi(1+\phi v)} - \frac{n}{\phi^{-2}}\log(1+\phi\gamma) + \frac{n\gamma(\bar{y}=\phi^{-1})}{1+\phi\gamma} = 0,$$

où $\bar{y} = \frac{1}{n}\sum_{i=1} n y_i$, $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ et $\mathscr{L}$ est la log-vraisemblance des données observées.
Les deux estimateurs posent problème. L'estimateur des moments est non défini quand la variance est égale à la moyenne, il peut être négatif quand la variance de l'échantillon est plus petit que celui de la moyenne et très très grand, quand la variance d'échantillonnage est légèrement plus grande que la moyenne empirique. La paramétrisation de la variance telle que $\gamma(1+\phi\gamma)$, plutôt que $\gamma(1+1/\phi\gamma)$ permet de contourner le problème de valeurs infinies quand la variance et la moyenne empirique sont égales et peut être calculé numériquement.

Paramétrisation ϕ ou $\frac{1}{\phi}$? Nous utiliserons la paramétrisation telle que $\mathbb{V}(Y) = \gamma + \phi\gamma^2$ permettant un lien direct entre le paramètre γ et le degré de surdispersion des données. Cette paramétrisation directe est la plus utilisée actuellement sauf dans le logiciel R, où la fonction **glm.nb** utilise $\frac{1}{\phi}$. La fonction *rnbinom* de R permet quant à elle de simuler une loi binomiale négative. Deux paramétrisations sont, là encore, disponibles. Celle que nous utiliserons est la paramétrisation *rnbinom(n,size,mu)* où mu est le paramètre de position et $size$ le paramètre de dispersion, avec la variance se définissant comme $mu + \frac{mu^2}{size}$. Pour faire la correspondance avec nos notations, $mu = \gamma$ et $size = \frac{1}{\phi}$.

2.3.3 Interprétation de la paramétrisation NB2 dans le cadre des données de séquençage

La figure 2.4 typiquement obtenue à partir de données de séquençage suggère une dispersion spécifique à chaque gène. McCarthy et al. [2012] ont les premiers proposé l'interprétation dans le cadre des données de séquençage.
Si on suppose que le comptage y_{ij} du gène i dans l'échantillon j suit une distribution binomiale négative de moyenne γ_{ij} et de variance $\gamma_{ij} + \phi_i\gamma_{ij}^2$ avec $\phi_i > 0$, alors le calcul du coefficient de variation s'obtient par la formule :

$$\mathrm{CV} = \frac{\sqrt{\gamma_{ij} + \phi_i\gamma_{ij}^2}}{\gamma_{ij}}.$$

En élevant au carré, on obtient :

$$\mathrm{CV}^2 = \frac{1}{\gamma_{ij}} + \phi_i,$$

ce qui s'interprète comme

$$\mathrm{CV}^2 = \mathrm{CV}^2_{\texttt{Technique}} + \mathrm{CV}^2_{\texttt{Biologique}}$$

Le paramètre de dispersion est alors égal au carré du coefficient de variation biologique. Le coefficient de variation biologique (CVB) représente le coefficient de variation qui resterait entre les répétitions biologiques si la profondeur de séquençage pouvait augmenter indéfiniment. Le coefficient de variation technique, quant à lui, diminue quand la profondeur de séquençage augmente. Le CVB représente la source principale de variation pour les gènes de fort comptage. Il est donc nécessaire de bien estimer le paramètre ϕ.

2.4 Tests multiples

2.4.1 Introduction

Lorsque l'on effectue un test statistique, deux types d'erreur peuvent être faites : l'erreur de type I (ou faux positif) si l'on rejette à tort l'hypothèse nulle testée et l'erreur de type II si l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. Lorsque l'on effectue non pas un mais plusieurs tests simultanément, typiquement dans une expérience de transcriptomique lorsque l'on teste pour chaque gène l'hypothèse nulle d'absence d'association entre un niveau d'expression et une condition biologique, le risque de faire des erreurs de type I augmente avec le nombre de tests effectués. Si dans une expérience dans laquelle aucune différence d'expression n'est attendue, un test est effectué au niveau $\alpha = 0.05$ pour chacun des 10000 gènes, on s'attend à rejeter $10000 * 0.05 = 500$ hypothèses nulles à tort. La définition d'une erreur de type I appropriée ainsi que des procédures adaptées aux tests multiples sont donc nécessaires. Dudoit et al. [2003] présente une revue dans le cadre des puces à ADN, et Roquain [2011] une étude du contrôle de l'erreur de type I dans le cadre des tests multiples.
Quatre étapes sont nécessaires à une analyse comprenant des tests multiples :

1. Choix des hypothèses à tester
2. Choix des tests individuels
3. Choix du critère de type I à contrôler
4. Choix de la procédure pour combiner les tests individuels

2.4.2 Formalisation du problème de tests multiples

Nous reprenons ici la présentation faite par Benjamini and Hochberg [1995] et reprise dans Dudoit et al. [2003]. Supposons que l'on teste m hypothèses nulles H_0^i, $i = 1, \cdots, m$. On suppose que les m hypothèses sont connues à l'avance et que m_0 et $m_1 = m - m_0$, respectivement le nombre d'hypothèses H_0 vraies et fausses sont des paramètres inconnus. On note R le nombre d'hypothèses rejetées à l'issue de la procédure de test.

Nombre de	non rejets	rejets	
H_0 vraies	U	V	m_0
H_0 non vraies	T	S	m_1
	$m - R$	R	m

TABLE 2.1 – Résultats possibles à l'issue d'une procédure de tests multiples comportant m hypothèses testées

L'objectif est de minimiser le nombre V de faux positifs ou erreurs de type I et le nombre T de faux négatifs ou erreurs de type II. Minimiser le nombre T de faux négatifs est équivalent à maximiser la puissance, définie comme la capacité à détecter les vrais

positifs.
L'approche standard au problème de tests multiples consiste à calculer pour chaque gène i une statistique de test puis à appliquer une procédure de test multiple pour déterminer les hypothèses à rejeter en garantissant un taux erreur de type I adapté et préalablement défini.

Probabilité critique ou pvalue Une probabilité critique $p(\mathrm{Y})$ peut être vue comme une normalisation particulière de la statistique de test associée telle que

— sous H_0, $p(\mathrm{Y}) \sim \mathcal{U}_{[0,1]}$
— sous l'alternative, $p(\mathrm{Y})$ est plutôt proche de 0.

On appelle probabilité critique brute (ou non ajustée), notée p_i, la probabilité critique associée au test de l'hypothèse H_0^i.

La probabilité critique ajustée associée au test de l'hypothèse H_0^i, notée $\tilde{p}_i$, est définie comme le niveau nominal de la procédure entière de test à laquelle H_0^i devrait être rejetée sachant toutes les autres statistiques de test impliquées.

Contrôle fort et faible Un contrôle fort [Dudoit et al., 2003; Westfall and Young, 1993] d'un taux de d'erreur de type I correspond à un contrôle quelques soient les combinaisons de vraies et fausses hypothèses nulles. Un contrôle faible correspond à un contrôle quand toutes les hypothèses nulles sont vraies, ce qui est peu réaliste en pratique. En effet, en règle générale, les hypothèses testées peuvent être vraies et d'autres fausses et on ne connait pas le sous-ensemble d'hypothèses fausses.

2.4.3 Le taux d'erreur par famille ou Family Wise Error Rate (FWER)

Définition Le Family Wise Error Rate (FWER) est la probabilité d'avoir au moins une erreur de type I (faux positif), c'est-à-dire par exemple de déclarer différentiellement exprimé au moins un gène non différentiellement exprimé.

$$\mathrm{FWER} := \mathbb{P}(\mathrm{V} > 0).$$

La procédure de Bonferroni est l'une des plus classiques pour contrôler le FWER. Elle consiste à réaliser chaque test au niveau $\alpha = \alpha^* / m$ ou à utiliser une probabilité critique ajustée $\tilde{p}_i = \min(1, p_i * m)$ avec $\mathrm{FWER} \leq \alpha^*$. Cette procédure effectue un contrôle fort (c'est-à-dire valide également lorsque $m_0 \neq m$). Elle est facile à mettre en oeuvre mais conservative et peu puissante. Quand le nombre de tests augmente, le FWER tend vers 1 avec un nombre V de faux positifs constant.

2.4.4 Le taux de fausses découvertes ou False Discovery Rate (FDR)

Plutôt que de contrôler la probabilité que l'étude produise une fausse découverte (faux positif) ou plus, on peut préfèrer contrôler la quantité de fausses découvertes parmi les découvertes. C'est ce qu'on appelle le FDR ou False Discovery Rate. Benjamini and Hochberg [1995] ont proposé une procédure pour un contrôle fort du FDR dans le cas de tests indépendants. Benjamini and Yekutieli [2001] ont montré que le contrôle était valable aussi dans le cas de certains types de dépendance (telle que la régression positive).

Définition Le FDR de Benjamini and Hochberg [1995] est la proportion attendue d'erreurs de type I parmi les hypothèses rejetées :

$$\text{FDR} := \mathbb{E}(\text{Q}) \text{ avec } \text{Q} = \frac{\text{V}}{\text{R}} \text{ si R} > 0 \text{ et Q} = 0 \text{ si R} = 0.$$

Lorsque V = 0 alors FDR = FWER, cela implique que les procédures qui contrôlent le FDR effectuent aussi un contrôle faible du FWER.

La procédure de Benjamini and Hochberg [1995] consiste dans un premier temps à ordonner les probabilités critiques brutes de la plus faible à la plus grande. On note $p^{(1)} \leq p^{(j)} \leq p^{(m)}$ ces probabilités ordonnées. Pour contrôler un FDR à un niveau α, il faut rejeter les hypothèses $\text{H}_0^{(j)}$ $(j = 1, \cdots, j^*)$ avec j^* tel que $j^* = \max\left\{j \text{ tel que} p^{(j)} \leq \frac{m}{j}\alpha\right\}$. Les probabilités critiques ajustées sont alors définies par $\tilde{p}_j = \min_{k=1,\cdots,m}\left\{\min(1, \frac{m}{k}p^{(k)})\right\}$.
Si les tests sont indépendants, la procédure de Benjamini and Hochberg [1995] implique $\text{FDR} \leq \frac{m_0}{m} \leq \alpha$, où m_0 le nombre d'hypothèses nulles est inconnu.
Storey et al. [2004] ont proposé l'estimateur suivant de m_0 :

$$\hat{m}_0(\lambda) = \frac{m - \text{R}(\lambda)}{(1-\lambda)}$$

où $m - \text{R}(\lambda)$ est le nombre d'hypothèses nulles acceptées par la procédure de tests et λ est un paramètre de réglage $\in [0, 1)$.

2.5 La régression locale ou 'lowess'

La régression locale initialement proposée par Cleveland [1979, 1981] est une méthode de régression non paramétrique locale pondérée, produisant une courbe lissée ajustée à un nuage de points. Le principe est le suivant : pour chaque point x_i, on ajuste dans un voisinage de ce point, un polynôme de faible degré (0, 1 ou 2). Les coefficients du polynôme sont calculés à l'aide de la méthode des moindres carrés pondérés. Plus un point est proche du centre du voisinage (point dont la réponse est estimée), plus il a un poids important. La fonction de pondération classiquement utilisée est $w(x) = (1 - |x|^3)^3 \mathbf{1}[|x| < 1]$. Les points du voisinage sont pondérés de façon décroissante par rapport à leur distance au centre du voisinage. La valeur prédite alors en x_i donne la valeur

prédite en ce même point par la régression lowess. Les paramètres importants sont les degrés du polynôme et la taille du voisinage. Le voisinage est choisi de façon à comporter une certaine fraction des données (appelée paramètre de lissage et notée f). Plus f est grand, plus l'ajustement est lissé.

Application à la normalisation de données issues de puces à ADN deux couleurs Cette méthode appliquée aux données de puces à ADN a été proposée par Yang et al. [2002] comme méthode de normalisation permettant d'éliminer les biais systématiques en lien avec l'intensité du signal. La façon la plus facile de visualiser les biais liés à l'intensité est la représentation graphique sous forme de MA-plot , variante du Bland-Attman Plot ou Tukey Mean-Difference plot.

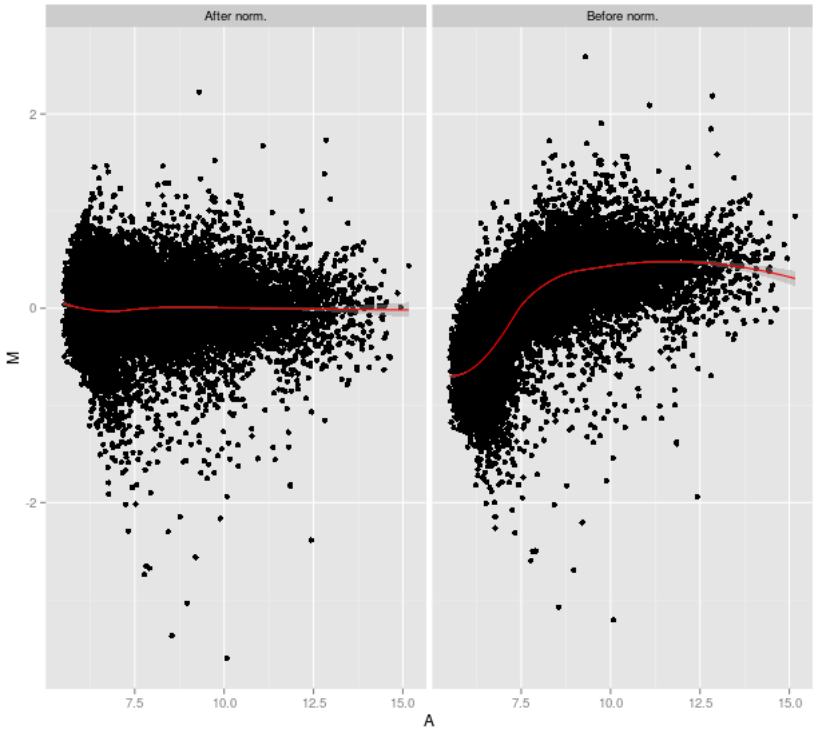


FIGURE 2.5 – Représentation MA-plot avant (à droite) et après (à gauche) une normalisation 'lowess'. La courbe rouge représente sur chacun des graphiques l'ajustement d'une régression lowess sur le nuage de points

Le MA-plot [Dudoit et al., 2002] représente pour une puce deux couleurs, le ratio M_i du signal marqué en Rouge R_i sur le signal marqué en Vert G_i pour chaque gène i de la puce exprimé en log base 2, soit $M_i = \log_2(\frac{R_i}{G_i})$ en fonction de A_i le log base 2 du produit des intensités dans les deux canaux $A_i = \log_2(\sqrt{(R_i * G_i)})$. Le MA-plot (figure 2.5 partie droite) constitue le point de départ de la normalisation lowess. La valeur du $M = (M_i)_i$ normalisé est obtenue en soustrayant à la valeur initiale l'ajustement $c(A)$ par une lowess

sur le MA-plot.

$$\begin{aligned} M' &= M - c(A) \\ &= \log_2\left(\frac{R}{k(A) * G}\right) \end{aligned} \tag{2.14}$$

Dans leur article, Yang et al. [2002] proposent d'utiliser la fonction *lowess* implémentée dans le logiciel R. Cette version a l'avantage de ne pas être affectée par un faible pourcentage de gènes différentiellement exprimés, qui apparaissent comme des éléments aberrants sur le MA-plot. Ils préconisent d'utiliser une valeur assez grande par exemple $f = 40\%$ pour le paramètre de lissage.
A noter qu'on peut utiliser les formules suivantes pour obtenir des valeurs normalisées dans chacun des canaux : $R'_i = R_i$ et $G'_i = G_i * 2^{c(A_i)}$.

2.6 Références

M. AlKhasawneh. Estimating the negative binomial dispersion parameter. *Asian Journal of Mathematics and Statistics*, 3 :1–15, 2010. 26

L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1) :164–171, 1970. 16

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *JRSSB*, 57(1) :289–300, 1995. 28, 30

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4) :1165–1188, 2001. 30

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725, 2000. 21

C. M. Bishop. *Pattern Recognition and Machine Learning*. New York : Wiley, 2006. 17

C. Bliss and R. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2) :176–200, 1953. 22

A. Cameron and P. Trivedi. Econometrics models based on count data : Comparisons and applications of some estimators. *Journal of Applied Econometrics*, 1 :29–53, 1986. 25

A. Cameron and P. Trivedi. *Regression Analysis of Count Data, 2nd edition*. Cambridge University Press, 2013. 21

W. Cleveland. Robust locally weighted regression and smoothing scatterplotsy. *Journal of the American Statistical Association*, 74(368) :829–836, 1979. 30

W. Cleveland. Lowess : A program for smoothing scatterplots by robust locally weighted regression. *Journal of the American Statistical Association*, 35(1) :54, 1981. 30

P. de Jong and G. Heller. *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008. 22

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 :1–38, 1977. 14

P. Devijver. Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3 (6) :369–373, 1985. 16

S. Dudoit, Y. Yang, M. Callow, and T. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, 12 : 1, 2002. 31

S. Dudoit, J. Popper Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18 :71–103, 2003. 28, 29

Z. Ghahramani and M. Beal. Variational inference for bayesian mixtures of factor analysers. In *Neural Information Processing Systems 12*, 2000. 17

M. Greenwood and G. Yule. An inquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society. Series A*, 83 :255–279, 1920. 25

J. Hilbe. *Negative Binomial Regression 2nd edition*. Cambridge University Press, 2011. 25

B. Jø rgensen. *The Theory of Dispersion Models*. Chapman & Hall \CRC, 1997. 26

N. Johnson, A. Kemp, and S. Kotz. *Univariate discrete distributions*. New York : John Wiley & Sons, 2005. 26

S. P. Krishna Saha. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61(1) :179–185, 2005. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3695660. 26

E. Lebarbier and T. Mary-Huard. Le critère bic : fondements théoriques et interprétation. *Journal de la Société Française de Statistique*, 147 :39–58, 2006. 20

J. Marioni, C. Mason, S. Mane, M. Stephens, and Y. Gilad. Rna-seq : an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9) :1509–17, 2008. 21

D. McCarthy, Y. Chen, and G. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40 :4288–4297, 2012. 27

G. McLachlan and D. Peel. *Finite Mixture Models, Willey Series in Probability and Statistics*. New York : John Wiley & Sons, 2000. 14, 23

R. M. Neal and G. E. Hinton. *Learning in Graphical Models*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-60032-3. URL http://dl.acm.org/citation.cfm?id=308574.308679. 17

J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64 :140–153, 2010. 17

S. Richards. Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45 :218–227, 2008. 22

M. Robinson and G. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2) :321–332, 2008. 21, 22

E. Roquain. Type i error rate control in multiple testing : a survey with proofs. *Journal de la Société Française de Statistique*, 152,2 :3–38, 2011. 28

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 03 1978. doi : 10.1214/aos/1176344136. URL http://dx.doi.org/10.1214/aos/1176344136. 20

M. Shaked. On mixtures from exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2) :192–198, 1980. 24

J. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates : A unified approach. *JRSSB*, 66 :187–205, 2004. 30

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2) :260–269, April 1967. ISSN 0018-9448. doi : 10.1109/TIT.1967.1054010. 17

P. Westfall and S. Young. *Resampling-Based Multiple Testing : Examples and Methods for p-Value Adjustment*. Wiley, 1993. 29

Y. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cdna microarray data : a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30 :e15, 2002. doi : 10.1093/nar/30.4.e15. 31, 32

Chapitre 3

Modèles à variables latentes pour l'analyse de données omiques

Sommaire

Le modèle à variables latentes le plus simple est le modèle de mélange. Dans ce modèle, les observations sont supposées indépendantes, chacune appartenant à une classe non observée. Les travaux présentés dans les sections 3.1 et 3.2 en sont des exemples : le premier dans un contexte de tests multiples et le deuxième dans le cadre de la définition d'un seuil d'hybridation pour des données issues de puces à ADN (chapitre 1 section1.4). Dans la section 3.3, nous proposons un modèle de chaînes de Markov cachées avec prise en compte de la dépendance entre les individus. Quand le nombre d'individus augmente, l'inférence exacte ne peut plus être envisagée. Nous proposons alors une inférence approchée basée sur une approche variationnelle (cf. section 2.1.3). Les notations utilisées sont les mêmes que celles présentées dans le chapitre introductif (chapitre 2 page 11). Plus de détails sur les résultats des méthodes peuvent être trouvés dans les articles associés aux sections 3.1 et 3.2, respectivement disponibles en annexe A et B.

3.1 FDR et FDR local

Cette section présente le travail que j'ai effectué en collaboration avec A. Bar-Hen, J.J Daudin et S. Robin [**Aubert, J** et al., 2004]. Ce travail a été développé dans le cadre de la recherche de gènes différentiellement exprimés entre plusieurs conditions à partir de données issues de puces à ADN mais est applicable dans tout contexte de tests multiples quelque soit la nature de la question ou la technologie utilisée. Nous utiliserons les notations proposées par Efron [2004], à savoir FDR pour le terme générique incluant le FDR classique (noté Fdr) et le FDR local (noté fdr).

3.1.1 Contexte

Afin de trouver les gènes impliqués dans un phénomène biologique particulier, on mesure l'expression de tous les gènes d'un organisme particulier chez plusieurs individus (répétitions biologiques) et dans plusieurs conditions. Pour chaque gène, on teste l'hypothèse nulle H_0^i = 'le gène i n'est pas différentiellement exprimé entre les conditions'. La proportion attendue de gènes faux positifs (cf. section 2.4) dans un ensemble de gènes, appelée False Discovery Rate (FDR) a été proposée pour mesurer la significativité statistique du test. De nombreuses procédures existent pour contrôler ce FDR. Le seuil choisi, généralement 5%, est arbitraire. Storey and Tibshirani [2003] ont introduit la *q-value* comme mesure de la significativité propre à un gène. Cependant cette mesure définie par la valeur du FDR atteinte pour une valeur observée donnée donne une vue trop optimiste de la probabilité qu'un gène soit un faux positif car seuls les gènes plus significatifs que celui pour lequel la q-value est calculée sont utilisés. C'est pourquoi nous avons proposé une mesure spécifique associée à chaque gène : le fdr local utilisant les gènes au voisinage de celui d'intérêt. On entend par voisinage, un ensemble de gènes pour lesquels les probabilités critiques associées sont proches de celle du gène considéré. Le fdr local

transforme le problème de tests multiples en un problème de classification à l'aide d'un modèle de mélange à deux composantes, représenté sur la figure 2.1.

3.1.2 Méthode

Modèle

On considère m tests simultanés. A chaque hypothèse testée H_0^i est associée une probabilité critique ou p-value (brute) P_i. Une distribution typique des probabilités critiques à l'issue des tests est représentée sur la figure 3.1. On peut supposer que les valeurs observées proviennent de deux classes dépendant du statut du gène i. Soit Z_i la variable latente correspondant au statut du gène, qui prend la valeur 1 si le gène est différentiellement exprimé (sous H_1) et 0 sinon. On peut supposer de façon assez naturelle que Z_i, $i = 1,\ldots,n$ suit une loi de Bernouilli de paramètre $1-\pi_0$. On utilisera par abus de langage indifféremment π_0 ou $\frac{m_0}{m}$. Les P_i étant des probabilités critiques, la distribution de P_i sachant $Z_i = 0$ est une uniforme $\mathscr{U}(0,1)$. On peut alors écrire le modèle de mélange à deux composantes suivant pour les probabilités observées :

$$\begin{aligned} f(p) &= \frac{m_0}{m} f_0(p) + \left(1 - \frac{m_0}{m}\right) f_1(p) \\ &= \frac{m_0}{m} + \left(1 - \frac{m_0}{m}\right) f_1(p), \end{aligned} \tag{3.1}$$

où f_1 correspond à la fonction de densité sous l'hypothèse H_1. Le modèle de mélange peut aussi s'écrire en terme de fonctions de répartition :

$$f(p) = \frac{m_0}{m} p + (1 - \frac{m_0}{m}) F_1(p).$$

A noter que les seules deux quantités inconnues sont m_0 et F_1.

Fdr

Notons $\texttt{Fdr}(p)$ le Fdr tel que toutes les hypothèses nulles avec une probabilité critique inférieure à p sont rejetées, dans ce cadre de modèle de mélange, $Fdr(p)$ se définit comme :

$$\texttt{Fdr}(p) = P(\text{'}H_0^i \texttt{ vraie'} | P_i \leq p) = \frac{\frac{m_0}{m} F_0(p)}{F(p)} = \frac{\frac{m_0}{m} p}{F(p)}.$$

Cette définition correspond à la celle de la q-value, qui correspond à la valeur du Fdr atteinte pour une valeur observée donnée.

En majorant $\frac{m_0}{m}$ par 1 et en remplaçant F par la fonction de répartition empirique $\hat{F}(p_i) = \frac{\texttt{order}(p_i)}{m}$, on obtient l'estimateur suivant pour le Fdr :

$$\widehat{\texttt{Fdr}}(p_i) \leq p_i \frac{m}{\texttt{order}(p_i)}.$$

Il s'agit de la même quantité que celle intervenant dans la procédure de Benjamini-Hochberg (cf. section 2.4).

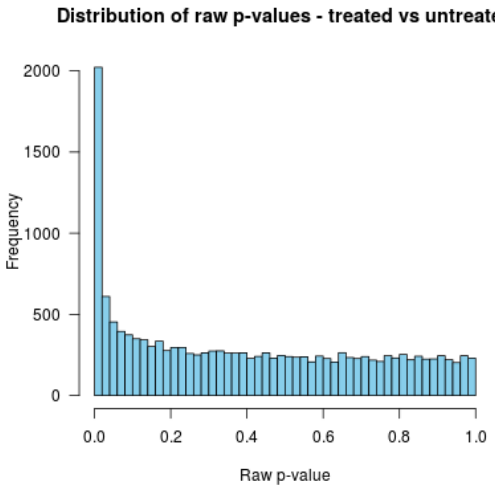


FIGURE 3.1 – Distribution des probabilités critiques à l'issue des tests

FDR local

Le FDR local, noté fdr, est quand à lui défini comme la probabilité qu'un gène soit sous H_0 sachant la probabilité critique qui lui est associée. On notera Z_i la variable latente telle que $Z_i = \mathbb{1}\{\text{'}H_1^i \texttt{ est vraie'}\}$, Z_i peut prendre deux modalités 0 en cas de non différence d'expression et 1 dans le cas contraire. Le FDR local peut alors s'écrire comme :

$$\begin{aligned} fdr(p) = P(\text{'}H_0^i \texttt{ vraie'}|P_i = p) &= \mathbb{E}(1 - Z_i|P_i) \\ \tau_{i0} &= \frac{\pi_0 f_0(p)}{f(p)}. \end{aligned} \tag{3.2}$$

Les difficultés résident en l'estimation de f ou F, qui n'a pas de forme paramétrique particulière spécifiée, à partir des probabilités critiques observées, et l'estimation de m_0. Nous proposons d'estimer la densité de probabilité à l'aide d'une fonction de lissage de la fonction de répartition empirique et utilisons l'estimateur de m_0 proposé par Storey et al. [2004] (chapitre 2 page 28), à savoir $\hat{m}_0(\lambda) = \frac{m - R(\lambda)}{1-\lambda}$; ce qui aboutit à une définition du FDR local en deux étapes :

1. une première étape d'estimation d'un FDR local brut en utilisant les différences successives entre les probabilités critiques observées ordonnées :

$$\widehat{\texttt{fdr}}_\lambda(p_i) = \begin{cases} m_0(\lambda)(p_i - p_{i-1}) \texttt{si } i > 1 \\ p_1 \texttt{si } i = 1 \end{cases}$$

2. une deuxième étape d'estimation d'un FDR local comme une valeur lissée basée

sur les données brutes afin de pallier la variabilité de l'estimateur $\widehat{\texttt{fdr}}_\lambda(p_i)$

$$\widehat{\texttt{fdr}}^s_\lambda(p_i) = g_i(\widehat{\texttt{fdr}}_\lambda(p_i))\texttt{avec} \quad j = 1, \cdots, m$$

avec g_i la valeur à la position i d'une fonction lissage.

Propriétés du fdr local brut Supposons que les probabilités critiques associées aux gènes non-différentiellement exprimés sont indépendantes, alors l'estimateur du fdr local brut a les propriétés suivantes :

1. Sous H_0^i et H_0^{i-1} et si $\mathbb{E}(\hat{m}_0) = m_0$ alors $\widehat{\texttt{fdr}}_\lambda(p_i)$ est sans biais de moyenne 1.
2. Sous H_0^i et H_0^{i-1} et si m_0 est connu, alors

$$\mathbb{V}(\widehat{\texttt{fdr}}_{m_0}(p_i)) = \frac{m_0^3}{(m_0+1)^2(m_0+2)}.$$

   Cette variance se rapproche de 1 si m_0 est assez grand, et nous fournit une borne inférieure de $\mathbb{V}(\widehat{\texttt{fdr}}_\lambda(p_i))$ si m_0 est inconnu.
3. A noter que la qvalue [Storey and Tibshirani, 2003] peut être vue comme la moyenne du fdr local des gènes ayant une probabilité inférieure à celle considérée p.

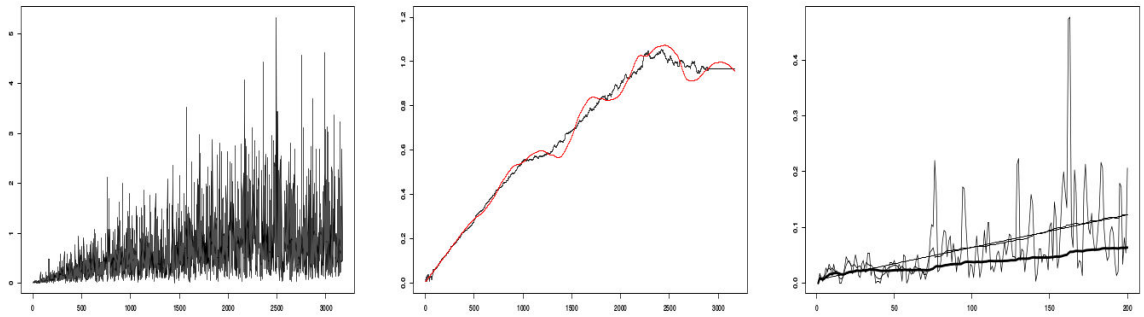
$$\begin{aligned} q_i = \texttt{Fdr}(p_i) &= \frac{\int_\infty^{p_i} \texttt{fdr}(\mathrm{P}_i) f(\mathrm{P}_i) d\mathrm{P}_i}{\int_\infty^{p_i} f(\mathrm{P}_i) d\mathrm{P}_i} \\ &= \mathrm{E}_f\left[\texttt{fdr}(\mathrm{P}_i) | \mathrm{P}_i <= p_i\right]. \end{aligned}$$

3.1.3 Résultats

$\widehat{\texttt{fdr}}^s_\lambda(p_i)$ fournit un indicateur pour choisir le seuil de significativité. Si l'on considère la courbe du fdr local contre l'ordre des gènes selon leur probabilité critique : un bon candidat serait un point avec une forte dérivée d'ordre 2, ce qui correspond sur les graphiques à un changement abrupt dans la pente (cf. figure 3.2(b)). L'intérêt de la méthode a été montré et discuté sur trois jeux de données bien connus.

3.1.4 Conclusions

Le fdr local associé à chaque gène mesure la probabilité que ce gène soit un faux positif. Nous avons proposé une façon de l'estimer. La courbe de l'estimateur lissé permet à la fois de mesurer l'information sur le nombre de gènes et leur significativité, et un moyen de choisir un seuil pour distinguer les gènes sous H_0 et sous H_1. Dans le cadre d'études génomiques ou transcriptomiques, cela permet de calculer un FDR pour un groupe de clones, séquences (d'un même gène) ou pour un groupe de gènes provenant de la même voie de signalisation ou positionné dans la même région chromosomique. Cet article a



(a) (b) (c)

FIGURE 3.2 – Graphiques des estimations du fdr local sur les données d'Hedenfalk avec en abscisse l'index des gènes ordonnés selon leur probabilité critique et en ordonnée l'estimation du fdr local. (a) : valeurs brutes, (b) : estimations lissées par moyenne mobile (sauts discrets), régression lowess (courbe lissée), (c) : zoom sur les 200 premiers gènes de (b) : valeurs brutes (sauts discrets), moyenne mobile et lowess (courbes lissées), q-value (courbe lissée moins épaisse)

été cité 43 fois (source : Web of Science®), à la fois dans des articles de développement de nouvelles méthodes, et dans des articles de recherche en biologie). Depuis de nouvelles méthodes ont vu le jour, permettant notamment d'estimer f_1 ou d'améliorer l'estimation de $\frac{m_0}{m}$. Les procédures de contrôle du FDR utilisent souvent comme ici des probabilités critiques en entrée. Cependant il peut être intéressant d'utiliser directement des statistiques de test. Cela peut permettre d'estimer f_0 à partir des données et de prendre en compte par exemple une dépendance entre les tests. Strimmer [2008] a proposé un article présentant dans un cadre unifié de l'ensemble des méthodes proposées et un package R *fdrtool* associé. Une fonction R pour calculer les estimations du fdr local est disponible dans le package R *anapuce*.

3.2 Modèle de mélange de gaussiennes tronquées pour définir un seuil d'hybridation

Ce travail présenté dans un rapport technique [Picard et al., 2008], disponible dans l'annexe B, est issu d'une collaboration avec F. Picard, S. Robin et M.-M. Martin-Magniette pour la partie statistique et des biologistes de l'ex-URGV (IPS2 Saclay) pour la partie biologique.

3.2.1 Contexte

Pour chacun des gènes d'un génome, on mesure à l'aide d'une technologie de type puce à ADN son niveau d'expression $Y_i, i = 1,\ldots,n$ dans une condition particulière. On aimerait pour chaque condition distinguer les gènes en fonction de leur niveau d'hybridation (faible, moyen, fort par exemple) et particulièrement les gènes non exprimés (non hybridés) des autres. Sur une puce, on observe un continuum de valeurs d'intensité parmi

les sondes dû notamment à la présence d'un fort bruit de fond. Il est donc difficile à l'oeil de distinguer les sondes faiblement exprimées de celles pour lesquelles on observe un signal dû à un bruit de fond mais pour lesquelles l'hybridation n'a pas lieu d'être (pas assez d'identité entre la sonde et la cible). Les procédures existantes sont basées sur l'estimation d'un bruit de fond local et l'utilisation d'un seuil arbitraire ou sur l'utilisation de sondes dites contrôles. Un critère souvent retenu est de considérer que les sondes ayant un signal inférieur à deux fois le bruit de fond ne sont pas exprimées. Dans ce travail nous proposons une approche différente en deux étapes. La première étape consiste à estimer la distribution de l'intensité du signal observé sur une puce à l'aide d'un modèle de mélange, et la deuxième à définir un seuil d'hybridation à partir des probabilités conditionnelles pour chaque sonde d'appartenir à chaque composante du mélange.

3.2.2 Méthode

Modèle

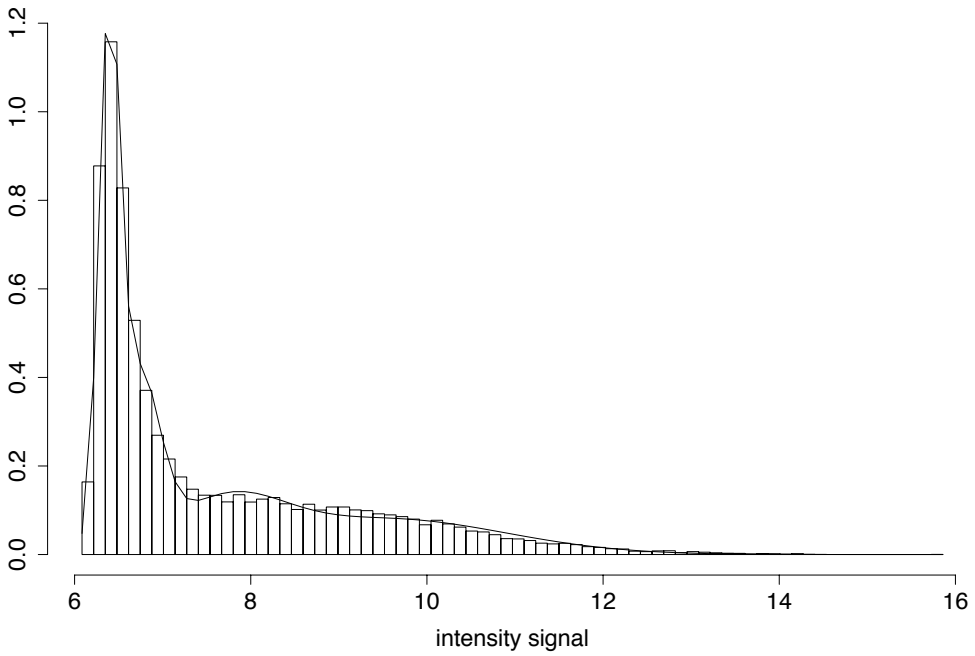


FIGURE 3.3 – Distribution de l'intensité du signal en log base 2 sur une puce

La figure 3.3 représente un histogramme typique des valeurs d'intensité des sondes observées en $\log_2$ sur une puce. Nous constatons que les données sont bornées par une valeur minimale correspondant au niveau d'autofluorescence de l'ADN et par une valeur maximale due à une saturation du signal et qu'elles sont asymétriques due à une forte proportion de sondes non hybridées. Après plusieurs analyses avec différentes distributions, nous avons finalement opté pour une modélisation de cet histogramme à l'aide

d'un modèle de mélange de gaussiennes tronquées. Les seuils de troncature à droite et à gauche sont déterminés a priori et correspondent au niveau minimal de signal détectable et au niveau de saturation du signal.

Soit Z_i la variable latente correspondant au statut du gène et relative à son niveau d'expression. Le nombre K de classes n'est pas connu a priori. Soit $f(y) = \sum_{k=1}^{K} \pi_k f_k(y;\theta_k)$ la densité de probabilité d'une observation dans le cas d'un modèle de mélange de gaussiennes non tronquées avec f_k gaussienne de moyenne μ_k et de variance σ^2. Après troncature à droite en $u = \max_i(y_i)$ et à gauche en $\ell = \min_i(y_i)$, on obtient $g(y) = f(y) / \int_\ell^u f(y)\mathrm{d}y$, qui est également un modèle de mélange :

$$g(y) = \sum_{k=1}^{K} \eta_k g_k(y)$$

avec les proportions de mélange $\eta_k = \pi_k \frac{\int_\ell^u f_k(y)\mathrm{d}y}{\int_\ell^u f(y)\mathrm{d}y}$ et $g_k(y) = \frac{f_k(y)}{\int_\ell^u f_k(y)\mathrm{d}y}$. Chaque composante g_k est une version tronquée de la composante originelle f_k.

La représentation sous forme de modèle graphique est la même que pour le modèle précédent (cf. figure 2.1).

Estimation

L'estimation des paramètres du modèle est faite via l'utilisation d'un algorithme EM. La log-vraisemblance complète $\mathscr{L}_T(\theta)$ s'écrit

$$\begin{aligned}\mathscr{L}_T(\theta) &= \sum_i \sum_k Z_{ik} \left[\log \eta_k + \log g_k(y)\right] \\ &= \sum_i \sum_k Z_{ik} \left[\log \eta_k + \log f_k(y) - \log \int_\ell^u f_k(y)\mathrm{d}y\right]. \end{aligned} \tag{3.3}$$

La différence principale entre l'expression de la log-vraisemblance dans ce modèle (3.3) et celle dans un modèle de mélange de gaussiennes non tronquées vient du terme de normalisation $\log \int_\ell^u f_k(y)\mathrm{d}y$. Ce terme ne complique pas l'étape E qui reste inchangée :

$$p(Z_i = k|Y) = \frac{\eta_k g_k(y)}{\sum_{k'} \eta_{k'} g_{k'}(y)} = \frac{\pi_k f_k(y)}{\sum_{k'} \pi_{k'} f_{k'}(y)}.$$

Il intervient cependant dans l'étape M. On obtient les équations de mises à jour suivantes :

$$\hat{\eta}_k = \frac{1}{G}\sum_i \tau_{ik},$$
$$\hat{\mu}_k = \frac{\sum_i \tau_{ik} y_i}{\sum_i \tau_{ik}} - m_k,$$
$$\hat{\sigma}_k^2 = \frac{\sum_i \tau_{ik}(y_i - \hat{\mu}_k)^2}{\sum_i \tau_{ik}} + H_k,$$

avec m_k et H_k les moments d'ordre 1 et 2 d'une gaussienne tronquée de moyenne 0 et de variance σ_k^2 tronquée en $\ell - \mu_k$ et $u - \mu_k$.
Les formules ne sont pas explicites et nécessitent la résolution d'équations de points fixes. Elles sont néanmoins assez similaires au cas non tronqué à m_k et H_k près.

Sélection de modèle

Nous envisageons une collection de modèles de mélange : non tronqués, tronqués à gauche, à droite, à droite et à gauche et choisissons le meilleur modèle à l'aide d'un critère BIC :

$$\mathrm{BIC}(\mathscr{M}_K\{\ell, u\}) = -2\log \mathscr{L}_T(Y; \pi, \theta | \mathscr{M}_K\{\ell, u\}) + (3K-1)\log(n)$$

où on note $\mathscr{L}_T(Y; \pi, \theta | \mathscr{M}_K\{\ell, u\})$ la vraisemblance du modèle de mélange $\mathscr{M}_K\{\ell, u\})$ à K composantes tronqué à gauche en ℓ et à droite en u. Le terme de pénalité $(3K-1)\log(n)$ dépend ici du nombre de classes K mais pas de la troncature en ℓ et u. Dans l'idéal, le modèle retenu aurait deux composantes : l'une correspondant aux sondes hybridées et la deuxième aux sondes non hybridées. La réalité étant plus complexe, le modèle à deux composantes est rarement celui sélectionné au profit d'un modèle à quatre ou cinq composantes. Il est alors nécessaire de définir un seuil d'hybridation afin de distinguer en pratique les sondes non hybridées des autres.

Définition du seuil d'hybridation

Une fois le modèle sélectionné, les composantes sont ordonnées de façon croissante selon leur paramètre de moyenne. La seule information certaine est que les sondes de la composante avec la plus grande moyenne, autrement dit de la $\hat{K}$ème composante, sont hybridées. Rien ne peut être dit sur la composante avec la plus faible moyenne étant donnée qu'il y a toujours ambiguïté entre les sondes faiblement exprimées et les sondes non hybridées. Si l'on considère comme hybridées seulement les sondes classées dans cette dernière composante selon la règle du Maximum A Posteriori (MAP), un seuil d'hybridation naturel se définit comme la valeur minimale du signal observé dans la composante de plus grande moyenne

$$T_{MAP} = \min\{y_i | k_i^* = \hat{K}\}.$$

Cette procédure étant vraiment conservative, nous proposons la procédure définie ci-dessous.

Pour chaque sonde i, k_i^* est déterminé selon la règle du MAP. Si la composante à laquelle la sonde est affectée n'est pas $\hat{K}$, alors on calcule la probabilité conditionnelle $\hat{\tau}_{si}$ d'appartenir à une composante de moyenne plus petite ou égale à celle de k_i^*. Le seuil d'hybridation est ensuite défini comme la première intensité telle qu'une des probabilités conditionnelles calculées soit plus grande que ϵ. En pratique, nous prenons $\epsilon = 10^{-4}$. Cette procédure aboutit à la définition ci-dessous du seuil d'hybridation :

$$T_\epsilon = \max\{y_i | k_i^* < \hat{K} \exists s \in \{1; \ldots; k_i^* - 1\}, \hat{\tau}_{si} \geq \epsilon\}.$$

Cette règle est illustrée sur la figure 3.4 pour un échantillon biologique.

3.2.3 Résultats

Les performances de notre méthode, appelée *MixThres*, ont été évaluées sur deux types de données. Le premier jeu de données correspond à l'hybridation de quatre échantillons biologiques (ARN messagers) sur une puce 'tiling' représentant le chromosome 4 d'*Arabidopsis thaliana*. Les sondes de cette puce couvrant à la fois les régions géniques et inter-géniques, les sondes situées dans les régions inter-géniques ne devraient pas hybridées. Ce jeu a donc servi aussi à vérifier la reproductibilité de la méthode, c'est-à-dire si les sondes déclarées hybridées dans un échantillon, le sont aussi dans un autre, et si les sondes situées dans les régions inter-géniques sont bien déclarées non hybridées pour l'ensemble des échantillons. Parmi les 21602 sondes que comporte la puce, 4681 ont été déclarées hybridées sur les 4 puces, et 13681 hybridées sur aucune puce, ce qui donne 88% de cohérence. De plus les seuils trouvés ainsi que le pourcentage de sondes déclarées hybridées varient de façon raisonnable (Annexe B Table 2). Un important travail d'investigation bioinformatique a été effectué suite à ces résultats, ce qui a permis de définir un ensemble de sondes situées dans les régions inter-géniques sans possibilité d'hybridation croisée. Parmi ces 3701 sondes, 49, soit 1.3% sont déclarés hybridées à tort. Le deuxième jeu de données correspond à 522 échantillons d'*Arabidopsis thaliana* provenant d'expériences de transcriptomique hybridés sur la puce CATMA et disponible dans la base CATdb [Gagnot et al., 2008]. Toutes les sondes déclarées hybridées par MixThres pour lesquelles aucune preuve connue d'existence de transcription existait ont été validées par des approches complémentaires. 88% des 465 potentiels nouveaux gènes découverts ont été ainsi validés [Aubourg et al., 2007].

3.2.4 Conclusions

Nous avons proposé une méthode pour définir de façon automatique un seuil d'hybridation basé sur les données. La méthode a été validée à la fois sur des données de

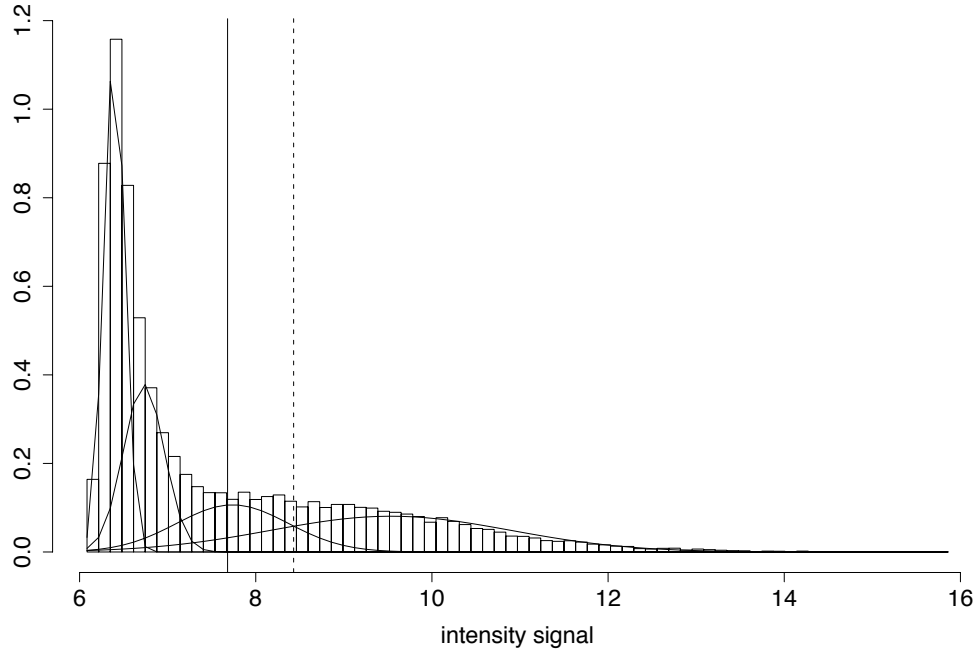


FIGURE 3.4 – Distribution de l'intensité d'un échantillon biologique avec le modèle sélectionné à 4 composantes. La droite verticale en trait plein correspond au seuil $T_\epsilon = 7.68$ avec $\epsilon = 10^{-4}$. La droite en pointillé indique le seuil $T_{MAP} = 8.43$.

puce d'expression et de puce à ADN (tiling array). Nous avons montré que la méthode a une bonne reproductibilité, que sa spécificité est supérieure à 97 % et sa précision à 88%. Un package R *MixThres* disponible à l'adresse http://www6.inra.fr/mia-paris/Production-Scientifique/Logiciel a été proposé afin de permettre l'utilisation de cette méthode. Elle a fait l'objet d'un rapport technique et a été utilisée dans plusieurs articles de biologie bien qu'elle n'ait pas été acceptée dans les journaux de bioinformatique.

3.3 Approche variationnelle dans un modèle de chaînes de Markov couplées

Ce travail est issu d'une collaboration avec X. Wang, E. Lebarbier et S. Robin dans le cadre du post-doctorat de X. Wang financé par le projet ANR CNV-Maize, coordonné par S. Nicolas (UMR Génétique Quantitative et Evolution - Le Moulon).

3.3.1 Contexte

Les modèles de Markov cachés fournissent un cadre statistique naturel pour la détection de variation du nombre de copies en génomique (cf. chapitre 1 section 1.2). Un processus de Markov caché (chapitre 2 section 2.1.2) est associé à chaque individu pour classer les régions génomiques selon leur statut (classiquement : gain, perte, normal). Ici nous nous intéressons à la détection de variants structuraux au sein de lignées de maïs. Ces lignées étant issues de sélection variétale, partagent un passé phylogénétique commun. De ce fait, on s'attend à ce que les variations structurales au sein de chacune des lignées ne soient pas complètement indépendantes de celles des autres lignées. Nous considérons un modèle de Markov caché, prenant en compte simultanément plusieurs processus cachés dépendants. Dans le cas d'un grand nombre de séries (individus), l'inférence par maximum de vraisemblance n'est pas possible. Nous proposons alors un algorithme d'inférence approchée fondé sur une approche variationnelle pour déduire des variations structurales dans des génomes de plantes.

3.3.2 Méthode

Modèle

Soient I individus ($i = 1,\dots,\mathrm{I}$), pour chacun desquels on observe une série de mesures $\mathrm{Y}_i = (\mathrm{Y}_{i,t})$. Ces mesures peuvent être typiquement issues de puces à ADN et sont supposées varier selon l'état de l'individu à un 'temps' donné $t = 1,\dots,\mathrm{T}$. On note $(\mathrm{Z}_{i,t})_t$ le processus caché pour l'individu i. $\mathrm{Z}_{i,t}$ peut prendre Q valeurs différentes (en général Q = 3 : -1 = 'perte', 0 = 'normal', 1 = 'gain'). Dans ce cas, l'espace des états du processus caché joint $(\mathrm{Z}_t)_t$, avec $\mathrm{Z}_t = (\mathrm{Z}_{1,t},\dots\mathrm{Z}_{\mathrm{I},t})$ consiste en $\mathrm{K} = \mathrm{Q}^{\mathrm{I}}$ valeurs possibles. Nous supposons

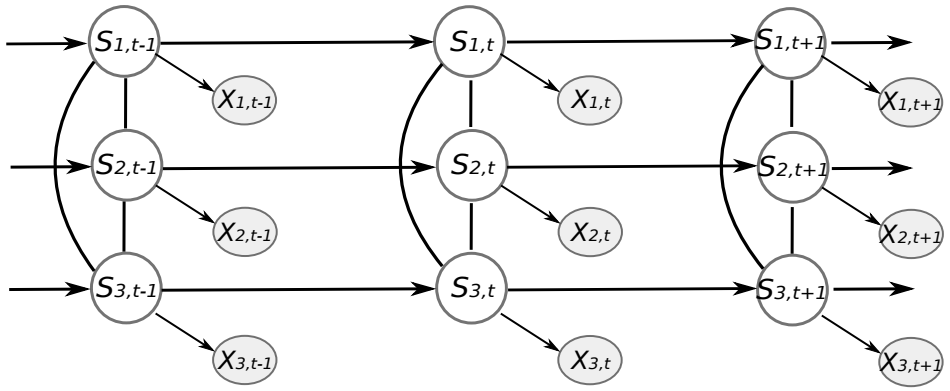


FIGURE 3.5 – Modèle graphique comportant à la fois des arêtes dirigées et non dirigées. Les arêtes dirigées représentent les relations de dépendance intra-individu. Les arêtes non dirigées représentent la corrélation génétique inter-individus. Sur cette figure, les variables cachées sont notées S et les observations X.

que Z est une chaîne de Markov d'ordre 1 et que les observations $Y_{i,t}$ sont échantillonnées conditionnellement à Z de la façon suivante :

$$Y_{i,t} = \sum_{q=1}^{Q} Z_{i,t}^{q} \mu_q + \varepsilon_{i,t},$$

où $Z_{i,t}^{q} = \mathbf{1}_{\{Z_{i,t}=q\}}$, μ_q correspond à la valeur moyenne de l'état q et les $(\varepsilon_{i,t})$ sont indépendants et identiquement distribués. On notera $\mu = (\mu_q)$.

Les différentes structures de dépendance que nous considérons sont résumées dans le modèle graphique de la figure 3.5.

La variation du statut caché le long du génome ainsi que la corrélation entre individus sont codées dans la matrice de transition P de taille K × K du processus caché joint (Z_t). Nous considérons que les probabilités de transition résultent du produit de deux termes : un premier tenant compte des transitions au sein d'un individu et un deuxième tenant compte des similarités entre individus (supposées constantes le long du génome) :

$$\mathbb{P}(Z_t = \ell | Z_{t-1} = k) := P_{k\ell} \propto \left(\prod_i \pi_{k_i,\ell_i} \right) W_\ell, \tag{3.4}$$

où

(*a*) π est une matrice de transition Q × Q (chaque ligne somme à 1) et k_i (resp. ℓ_i) représente le statut caché de l'individu i quand l'état caché joint est k (resp. ℓ) ;

(*b*) la relation de dépendance entre les individus est codée dans les coefficients

$$W_\ell = \prod_{i,j \neq i} \omega^{s_{ij} \mathbf{1}_{\{\ell_j \neq \ell_i\}}}, \tag{3.5}$$

où $\omega < 1$ et s_{ij} représente la similarité (ici la proximité phylogénétique) entre les individus i et j ;

On suppose de plus que l'état initial $Z_1 = (Z_{i1})_i$ a pour distribution

$$\mathbb{P}(Z_1 = \ell) \propto \left(\prod_i m_{\ell_i}\right) W_\ell$$

où (m_q) est la distribution des états $1 \leq q \leq Q$.

Comme la distribution initiale et les transitions (3.4) ne sont pas normalisées, la loi des processus cachés Z s'écrit :

$$\mathbb{P}(Z) = \frac{1}{C} \prod_\ell \left[\left(\prod_i m_{\ell_i}\right) W_\ell\right]^{Z_1^\ell} \prod_{\substack{t\geq 2 \\ k,\ell}} \left[\left(\prod_i \pi_{k_i,\ell_i}\right) W_\ell\right]^{Z_{t-1}^k Z_t^\ell}$$

où C est une constante de normalisation. On a alors :

$$\begin{aligned} \log \mathbb{P}(Y,Z) \quad = \quad & \sum_{i,q} Z_{i,1}^q \log m_q + \sum_{i,t\geq 2,q,r} Z_{i,t-1}^q Z_{i,t}^r \log \pi_{qr} \\ & + \sum_{i,t,r} Z_{i,t}^r \sum_{j\neq i} (1 - Z_{j,t}^r) s_{ij} \log \omega \\ & + \sum_{i,t,r} Z_{i,t}^r \log \phi_r(Y_{i,t}) - \log C \end{aligned}$$

Inference

Dans cette section, nous nous intéressons à l'inférence des paramètres à Q et ω fixés. Nous proposons d'estimer les paramètres à l'aide d'un algorithme type EM [Dempster et al., 1977]. Dans le cas des HMM, l'étape E peut être faite à l'aide d'un algorithme type forward-backward (chapitre 2 section 2.1.2). Cette étape est l'étape critique pour l'inférence du modèle décrit précédemment. En effet, trois situations différentes peuvent être envisagées :

- (i) si on ne prend pas en compte la dépendance entre les individus, i.e. si on pose $\omega = 1$ dans (3.5), alors les états cachés de chaque individu sont indépendants et l'étape E peut être réalisée de façon classique pour chacun des individus.
- (ii) si on tient compte de la proximité phylogénétique entre les individus mais que le nombre d'états cachés et le nombre d'individus sont tous les deux petits, $K = Q^I$ inférieur à quelques dizaines, le modèle global peut être considéré comme un seul HMM et l'étape E peut être réalisée via une récursion Forward-Backward de complexité TK^2.
- (iii) si on tient compte de la proximité phylogénétique et que K ou I est trop grand, une alternative à l'étape E telle qu'effectuée dans les cas précédents doit être envisagée.

Dans les deux premiers cas, un algorithme EM standard peut être utilisé. Notre travail s'est donc focalisé sur le troisième. Dans le cas où K est grand, la distribution conditionnelle de Z sachant les données observées Y devient intractable et l'étape E classique n'est

plus faisable. Nous suivons alors une approche variationnelle, qui vise à maximiser une borne inférieure de la vraisemblance $\mathscr{J}(\mathrm{Y},\theta,\widetilde{\mathbb{P}}) := \widetilde{\mathbb{E}}\log\mathbb{P}(\mathrm{Y},\mathrm{Z}) - \widetilde{\mathbb{E}}\log\widetilde{\mathbb{P}}(\mathrm{Z})$ (chapitre 2 section 2.1.3), où $\widetilde{\mathbb{E}} = \mathbb{E}_{\widetilde{\mathbb{P}}}$. L'étape VE qui en découle revient à mettre à jour la distribution conditionnelle approchée $\widetilde{\mathbb{P}}(\mathrm{Z})$ selon

$$\widetilde{\mathbb{P}}^{h+1} = \arg\max_{\widetilde{\mathbb{P}}} \mathscr{J}(\mathrm{Y},\theta^{h+1},\widetilde{\mathbb{P}}) = \arg\min_{\widetilde{\mathbb{P}}} \mathrm{KL}\left[\widetilde{\mathbb{P}}(\mathrm{Z})||\mathbb{P}(\mathrm{Z}|\mathrm{Y};\theta)\right].$$

Nous adoptons l'approche générale proposée par Saul and Jordan [1996] et adaptée pour les HMM couplés par Ghahramani and Jordan [1997], forçant $\widetilde{\mathbb{P}}$ à être un produit de chaînes de Markov indépendantes.

$$\widetilde{\mathbb{P}}(\mathrm{Z}) = \prod_i \widetilde{\mathbb{P}}(\mathrm{Z}_i) \qquad \text{où} \qquad \widetilde{\mathbb{P}}(\mathrm{Z}_i) = \prod_i \widetilde{\mathbb{P}}(\mathrm{Z}_{i,1}) \prod_{t\geq 2} \widetilde{\mathbb{P}}(\mathrm{Z}_{i,t}|\mathrm{Z}_{i,t-1}).$$

Nous adoptons la même paramétrisation que Ghahramani and Jordan [1997] :

$$\widetilde{\mathbb{P}}(\mathrm{Z}_i) = \frac{1}{\tilde{\mathrm{C}}_i}\left(\prod_q (m_q h_{i1}^q)^{\mathrm{Z}_{i1}^q}\right)\prod_{t\geq 2}\left(\prod_{q,r}(\pi_{qr}h_{it}^r)^{\mathrm{Z}_{i,t-1}^q \mathrm{Z}_{i,t-1}^r}\right)$$

où $\tilde{\mathrm{C}}_i$ est une constante de normalisation assurant que les $\widetilde{\mathbb{P}}(\mathrm{Z}_i)$ somment à 1. Les paramètres variationnels h_{it}^r peuvent être vus comme des termes de corrections selon une chaîne de Markov de paramètres (m,π).

On note $\tau_{it}^r = \widetilde{\mathbb{E}}\mathrm{Z}_{it}^r$ et $\Delta_{it}^{qr} = \widetilde{\mathbb{E}}(\mathrm{Z}_{i,t-1}^q \mathrm{Z}_{i,t}^r)$ et

$$\log\Omega_{it}^r = \log w\left[\sum_{j\neq i} s_{ij}(1-\tau_{jt}^r)\right].$$

En utilisant les propriétés de factorisation de la distribution approchée proposée $\widetilde{\mathbb{P}}$, la borne inférieure $\mathscr{J}(\mathrm{X},\theta,\widetilde{\mathbb{P}}^h)$ donnée par (4.2) s'écrit

$$\begin{aligned}
\mathscr{J}(\mathrm{Y},\theta,\widetilde{\mathbb{P}}^h) &= \sum_{ir}\tau_{i1}^r[\log m_r + \log\phi_r(\mathrm{Y}_{i1}) - \log(m_q h_{i1}^r)] \\
&\quad + \sum_{i,t\geq 2,r}\tau_{it}^r[\log\Omega_{it}^r + \log\phi_r(\mathrm{Y}_{it})] - \log\mathrm{C} + \sum_i \log\tilde{\mathrm{C}}_i \\
&\quad + \sum_{i,t\geq 2,q,r}\Delta_{it}^{qr}[\log\pi_{qr} - \log(\pi_{qr}h_{it}^r)] \\
&= \sum_{itr}\tau_{it}^r[\log\phi_r(\mathrm{Y}_{it}) + \log\Omega_{it}^r - \log h_{it}^r] - \log\mathrm{C} + \sum_i \log\tilde{\mathrm{C}}_i
\end{aligned}$$

parce que $\widetilde{\mathbb{E}}(\mathrm{Z}_{it}\mathrm{Z}_{jt}^r) = \tau_{it}\tau_{jt}^r$ pour tout $i \neq j$ et parce que $\sum_q \Delta_{it}^{qr} = \tau_{it}^r$.

Etape VE L'étape VE consiste à trouver à la fois la valeur optimale pour les paramètres variationnels (h_{it}^r) et à calculer les moments conditionnels approchés τ_{it}^r et Δ_{it}^{qr}. Suivant

Ghahramani and Jordan [1997], Annexe D, on obtient

$$\frac{\partial \mathscr{J}(\mathrm{Y},\theta,\widetilde{\mathbb{P}}^h)}{\partial \log h^r_{it}} = \left[\log\phi_r(\mathrm{Y}_{it}) + \log\Omega^r_{it} - \log h^r_{it}\right]\frac{\partial \tau^r_{it}}{\partial \log h^r_{it}} - \tau^r_{it} + \tau^r_{it}$$

comme C ne dépend pas de h^r_{it} et $\partial \log \tilde{\mathrm{C}}_i / \partial \log h^r_{it} = \tau^r_{it}$. Cette dérivée s'annule pour

$$h^r_{it} = \phi_r(\mathrm{Y}_{it})\Omega^r_{it}. \tag{3.6}$$

Les moments conditionnels, qui ne dépendent pas des constantes de normalisation $\tilde{\mathrm{C}}_i$, sont ensuite calculés en utilisant un algorithme forward-backward indépendant pour chaque individu i :

— Etape Forward : soit $\mathrm{F}^q_{i,1} \propto m_q h^q_{i1}$ et, pour $t \geq 2$, on calcule

$$\mathrm{F}^r_{i,t} \propto \sum_q \mathrm{F}^q_{i,t-1}\pi_{qr}h^r_{it};$$

— Etape Backward : $\tau^r_{i\mathrm{T}} = \mathrm{F}^r_{i,\mathrm{T}}$ et pour $1 \leq t \leq \mathrm{T}-1$, on calcule

$$\mathrm{G}^r_{i,t+1} = \sum_q \mathrm{F}^q_{i,t}\pi_{qr}, \quad \Delta^{qr}_{it} = \pi_{qr}\frac{\tau^r_{i,t+1}}{\mathrm{G}^r_{i,t+1}}\mathrm{F}^q_{i,t}, \quad \tau^q_{it} = \sum_r \Delta^{qr}_{it}.$$

3.3.3 Etudes de simulation

Afin d'évaluer les performances, à la fois en terme de temps de calcul et de précision, de la méthode d'inférence approchée que nous proposons, appelée *CHMM-VEM* pour Variational EM for Coupled HMM, nous avons effectué deux études de simulation. L'étude 1 compare le temps de calcul obtenu par rapport à une version exacte de l'algorithme EM, appelée *CHMM-EM* pour EM for Coupled HMM. La deuxième étude illustre l'importance de prendre en compte la dépendance et compare notre modèle à un modèle HMM indépendant. Afin de ne pas biaiser la comparaison, nous supposons que les paramètres d'émission sont communs à toutes les séries. Ils sont estimés par un algorithme EM. Ce modèle est noté *iHMM-EM*. Notons qu'il est équivalent au modèle *CHMM-EM* avec $\omega = 1$.

Plan de simulation et critères de comparaison

Plan de simulation Dans l'étude 1, nous considérons un nombre croissant d'individus $\mathrm{I} \in \{2,3,4,5\}$. Le nombre d'individus est fixé à $\mathrm{I} = 10$ dans l'étude 2. Pour les deux études, la longueur de la série est fixée à $\mathrm{T} = 1000$ points et le nombre d'états cachés à $\mathrm{Q} = 3$. Nous utilisons des lois d'émission gaussiennes de moyenne respectivement -1, 0 et 1 et considérons une séquence de bruit croissant : $\sigma \in \{0.3, 1, 1.2\}$. A noter que plus σ est grand, plus le problème de détection se complique.

Le terme de corrélation W_ℓ utilisé dans (3.5) dépend à la fois des similarités et du paramètre ω. Ici, dans le but de se rapprocher au plus des situations réelles, nous tirons la matrice de similarité $(s_{i,j})_{i,j}$ ($[I \times I]$ à partir d'une vraie matrice d'apparentement issue de 336 lignées de maïs [Darracq et al., 2016], qui sera utilisée dans l'appication sur données réelles. Pour ω, nous considérons deux valeurs correspondant à deux niveaux de corrélation entre les profils : un cas de dépendance que nous qualifions de modérée (telle que $-20\log\omega = 7$) et un cas de dépendance faible (telle que $-20\log\omega = 4$). Nous simulons ensuite les états cachés de la façon suivante :

(1) nous fixons les positions altérées centrales toutes les 50 positions, c'est-à-dire aux positions $25, 75, \ldots, 975$;

(2) autour de chacune de ces positions centrales, nous fixons une fenêtre de longueur distribuée selon une loi de Poisson de moyenne 15, de telle sorte que les altérations soient de longueur variable ;

(3) pour chaque fenêtre, nous échantillonnons ℓ ($1 \leq \ell \leq K$) statuts individuels avec une probabilité proportionnelle à W_ℓ (3.5), de telle sorte que chaque altération ne soit pas portée que par tous les individus.

Nous simulons chaque configuration (σ, ω) 100 fois. Pour chaque jeu de données simulées, nous faisons tourner à la fois iHMM$-$EM et CHMM$-$EM, les loci sont classés en utilisant l'algorithme de Viterbi.

Critères de comparaison L'étude 1 se base sur le temps de calcul mesuré en seconde. A noter que l'algorithme forward-backward est codé en C++, et que le reste est implémenté en langage R [R Core Team, 2016]. Nous considérons la classification normale 0 et altérée(-1 ou 1). Afin d'évaluer la performance des modèles, nous considérons les critères ci-dessous :

— le taux de faux positifs (FPR) : proportion d'altérations détectées à tort,
— le taux de faux négatifs (FNR) : proportion de statuts 0 (normaux) estimés de façon erronée parmi les statuts estimés normaux,
— l'exactitude de classification : proportion de statuts estimés correctement.

Pour chaque configuration, nous considérons la moyenne de ces critères sur les 100 jeux de données simulées.

Choix du paramètre ω Comme dit précédemment, la procédure proposée ne permet pas d'estimer le paramètre ω. Afin de le sélectionner, nous proposons la stratégie suivante : faire varier ω sur sur une grille de valeurs et sélectionner celui qui minimise la somme des carrés des résidus pondérés (RSS pour Residuals Sum of Squares) :

$$\mathrm{RSS}_\omega = \sum_{i,t,r} \tau_{i,t}^r (y_{i,t} - \mu_r)^2.$$

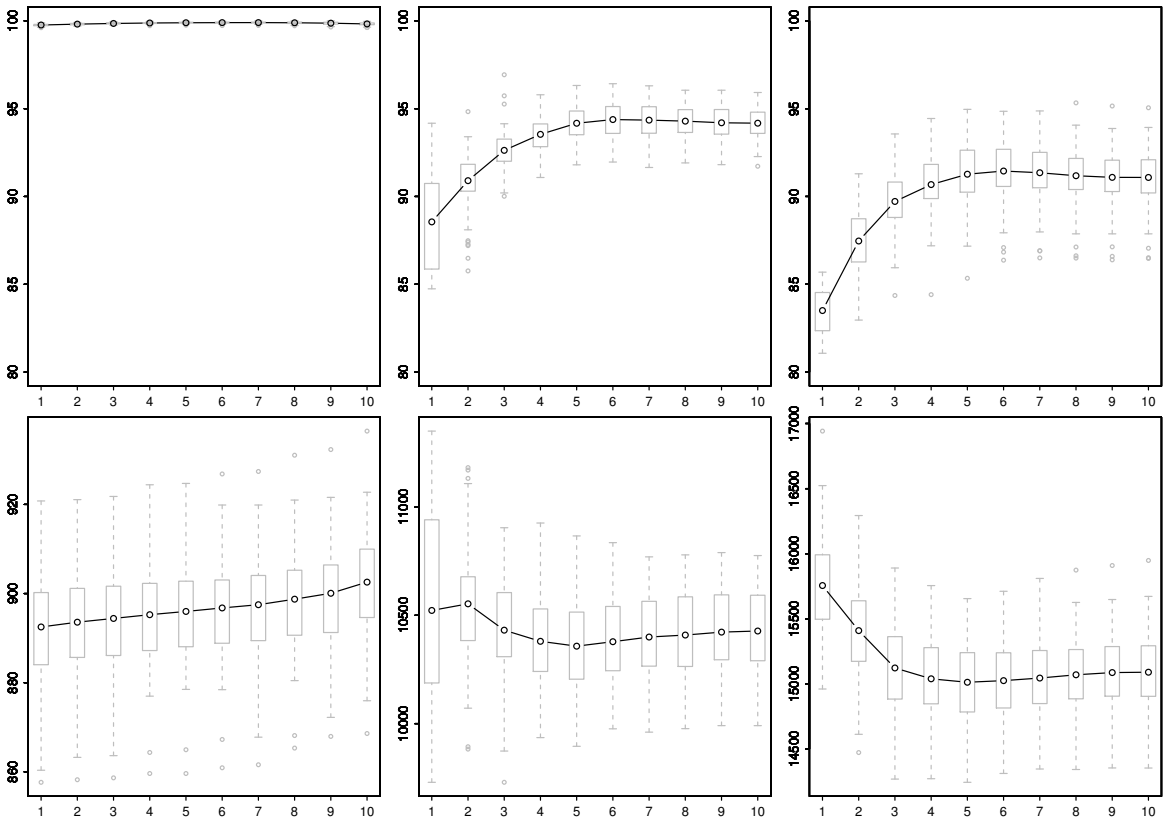


FIGURE 3.6 – Boxplot de l'exactitude de classification (%, en haut) et du critère RSS_ω (bas) pour différentes valeurs de $\omega \in \{e^{-k/20} | k = 1,2,\ldots,10\}$. A gauche : $\sigma = 0.3$, au milieu : $\sigma = 1$, à droite : $\sigma = 1.2$.

La figure 3.6 donne à la fois l'exactitude de classification et le critère RSS_ω pour différentes valeurs de $\omega \in \{e^{-k/20} | k = 1,2,\ldots,10\}$, différentes valeurs σ, I = 10 et dans un cas de dépendance faible. Nous observons que quand σ est faible, le problème de segmentation est évident et le choix de ω n'affecte pas l'exactitude de classification. Pour de plus grandes valeurs de σ, nous observons que la courbe du RSS_ω a un minimum et que ce dernier est proche du maximum d'exactitude de classification.

Etude 1

Seuls les résultats dans le cas de dépendance faible avec $\sigma = 1$ sont présentés, les autres configurations amènent aux mêmes conclusions. La table 3.1 donne la médiane du temps de calcul en secondes sur un PC avec 3.2GHz, pour un nombre croissant d'individus. La figure 3.7 représente l'exactitude de classification pour le cas I = 3 individus.

Comme attendu, *CHMM-EM* donne des résultats légèrement meilleurs en terme d'exactitude de classification que *CHMM-VEM* suivi par *iHMM-EM*. Cependant, le temps de calcul de *CHMM-EM* augmente de façon exponentielle avec le nombre d'individus I et ne peut pas être utilisé pour un plus grand nombre de lignées, même assez faible. A noter que *CHMM-VEM* est bien plus rapide que *iHMM-EM*. Nous observons déjà que la prise en compte de la dépendance entre les individus améliore la classification.

TABLE 3.1 – Temps de calcul (en secondes) en fonction du nombre d'individus I pour les 3 procédures

I	*iHMM-EM*	*CHMM-VEM*	*CHMM-EM*
2	0.8	0.4	2.0
3	1.1	0.5	11.2
4	1.2	0.6	79.4
5	1.6	0.8	920.2

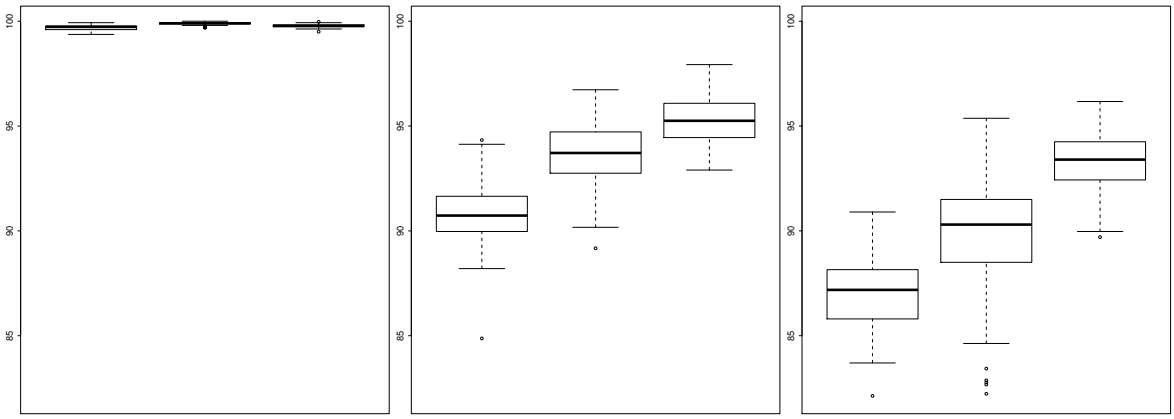


FIGURE 3.7 – Exactitude de classification (%) *iHMM-EM* (à gauche), *CHMM-VEM* (au centre) et *CHMM-EM* (à droite) pour I = 3. A gauche : $\sigma = 0.3$. Au centre : $\sigma = 1$. A droite : $\sigma = 1.2$

Etude 2

Pour chaque configuration, ω a été choisi selon la stratégie décrite dans le paragraphe 3.3.3. Sur la Figure 3.8, on observe que quand σ est petit, c'est-à-dire quand le problème de classification est facile, à la fois *iHMM-EM* et *CHMM-VEM* donnent de très bons résultats. Cependant, quand σ augmente, *CHMM-VEM* donne de meilleurs résultats que *iHMM-EM* quelque soit la dépendance, montrant ainsi l'importance de prendre en compte une dépendance existante entre les individus. La différence entre les procédures s'observe d'autant plus que le niveau de dépendance est grand.

3.3.4 Application sur données réelles

Description des données

Nous considérons un jeu de données issues de puces de génotypage Illumina sur un panel de I = 336 lignées de maïs [Darracq et al., 2016]. Les données dont nous disposons correspondent à des valeurs de LRR (Log Red Ratio), calculées à l'aide du logiciel GenomeStudio d'Illumina :

$$Y_{it} = \log_2 \frac{R_{it}^{\texttt{observé}}}{R_t^{\texttt{attendu}}},$$

où $R_{it}^{\texttt{observé}}$ est l'intensité du signal normalisé à la position (locus) t dans la lignée i et $R_t^{\texttt{attendu}}$ est l'intensité de référence à la position t. Dans ce panel, seulement deux cas

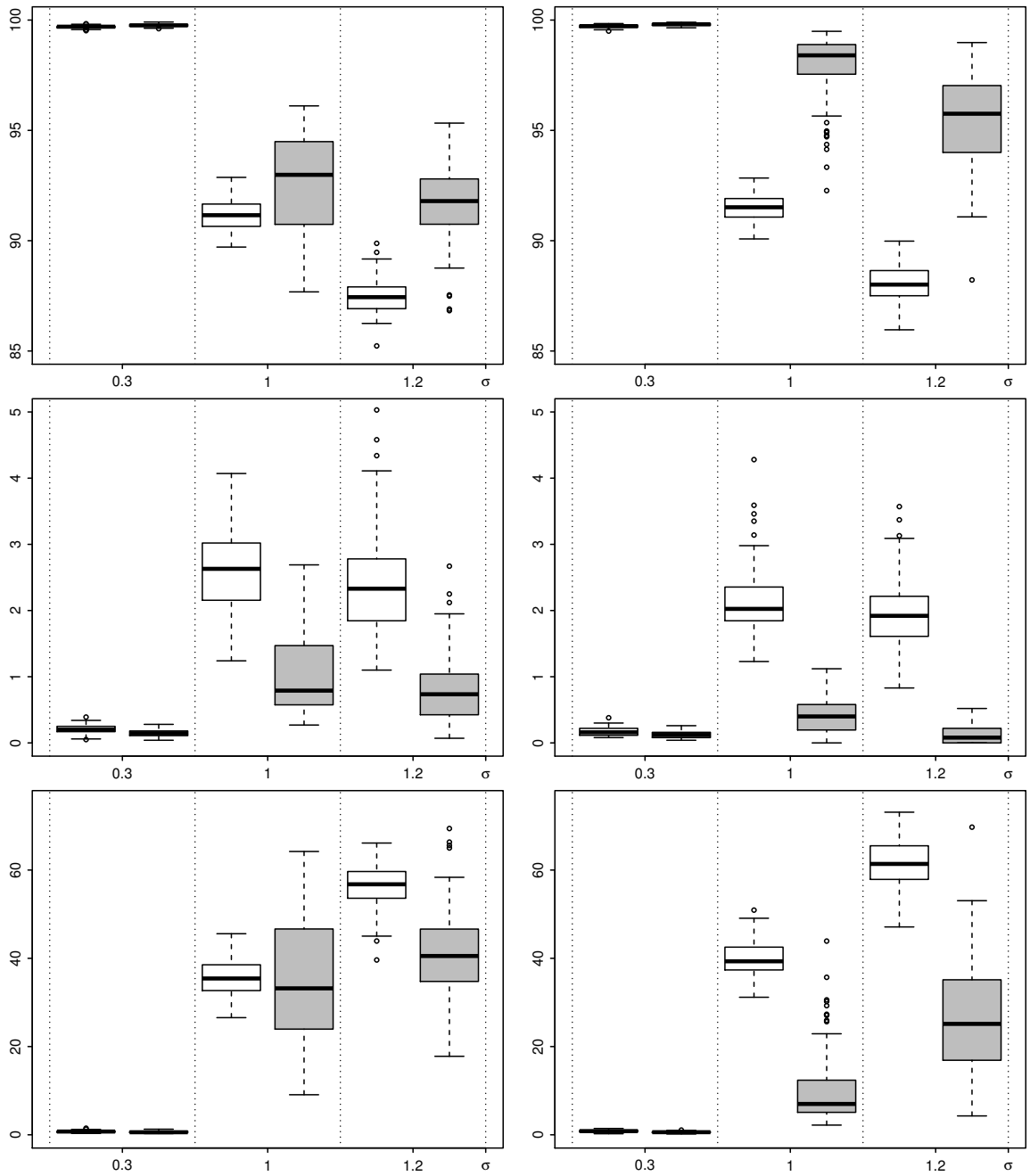


FIGURE 3.8 – A gauche : cas de faible dépendance. A droite : cas de dépendance modérée. Boîtes à moustaches de l'exactitude de classification (en %, en haut), du FPR (%, au centre) et du FNR (%, en bas) pour différentes valeurs de σ (axe des abscisses). Pour chaque σ, on représente *iHMM-EM* (boîte blanche) et *CHMM-VEM* (boîte grise).

TABLE 3.2 – Comparaison des classifications obtenues (nombre de loci classés respectivement dans l'état 'délété' et 'normal') en analysant les 4 groupes séparément ou ensemble (un groupe)

		4 groupes	
		Délétion	Normal
1 groupe	Délétion	1469821	49679
	Normal	59456	17082820

sont envisageables : soit la lignée i testée partage un locus t avec le génome de référence et Y_{it} est proche de 0 (cas normal) ou le locus t n'existe pas dans le génome de la lignée i et Y_{it} est inférieur à 0 (cas altéré).
En plus de ces données, nous disposons de la matrice d'apparentement [Astle and J.Balding, 2009] $[s_{ij}]$ entre les lignées.

Résultats

Comme expliqué précédemment, nous nous attendons à avoir Q = 2 états. Afin de valider cette valeur, nous avons ajusté un modèle HMM pour chaque lignée avec respectivement Q = 2, 3 et 4 états. La figure 3.9 confirme que deux modes apparaissent, justifiant notre choix de Q = 2.

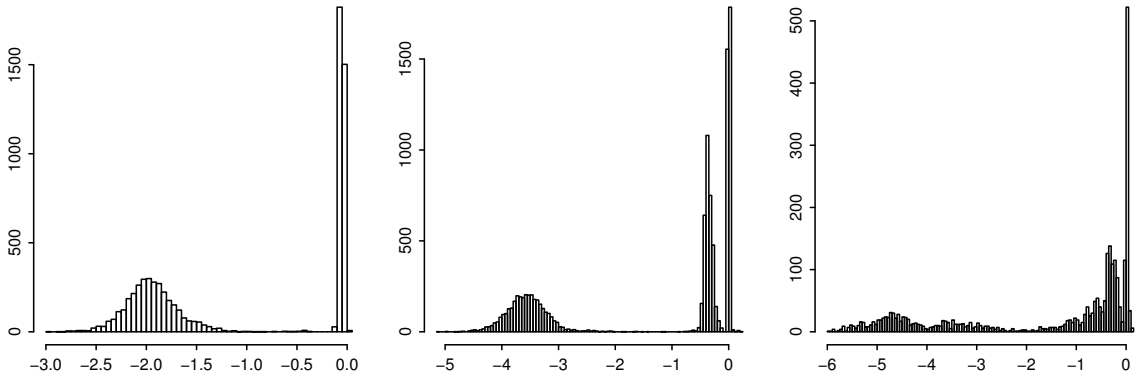


FIGURE 3.9 – Histogramme de la moyenne estimée par des HMM indépendants pour les 336 lignées, avec Q = 2 à gauche, Q = 3 (au centre) et Q = 4 à droite.

Comme vu dans la section 3.3.3, la prise en compte de la dépendance entre individus quand elle existe améliore la classification, et ce, d'autant plus que la corrélation entre individus est forte. Afin d'obtenir des groupes bien corrélés, nous avons divisé 336 individus en 4 groupes définis à partir d'une classification ascendante hiérarchique et avons analysé chaque groupe indépendamment. Ces groupes sont représentés sur la figure 3.10. Nous observons dans le tableau 3.2 qu'analyser plusieurs groupes, bien corrélés entre eux, séparément permet de détecter près de 10^4 délétions supplémentaires par rapport au cas où tous les individus sont analysés conjointement.

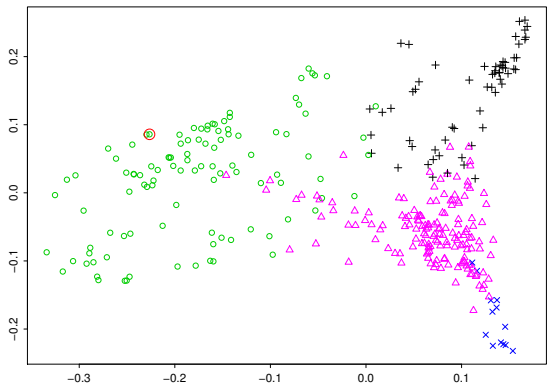


FIGURE 3.10 – Positionnement des 336 lignées à partir de leur matrice d'apparentement. Chaque symbole représente un groupe différent. La lignée Fν2, lignée de référence française, est entourée en rouge.

La figure 3.11 représente la corrélation entre la matrice de similarité originelle et la matrice de corrélation entre les lignées estimée respectivement par *iHMM-EM* et *CHMM-VEM*. On remarque sur cette figure que le fait de prendre en compte la dépendance entre les individus dans la méthode *CHMM-VEM* permet d'estimer une matrice de corrélation plus proche de la structure de similarité initiale.

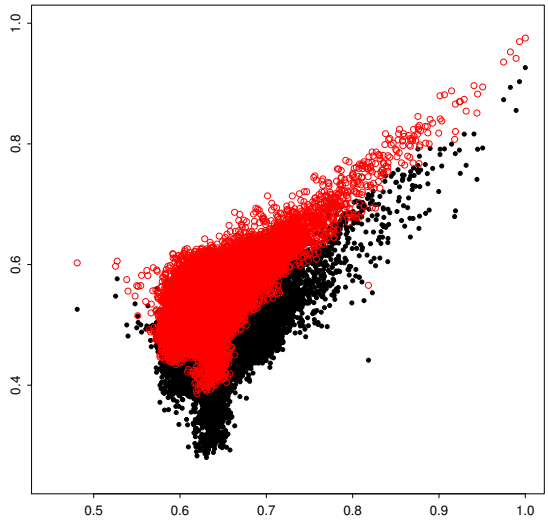


FIGURE 3.11 – Corrélation entre la matrice de similarité donnée et la matrice de corrélation estimée respectivement par les méthodes *iHMM-EM* (en noir) et *CHMM-VEM* (en rouge)

La figure 3.12 donne le nombre de loci classés comme déletés spécifiques à chacune des deux méthodes *iHMM-EM* et *CHMM-VEM*, et communs aux deux méthodes. Nous observons que la méthode *iHMM-EM* détecte plus de loci en délétion que la méthode *CHMM-VEM*. En l'absence de connaissance de la vérité, il est difficile de dire quelle méthode est la meilleure. Il faudrait pouvoir valider ou invalider les locis trouvés spécifiquement par chacune des méthodes. Toutefois, l'étude de simulation précédente nous pousse à penser qu'une partie des locis détectés par la méthode *iHMM-EM* pourraient

TABLE 3.3 – Exactitude de classification des méthodes *iHMM-EM* et *CHMM-VEM*. I : taille du panel. $\overline{s}_I$: apparentement moyen au sein du panel considéré. FPR et FNR sur les 58 altérations de Fv2.

	iHMM-EM	*CHMM-VEM*				
I	1	6	49	80	153	336
$\overline{s}_I$	1.00	0.75	0.71	0.67	0.65	0.64
FPR(%)	12.68	10.43	10.02	9.32	8.89	8.95
FNR(%)	24.14	24.14	24.14	25.86	25.86	25.86

être des faux positifs.

Nous disposons tout de même de connaissances partielles pour comparer la qualité de la classification. En effet, 58 loci ont été détectés par séquençage (Stéphane Nicolas, communications personnelles) dans la lignée Fv2, lignée de référence française. Nous utilisons ces $loci$ pour comparer les performances en terme de classification des deux méthodes. Nous étudions aussi comment le nombre de lignées choisies dans le panel pour une analyse conjointe, influence les résultats. Pour ce faire, nous avons ordonné les lignées de façon décroissante en fonction de leur apparentement à la lignée Fv2 et défini alors différents panels de taille croissante. Nous observons (tableau 3.3) qu'une analyse jointe de lignées corrélées réduit la proportion de délétions détectées à tort.

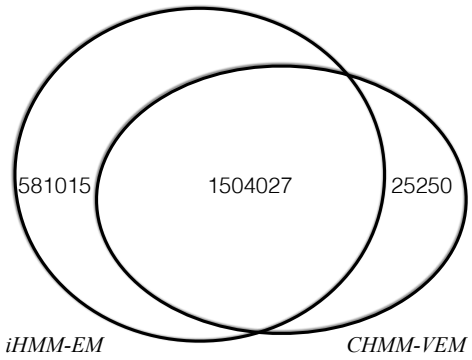


FIGURE 3.12 – Diagramme de Venn des locis classés en délétion par respectivement *iHMM-EM* et *CHMM-VEM*.

3.3.5 Discussion

Nous avons proposé une méthode permettant de prendre en compte la dépendance entre séries ainsi que deux algorithmes d'inférence associés. Nous avons confirmé que prendre en compte la dépendance entre les séries, lorsqu'elle existe améliore la classification des individus. L'article associé à cette méthode est en cours de rédaction. Une extension de la méthode aux données appariées (issues notamment de puces à deux couleurs) a également été proposée. Le modèle proposé peut facilement être étendu à des données de séquençage via la spécification d'une loi d'émission adaptée à la nature des

données, comme une loi de Poisson ou binomiale négative.

3.4 Références

W. Astle and D. J.Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24 :451–471, 2009. 56

S. Aubourg, M.-L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, P. Jullien, M. Ingouff, V. Thareau, T. Scheix, A. Lecharny, and J. Renou. Analysis of catma transcriptome data identifies hundreds of novel functional genes and improves gene models in the arabidopsis genome. *BMC Genomics*, 8, 2007. 45

A. Darracq, C. Vitte, S. Nicolas, J. Duarte, J. Pichon, **Aubert, J.**, X. Wang, T. Mary-Huard, C. Chevalier, A. Charcosset, M. Lepaslier, P. Rogowsky, and J. Joets. Sequence analysis of european maize inbred line fv2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *submitted*, 2016. 52, 54

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 :1–38, 1977. 49

B. Efron. Large-scale simultaneous hypothesis testing : the choice of a null hypothesis. *J Amer Statist Assoc*, 99, 2004. doi : 10.1198/016214504000000089. URL http://dx.doi.org/10.1198/016214504000000089. 37

S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J. Renou, A. Lecharny, and V. Brunaud. Catdb : a public access to arabidopsis transcriptome data from the urgv-catma platform. *Nucleic Acids Research*, pages 986–990, 2008. 45

Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine learning*, 29 (2-3) :245–273, 1997. 50, 51

F. Picard, M.-L. Martin-Magniette, S. Gagnot, V. Brunaud, **Julie Aubert**, V. Gendrel, S. Robin, M. Caboche, A. Lecharny, and V. Colot. Mixthres : mixture models to define a hybridization threshold in dna microarray experiments. Technical Report 20, SSB, october 2008. 41

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. 52

L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. *Advances in neural information processing systems*, pages 486–492, 1996. 50

J. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100,16 : 9440–9445, 2003. doi : 10.1073/pnas.1530509100. 37, 40

J. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates : A unified approach. *JRSSB*, 66 :187–205, 2004. 39

K. Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9 (1) :1–14, 2008. ISSN 1471-2105. doi : 10.1186/1471-2105-9-303. URL http://dx.doi.org/10.1186/1471-2105-9-303. 41

Aubert, J, A. Bar-Hen, J.-J. Daudin, and S. Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics*, 5 (1) :125, 2004. ISSN 1471-2105. doi : 10.1186/1471-2105-5-125. URL http://www.biomedcentral.com/1471-2105/5/125. 37

Chapitre 4

Modèle à blocs latents pour l'analyse de données de comptage surdispersées - Application en écologie microbienne

Sommaire

L'écologie microbienne à l'ère des nouvelles technologies de séquençage L'écologie microbienne s'intéresse à la façon dont les microorganismes interagissent avec leur environnement, entre eux et avec leur hôtes. Les nouvelles technologies de séquençage (cf. section 1.4) permettent d'aller plus loin dans l'étude de la caractérisation et du fonctionnement des communautés microbiennes dans et avec leur environnement en atteignant les microorganismes non cultivables. Le méta-codebarres ou séquençage d'amplicons de gènes marqueurs permet de répondre à la question 'qui est présent ?'. La métagénomique, la métatranscriptomique ou la génomique basée sur des cellules uniques permettent d'aller plus loin et d'explorer également les aspects fonctionnels et de commencer à répondre à la question 'pour quoi faire ?'.
Les données issues de ces nouvelles technologies posent un certain nombre de défis. Le lecteur peut se référer aux articles de Nayfach and Pollard [2016]; Segata et al. [2012] pour une description de ces défis. Un des grands intérêts de ces données résident dans la possibilité de les comparer, de comparer les communautés bactériennes entre les échantillons biologiques d'une même étude, voire même entre différentes études. Cela nécessite une harmonisation des pratiques et une quantification précise, ainsi que des méthodes statistiques adaptées.

Données analysées Dans les travaux présentés dans cette thèse, relatifs aux données issues de séquençage à haut débit, nous ne nous intéressons pas à la construction de données résumées à partir des données brutes sorties des séquenceurs, mais à l'analyse statistique de données résumées. On entend par données résumées, une matrice de comptage résultant d'un traitement bioinformatique donnant le nombre de séquences associées à chacune des unités d'intérêt dans chacun des échantillons environnementaux dont ils sont issus. Une unité d'intérêt peut être un OTU (en séquençage d'amplicons de gènes marqueurs ou métagénomique) ou un gène (en RNA-Seq ou métatranscriptomique). Un OTU est une unité taxonomique d'intérêt servant de proxy pour les espèces de microorganismes. Dans nos exemples, les unités d'intérêt sont en ligne et les échantillons biologiques en colonne. Cette matrice de comptages peut contenir un grand nombre de zéros. Un zéro peut être interprété comme l'absence de présence (ou d'expression) d'un OTU (ou gène), ou une présence (ou expression) qui n'a pas pu être détectée du fait d'un échantillonnage trop faible.

Analyse différentielle d'abondance ou d'expression Ces données sont par nature des données de comptage. Une présentation des lois adaptées à ces données est disponible en section 2.3. SMarek Gierlinski et al. [2015] ont testé sur des données réelles de RNA-Seq issues de 48 cultures cellulaires de *S. cerevisiae* l'adéquation des distributions binomiale négative (supposées dans les packages Bioconductor DESeq2 [Love et al., 2014] et edgeR [Robinson et al., 2009]), normale et log-normale (supposée dans le package Bioconductor limma [Law et al., 2014]). Ils confirment que, dans le cadre de l'analyse diffé-

rentielle d'expression à partir de données issues du séquençage de l'ARN, la distribution binomiale négative est adaptée aux comptages des lectures pour la grande majorité des gènes. La distribution log-normale pourrait être adaptée mais pose problème quand un ou plusieurs échantillons contiennent des zéros. Les auteurs recommandent donc l'utilisation d'outils basés sur la distribution binomiale négative. McMurdie and Holmes [2014] recommandent l'utilisation de méthodes basées sur des modèles hiérarchiques, comme le modèle Poisson-Gamma ou Béta-binomial pour l'analyse de données de microbiome, après une standardisation utilisant une stabilisation de la variance adéquate. Ils insistent sur le fait que les méthodes de raréfaction, de normalisation par sous-échantillonnage (down-sampling) sont à proscrire. Jonsson et al. [2016] comparent 14 méthodes pour l'analyse différentielle de données de microbiome et confirme les résultats de McMurdie and Holmes [2014], à savoir, que les méthodes basées sur la loi binomiale négative telles que celles implémentées dans DESeq2 et edgeR sont les plus performantes.

Par ailleurs, ces données peuvent comporter un nombre important de zéros. Cela est d'autant plus vrai pour les données de microbiome, et lorsque le niveau taxonomique est bas. Les modèles avec excès de zéros peuvent constituer une alternative intéressante à ceux basés sur la distribution binomiale négative pour rechercher les gènes ou OTUs différentiellement exprimés ou abondants entre plusieurs conditions. Paulson et al. [2013], par exemple, ont développé un modèle de mélange de lois normales avec excès de zéros pour prendre en compte les biais de sous-échantillonnage des communautés microbiennes. Dans le cadre de l'analyse différentielle d'expression issues de données de séquençage de l'ARN, on peut citer le package ShrinkBayes [Van de Wiel et al., 2014] qui implémente notamment des lois de Poisson ou binomiale négative avec excès de zéros. Plus récemment, Zhang et al. [2016] ont proposé un modèle de régression de binomiale négative avec excès de zéros pour identifier les taxons différentiellement abondants entre plusieurs conditions. Chen and Li [2016] proposent un modèle de régression Béta avec excès de zéros et effets aléatoires en deux parties, pour tester l'association entre les abondances microbiennes et des covariables cliniques pour des données de microbiome longitudinales. Leur modèle comprend une composante logistique pour modéliser la présence/absence d'un microbe dans les échantillons et une composante Béta pour modéliser l'abondance des microbes présents. Chaque composante comprend un effet aléatoire pour tenir compte des corrélations entre les mesures répétées d'un même patient.

Classification croisée Ici, nous ne cherchons pas à trouver les variables (OTUs) statistiquement significatives entre un ou plusieurs groupes (correspondant à plusieurs échantillons biologiques) mais nous cherchons à explorer les données par une méthode de classification croisée probabiliste. Ce chapitre présente un modèle à blocs latents modélisant une structure sous-jacente par bloc de lignes (groupes d'OTUs) et colonnes (groupes d'échantillons environnementaux) adapté à des données de comptage issue de l'écolo-

gie microbienne. Tout comme dans le chapitre précédent, une inférence basée sur une approche variationnelle a été proposée. Le vocabulaire utilisé est celui défini dans le chapitre 1 et les notations utilisées sont les mêmes que celles présentées dans le chapitre 2. La suite de cette partie est rédigée en anglais et correspond à l'article en cours de rédaction.

4.1 Introduction

Biclustering aims at simultaneously partition a data matrix into several homogenous blocks constituted by rows and columns groups. It allows to summarize a dataset into a smaller one keeping the same structure. Biclustering approaches have proven useful to discover local patterns in which a subset of features exhibit similar values over a subset of samples. This technique has been applied in various fields such as bioinformatics for example to discover subgroups of co-expressed genes across a subset of biological conditions [Madeira and Oliveira, 2004; Oghabian et al., 2014], or collaborative filtering for example to cluster people based on their similarity of interest (books, films, music) and allows recommendations of new items for people belonging to the same co-cluster [de Castro et al., 2007; Hofmann, 2004]. Two different strategies can be used to this aim : model-based or algorithmic methods. Model-based approaches as mentioned by Warton et al. [2015] has the advantages to be interpretable and flexible. Bouveyron and Brunet [2014] and Melnykov [2016] provided detailed reviews about finite mixture models and model-based clustering. In such models, one challenge is the choice of an adequate emission distribution.

In this paper, we are interested in microbial ecological data provided by high-throuput sequencing. Amplicon-based sequencing and shotgun metagenomics study microbial communities directly from environmental samples. They provide count matrices where rows correspond to operational taxonomic unit (OTU), proxy for species, say bacteria or fungus and columns to sampling units, such as human guts, plant rhizospheres or marine stations for example. One major goal of these projects is to find associations between bacteria communities and sampling units. This goal may be achieved via block-clustering. Recent studies [Lindner and Renard, 2015; White et al., 2009] show that metagenomics data are overdispersed and that Poisson-Gamma mixtures are well adapted for these data [Jonsson et al., 2016; McMurdie and Holmes, 2014]. In this paper, we propose a model for biclustering using mixtures adapted to overdispersed count data. We use the latent block model (LBM) framework introduced by Govaert and Nadif [2012] and for which a recent review is available in Brault and Mariadassou [2015]. We extend a Poisson model quite natural when dealing with abundance matrices to a Poisson-Gamma model, and parametrize our model such that row and columns effects and use of replicates are taken into account . The inference step aims to estimate both the hidden variables and the parameters. Since latent variables are not independent conditionally on observed variables, the classical maximum likelihood inference is intractable and we propose a generalized

Variational Expectation-Maximization algorithms for inference following the strategy of Govaert and Nadif [2010].

The interest of the proposed methodology will be illustrated on three 16S or 18S rRNA amplicon-based datasets. The first one is the GlobalPatterns dataset [Caporaso et al., 2011] which includes biological samples from different environnements and known mock communities. The second one is the MetaRhizo dataset which aims to study the plant-microbial communities interactions in the rhizosphere, i.e. the region of soil directly influenced by root secretions and associated soil microorganisms. The third dataset concerns the oak phyllosphere microbiota [Jakuschkin et al., 2016] which is composed of both fungi and bacteria.

The model framework is described in Section 2, followed by the presentation of the parameter inference strategy in Section 3. We propose a criterion for choosing the number of co-clusters in Section 4. Finally Section 5 presents the application of our model on the three datasets described above.

4.2 Model

4.2.1 Latent Block Model

We consider a $n \times m$ random matrix $\mathbf{Y}$ with entries Y_{ij} representing the count of the feature i in the sample j. The Latent Block Model (LBM) introduced by Govaert and Nadif [2012] associates to each feature i (respectively to each sample j) a latent variable $\mathbf{Z}_i$ (resp. $\mathbf{W}_j$) drawn from a multinomial distribution $\mathcal{M}(1;\pi = (\pi_1,\ldots,\pi_K))$ (resp. $\mathcal{M}(1;\rho = (\rho_1,\ldots,\rho_G))$) where π (resp. ρ) denotes the vector of class proportions in rows (resp. in columns). Only one component of the vector $\mathbf{Z}_i$ and one component of the vector $\mathbf{W}_j$ are not null such that $Z_{ik} = 1$ if feature i belongs to class k and $W_{jg} = 1$ if sample j belongs to class g. So, for all $i \in \{1,\ldots,n\}$, $\sum_{k=1}^{K} Z_{ik} = 1$ and, for all $j \in \{1,\ldots m\}$, $\sum_{g=1}^{G} W_{jg} = 1$.
Y_{ij} are drawn independently from some parametric distribution $F(\gamma_{ij\mathbf{Z}_i\mathbf{W}_j})$. According to this model, the latent variables $(\mathbf{Z}_i)_i$ (respectively $(\mathbf{W}_j)_j$) are independent and identically distributed. Given this latent structure, all the entries Y_{ij} are supposed to be independent.

We denote by θ the vector of model parameters containing the mixing proportions π, ρ and the vector of emission parameters $\gamma = \left(\gamma_{ij\mathbf{Z}_i\mathbf{W}_j}\right)_{i,j,\mathbf{Z}_i,\mathbf{W}_j}$.

Denoting $f(.;\gamma)$ the probability distribution function of $F(\gamma)$, the complete likelihood of $(\mathbf{Y},\mathbf{Z},\mathbf{W})$ factorizes as :

$$\begin{aligned} p(\mathbf{Y},\mathbf{Z},\mathbf{W};\theta) &= p(\mathbf{Z};\pi)\,p(\mathbf{W};\rho)\,p(\mathbf{Y};\mathbf{Z},\mathbf{W},\gamma) \\ &= \prod_{i,k} \pi_k^{Z_{ik}} \prod_{j,g} \rho_g^{W_{jg}} \prod_{i,j,k,g} f(Y_{ij};\gamma_{ijZ_{ik}W_{jg}})^{Z_{ik}W_{jg}}. \end{aligned} \tag{4.1}$$

As samples may correspond to replicates of a same condition we would like to cluster together, we introduce a new subscript $r = 1,\ldots,R_j, j = 1,\ldots,m$ for replicate hierarchi-

zed in condition j. In case of replication, the same framework is used with multivariate $(\mathbf{Y}_{ij})_{i,j} = (\mathrm{Y}_{ijr})_{i,j,r}$, where the replicates $\{\mathrm{Y}_{ijr}\}_r$ are supposed to be independent conditionally on $\mathbf{Z}_i$ and $\mathbf{W}_j$ with respective distribution $\mathrm{F}(\gamma_{ijr\mathbf{Z}_i\mathbf{W}_j})$. Note that the parameter γ does not only depend on $\mathbf{Z}_i$ and $\mathbf{W}_j$ but also directly on i and j, which allows us to introduce both a microorganism specific effect and a sample specific effect.

4.2.2 Poisson-Gamma Latent Block Model

In our motivating examples, observed data are count data that display overdispersion with respect to (wrt) the Poisson distribution. In this article we focus on the negative binomial distribution which is the reference law for data from Next Generation Sequencing [Anders and Huber, 2010; Lindner and Renard, 2015; Robinson et al., 2009; White et al., 2009]. We use the Poisson-Gamma parametrization $\mathscr{P}(\gamma_{ijZ_{ik}W_{jg}}U_{ijr})$ with $(U_{ijr})_{i,j,r}$ distributed according to a gamma distribution with the same scale and shape parameter a, of the negative-binomial distribution with mean $\gamma_{ijZ_{ik}W_{jg}}$ and variance $\gamma_{ijZ_{ik}W_{jg}}(1+1/a)$ so that overdispersion is due to the third hidden layer $\mathbf{U}$ corresponding to unobserved heterogeneity. This model based on multiplicative heterogeneity increases variability in the Poisson mean but in expectation leaves the Poisson mean unchanged. This leads to increased probabilities of occurrences of low and high counts.

To account for the specificities of the data, we introduce row and column effects, denoted by $\mu = (\mu_i)_i, i = 1,\dots,n$ and $\nu = (\nu_{jr})_{j,r}, j = 1,\dots,m, r = 1,\dots,\mathrm{R}_j$ respectively. The row effects may be interpreted as a specific microorganism effect (e.g. its mean abundance in a whole range of media) while the column effect may represent a sampling effort. This leads to the following model :

1. $(\mathbf{Z}_i)_i$ independent and identically distributed $\sim \mathscr{M}(1;\pi = (\pi_1,\dots,\pi_\mathrm{K}))$
2. $(\mathbf{W}_j)_j$ independent and identically distributed $\sim \mathscr{M}(1;\rho = (\rho_1,\dots,\rho_\mathrm{G}))$
3. $(\mathrm{Y}_{ijr})_{i,j,r}$ independent $|(\mathbf{Z}_i)_i,(\mathbf{W}_j)_j \sim \mathscr{P}(\mu_i\nu_{jr}\alpha_{\mathbf{Z}_i\mathbf{W}_j}\mathrm{U}_{ijr})$
4. $(\mathrm{U}_{ijr})_{i,j,r}$ independent and identically distributed $\sim \mathscr{G}(a,a)$

where $\mathscr{P}$ is the Poisson distribution, $\mathscr{G}$ is the gamma distribution. We denote $\alpha = (\alpha_{kg})$.

As discussed in Brault and Mariadassou [2015] and Keribin et al. [2014], this model, like mixture models, is not identifiable but an identifiability up-to a label permutation is sufficient. Here ν is supposed to be given.

Denoting $\underline{\mathrm{H}} = (\mathbf{Z},\mathbf{W},\mathbf{U}) = ((\mathbf{Z}_i)_i,(\mathbf{W}_j)_j,(\mathrm{U}_{ijr})_{i,j,r})$ the set of hidden variables, we can

write the joint log-likelihood of $(\mathbf{Y},\mathbf{H})$ (also called *complete* log-likelihood) as :

$$\begin{aligned}
\log p(\mathbf{Y},\mathbf{H};\theta) &= \log p(\mathbf{Z};\pi) + \log p(\mathbf{W};\rho) + \log p(\mathbf{U};a) + \log p(\mathbf{Y};\mathbf{H},\mu,\nu,\alpha) \\
&= \sum_{i,k} Z_{ik}\log\pi_k + \sum_{j,g} W_{jg}\log\rho_g + \sum_{i,j,r}\left(a\log a - \log\Gamma(a) + (a-1)\log U_{ijr} - aU_{ijr}\right) \\
&- \sum_{i,j,r,k,g} Z_{ik}W_{jg}U_{ijr}\left(\mu_i\nu_{jr}\alpha_{kg}\right) + \sum_{i,j,r,k,g} Z_{ik}W_{jg}Y_{ijr}\log\left(\mu_i\nu_{jr}\alpha_{kg}\right) \\
&+ \sum_{i,j,r,k,g} Z_{ik}W_{jg}Y_{ijr}\log U_{ijr} - \sum_{i,j,r,k,g} Z_{ik}W_{jg}\log y_{ijr}!
\end{aligned}$$

where $\Gamma(.)$ is the Gamma function.

4.3 Inference

4.3.1 Variational approximation principle

Maximum likelihood inference of incomplete data model$(\mathbf{Y};\mathbf{H},\theta)$ is traditionally achieved via the Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. It iteratively maximizes the likelihood of the observed data alternating two steps. The E-step is devoted to the calculation of the conditional distribution of the latent variables given the data, and the M-step to the maximization of the conditional expectation of the complete log-likelihood. Unfortunately, in the case of latent block models, the latent variables are not independent given the observations (see Figure 4.1).

Govaert and Nadif [2008] propose a variational approach [Jordan et al., 1999] that aims at maximizing a lower bound of the log-likelihood of the observed data using a tractable distribution $q(\mathbf{H})$. Indeed, for any distribution q for $\mathbf{H}$, we have that :

$$\begin{aligned}
\log p(\mathbf{Y};\theta) &\geq \log p(\mathbf{Y};\theta) - \mathcal{KL}\left[q(\mathbf{H})||p(\mathbf{H}|\mathbf{Y};\theta)\right] \\
&= \mathbb{E}_q[\log p(\mathbf{Y},\mathbf{H};\theta)] + \mathcal{H}[q(\mathbf{H})] =: \mathcal{J}(\mathbf{Y},\theta,q) \qquad (4.2)
\end{aligned}$$

where $\mathcal{KL}[q(\mathbf{H})||p(\mathbf{H}|\mathbf{Y};\theta)]$ is the Küllback-Leibler divergence between distributions $q(\mathbf{H})$ and $p(\mathbf{H}|\mathbf{Y};\theta)$, $p(\mathbf{H}|\mathbf{Y};\theta)$ is the true conditional distribution of the latent variables $\mathbf{H}$ given the data $\mathbf{Y}$ and $q(\mathbf{H})$ is an approximation of this distribution. $\mathcal{H}[q(\mathbf{H})]$ is the Shannon entropy of distribution q.

The inference strategy then consists in maximizing the lower bound $\mathcal{J}(\mathbf{Y},\theta,q)$ wrt the parameter θ using a variational version of the EM algorithm, alternating two steps. The first one (VE-step) computes the approximate conditional distribution q given the current parameters and the second (M-step) uses this approximate distribution to infer the parameters maximizing the lower bound.

The quality of this approximation mostly relies on the class of approximating distributions within which q is searched for. We consider the class $\mathcal{Q}$ of factorisable distributions

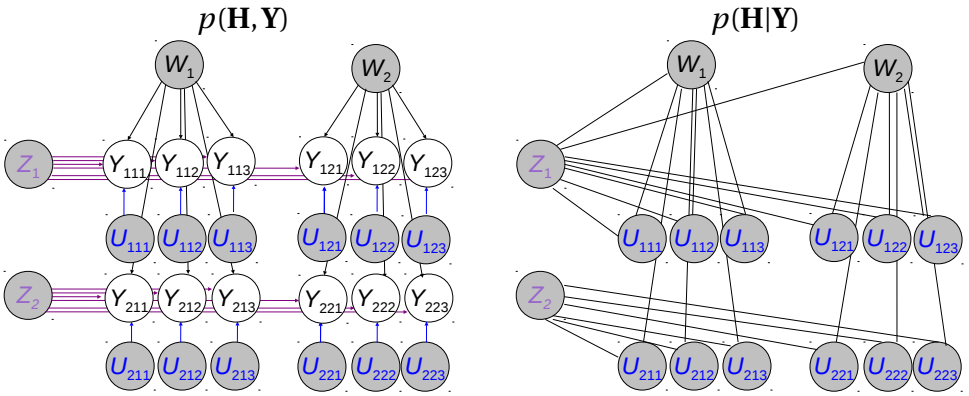


FIGURE 4.1 – Graphical representation of the dependency structure. Left : Latent space model as a directed probabilistic graphical model. Right : conditional distribution of the latent variables as an undirected probabilistic graphical model. Legend : Observed variables (filled white), latent variables (filled gray)

and look for the best approximation in term of Küllback-Leibler divergence in this class :

$$\mathcal{Q} = \{q : q(\mathbf{H}) = \prod_{\ell} q_\ell(\mathbf{H}_\ell) = q_1(\mathbf{Z})\, q_2(\mathbf{W})\, q_3(\mathbf{U})\}. \tag{4.3}$$

This corresponds to an approximation framework developed initially in physics called mean field theory [Jaakkola, 2000].

4.3.2 Inference in the Poisson-Gamma mixture model

Variational E-step or approximate conditional distributions

We denote $\mathbb{E}_q$ the expectation wrt distribution q. We further define $s_{ik} := \mathbb{E}_q(Z_{ik})$ and $t_{jg} := \mathbb{E}_q(W_{jg})$. Because of the factorization properties of the distributions from class $\mathcal{Q}$ defined in (4.3), the lower bound $\mathcal{J}(\mathbf{Y}, \theta, q)$ given in (4.2) writes

$$\begin{aligned}
\mathcal{J}(\mathbf{Y}, \theta, q) \;=\; & \sum_{i,k} s_{ik} \log \pi_k + \sum_{j,g} t_{jg} \log \rho_g + \sum_{i,j,r} \left(a \log a - \log \Gamma(a) + (a-1)\mathbb{E}_q(\log U_{ijr}) - a\mathbb{E}_q(U_{ijr}) \right) \\
& - \sum_{i,j,r,k,g} s_{ik} t_{jg} \mathbb{E}_q(U_{ijr}) \left(\mu_i \nu_{jr} \alpha_{kg} \right) + \sum_{i,j,r,k,g} s_{ik} t_{jg} y_{ijr} \log \left(\mu_i \nu_{jr} \alpha_{kg} \right) \\
& + \sum_{i,j,r,k,g} s_{ik} t_{jg} Y_{ijr} \mathbb{E}_q(\log U_{ijr}) - \sum_{i,j,r,k,g} s_{ik} t_{jg} \log y_{ijr}! \\
& - \sum_{i,j} s_{ik} \log s_{ik} - \sum_{j,g} t_{jg} \log t_{jg} + \mathcal{H}[q(\mathbf{U})].
\end{aligned} \tag{4.4}$$

The optimization of this lower bound wrt the approximate distribution of each hidden variable and the constraints $(s_{ik})_k \in [0,1]^K$ and $\sum_k s_{ik} = 1$ for all i, similarly for $(t_{jg})_g$, produces multinomial distributions for $q_1(\mathbf{Z})$ and $q_2(\mathbf{W})$ and a gamma distribution for $q_3(\mathbf{U})$ (see Appendix for details). And the approximate conditional moments s_{ik}, t_{jg}, $\mathbb{E}_q(U_{ijr})$ and $\mathbb{E}_q(\log U_{ijr})$ result from closed-form interdependent variational update equations

which must be solved iteratively :

$$s_{ik} \propto \pi_k \prod_{j,g} \exp(\mathrm{A}_{ijrkg})^{\mathbb{E}_q(\mathrm{W}_{jg})}, \qquad t_{jg} \propto \rho_g \prod_{i,k} \exp(\mathrm{A}_{ijrkg})^{\mathbb{E}_q(\mathrm{Z}_{ik})},$$

$$\mathbb{E}_q(\mathrm{U}_{ijr}) = \frac{\widetilde{a}_{ijr}}{\widetilde{b}_{ijr}}, \qquad \mathbb{E}_q(\log \mathrm{U}_{ijr}) = \psi(\widetilde{a}_{ijr}) - \log \widetilde{b}_{ijr}$$

$$\mathrm{A}_{ijrkg} = \mu_i \nu_{jr} \alpha_{k,g} \mathbb{E}_q(\mathrm{U}_{ijr}) + y_{ijr} \mathbb{E}_q(\log \mathrm{U}_{ijr}),$$

$$\widetilde{a}_{ijr} = a + \sum_{i,j,r} s_{ik} t_{jg} \mu_i \nu_{jr} \alpha_{kg}, \qquad \widetilde{b}_{ijr} = a + \sum_{k,g} s_{ik} t_{jg} y_{ijr}$$

and ψ the digamma function (i.e. the logarithmic derivative of the Gamma function).

Variational M-step

Keeping now q fixed, we maximize the quantity $\mathscr{J}(\mathrm{Y}, \theta, q)$ wrt θ and the constraint $(\mathrm{KG})^{-1} \sum_{k,g} \alpha_{kg} = 1$ and update the parameter value to this maximiser. This maximization is quite easy as the entropy term $\mathscr{H}[q(\mathbf{H})]$ does not involve θ. This leads to

$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} s_{ik}, \qquad \widehat{\rho}_g = \frac{1}{m} \sum_{j=1}^{m} t_{jg},$$

$$\widehat{\alpha}_{kg} = \frac{\mathrm{KG} \beta_{kg}}{\sum_{k,g} \beta_{kg}}, \qquad \widehat{\mu}_i = \frac{\sum_{j,r,k,g} s_{ik} t_{jg} y_{ijr}}{\sum_{j,r,k,g} s_{ik} t_{jg} \alpha_{kg} \nu_{jr} \mathbb{E}_q(\mathrm{U}_{ijr})}$$

with $\beta_{kg} = \sum_{i,j,r} s_{ik} t_{jg} y_{ijr} \big/ \left[\sum_{i,j,r} s_{ik} t_{jg} \mu_i \nu_{jr} \mathbb{E}_q(\mathrm{U}_{ijr})\right]$. The scale parameter $\widehat{a}$ of the Gamma distribution is solution of the following equation

$$n \mathrm{R}_+ (1 + \log \widehat{a} - \psi(\widehat{a})) + \sum_{i,j,r} \mathbb{E}_q(\log \mathrm{U}_{ijr}) - \sum_{i,j,r} \mathbb{E}_q(\mathrm{U}_{ijr}) = 0 \tag{4.5}$$

(with $\mathrm{R}_+ = \sum_{j=1}^{m} \mathrm{R}_j$) that can be solved numerically.

The estimation of the dispersion parameter a and its potential biases, such as unstable variance have been discussed in numerous papers. As a consequence, this parameter is sometimes assumed to be fixed [Cameron and Trivedi, 2013]. In our case, we observed that the EM algorithm may have an erratic behavior when the parameter a is updated at each M step. In such case, we choose to optimize a over a preset grid of values, for which we run a series of VEM.

4.3.3 Classification

In model-based clustering, classification most often relies on the conditional distribution of the hidden labels given the observed data $\mathbf{Y}$. However, the conditional distributions $p(\mathbf{Z}|\mathbf{Y})$ and $p(\mathbf{W}|\mathbf{Y})$ are not tractable in LBM. Still, s_{ik} and t_{jg} are the optimal approxima-

tions (in terms of Küllback-Leibler divergence) of $p(\mathbf{Z}_i = k|\mathbf{Y})$ and $p(\mathbf{W}_j = g|\mathbf{Y})$, respectively. We classify the observation according to the maximum a posteriori rule applied to these approximate probabilities and we define

$$\widehat{\mathbf{Z}}_i = \arg\max_k s_{ik}, \qquad \widehat{\mathbf{W}}_j = \arg\max_g t_{jg}.$$

4.4 Model selection

4.4.1 Standard BIC and ICL criteria

In practice the number of co-clusters is generally unknown and should be chosen according to an adequate selection model procedure. BIC (Bayesian Information Criterion, Schwarz [1978]) is the most used criterion in the context of mixture model and is defined for model $\mathscr{M}$ as $\text{BIC}(\mathscr{M}) = \log p(\mathbf{Y}; \widehat{\theta}_{\mathscr{M}}) - \text{pen}(\mathscr{M})$ where $\text{pen}(\mathscr{M}) = d_{\mathscr{M}} \log(\text{N})/2$, $d_{\mathscr{M}}$ is the number of independent parameters and N the number of data. In the context of classification, Biernacki et al. [2000] proposed to add a penalty term that accounts for the uncertainty of the hidden variables $\mathbf{H}$. They define the criterion $\text{ICL}(\mathscr{M}) = \text{BIC}(\mathscr{M}) - \mathscr{H}[p(\mathbf{H}|\mathbf{Y})]$. Interestingly, this amounts simply to penalize the maximized objective function of the EM algorithm : $\text{ICL}(\mathscr{M}) = \mathbb{E}[\log p(\mathbf{Y}, \mathbf{H}; \widehat{\theta}_{\mathscr{M}})|\mathbf{Y}] - \text{pen}(\mathscr{M})$.
Still, these standard criteria do not apply to LBM because (i) the log-likelihood $\log p(\mathbf{Y}; \theta)$, is intractable (ii) the objective function of VEM is not the same as this of EM and (iii) the form of the penalty needs to be adapted to the bi-clustering setting.

4.4.2 Variational BIC and ICL for the Poisson-Gamma LBM

For the three issues mentioned above, we follow the line of Keribin et al. [2014] for LBM and Daudin et al. [2008] for the stochastic block-model (SBM, which is a special case of LBM). Regarding point (iii), we define the following penalty for an LBM $\mathscr{M}_{\text{K,G}}$ with K and G groups :

$$\text{pen}(\mathscr{M}_{\text{K,G}}) = (\text{K}-1)\log(n)/2 + (\text{G}-1)\log(m)/2 + (\text{KG}-1)\log(nm)/2.$$

Note that the penalty terms related to the parameters μ_i, ν_{jr} and a do not appear here as they do not depend on K an G.
Both points (i) and (ii) are due to the intractability of the conditional distribution $p(\mathbf{Z}, \mathbf{W}, \mathbf{U}|\mathbf{Y})$. Again, following Keribin et al. [2014] and Daudin et al. [2008], we replace all quantities with their respective variational approximations, that is $\log p(\mathbf{Y}, \widehat{\theta}) \simeq \mathscr{J}(\mathbf{Y}, \widehat{q}, \widehat{\theta})$ (as defined in (4.2), $\mathscr{H}[p(\mathbf{Z}|\mathbf{Y})] \simeq \mathscr{H}[\widehat{q}_1(\mathbf{Z})]$, etc., denoting $\widehat{q}$ the optimal approximate conditional distribution. This results in the variational-BIC criterion :

$$v\text{BIC}(\mathscr{M}_{\text{K,G}}) = \mathscr{J}(\mathbf{Y}, \widehat{q}_{\text{K,G}}, \widehat{\theta}_{\text{K,G}}) - \text{pen}(\text{K,G}).$$

Three different sets of hidden variables (**Z**, **W** and **H**) are involved in the Poisson-Gamma LBM. In a classification context, it may seem desirable to penalize only for the entropy of the classification variables **Z** and **W**. Because the approximate conditional joint entropy can be decomposed as $\mathscr{H}[\widehat{q}(\mathbf{Z},\mathbf{W},\mathbf{U})] = \mathscr{H}[\widehat{q}_3(\mathbf{Z})] + \mathscr{H}[\widehat{q}_2(\mathbf{W})] + \mathscr{H}[\widehat{q}_3(\mathbf{U})]$, we get a first variational ICL criterion :

$$\begin{aligned} v\mathrm{ICL}_1(\mathscr{M}_{\mathrm{K,G}}) &= \mathscr{J}(\mathbf{Y}, \widehat{q}_{\mathrm{K,G}}, \widehat{\theta}_{\mathrm{K,G}}) - \mathscr{H}[\widehat{q}_1(\mathbf{Z})] - \mathscr{H}[\widehat{q}_2(\mathbf{W})] - \mathrm{pen}(\mathrm{K},\mathrm{G}) \\ &= \mathbb{E}_{\widehat{q}}[\log p(\mathbf{Y},\mathbf{Z},\mathbf{W},\mathbf{U};\widehat{\theta}_{\mathrm{K,G}})] + \mathscr{H}[\widehat{q}_3(\mathbf{U})] - \mathrm{pen}(\mathrm{K},\mathrm{G}). \end{aligned}$$

In practice, because the entropies $\mathscr{H}[\widehat{q}_1(\mathbf{Z})]$ and $\mathscr{H}[\widehat{q}_2(\mathbf{W})]$ turn out to be very small (all s_{ik} and t_{jg} are very close to 0 or 1), we observed very little difference between $v\mathrm{BIC}(\mathscr{M}_{\mathrm{K,G}})$ and $v\mathrm{ICL}_1(\mathscr{M}_{\mathrm{K,G}})$.
In the next section, we will present an example with very large over-dispersion, suggesting a large value for $\mathscr{H}[\widehat{q}_3(\mathbf{U})]$. This situation suggest to penalize for the entropy of all hidden variables, leading to second variational ICL :

$$\begin{aligned} v\mathrm{ICL}_2(\mathscr{M}_{\mathrm{K,G}}) &= \mathscr{J}(\mathbf{Y}, \widehat{q}_{\mathrm{K,G}}, \widehat{\theta}_{\mathrm{K,G}}) - \mathscr{H}[\widehat{q}(\mathbf{Z},\mathbf{W},\mathbf{U})] - \mathrm{pen}(\mathrm{K},\mathrm{G}) \\ &= \mathbb{E}_{\widehat{q}}[\log p(\mathbf{Y},\mathbf{Z},\mathbf{W},\mathbf{U};\widehat{\theta}_{\mathrm{K,G}})] - \mathrm{pen}(\mathrm{K},\mathrm{G}). \end{aligned}$$

4.5 Applications

This section presents applications on three real 16S or 18S rRNA amplicon-based datasets. Our purpose is to understand the structure of the relationships between plants and bacteria communities living in their rhizosphere (MetaRhizo dataset), between bacteria communities living in various environments and the environmental samples themselves (GlobalPatterns dataset), and finally between bacteria and fungi living in the phyllosphere and leaves themselves (*Erysiphe alphitoides* pathobiome dataset) respectively.

4.5.1 MetaRhizo dataset

The MetaRhizo dataset includes 483 biological samples corresponding to different genotypes of the plant and 288 bacteria at genus level (C. Mougel, personnal communications). The total counts per sample go from 29410 to 33840 number of sequences. We perform the LBM defined in section 4.2 with an estimation of the parameter a.

Figure 4.2 illustrates the model selection criterion defined in (4.4.2) according to the penalty term. The selected model corresponds to K = 16 groups of bacteria and G = 12 groups of plants. Even if a row effect, corresponding to a specific bacterium effect, is included in the model, the groups in rows seem to be explained by the mean abundance level of bacteria (see Figure 4.3 top right). In order to understand this, we calculate for each biological sample the Shannon diversity Index, a common measure of diversity using the

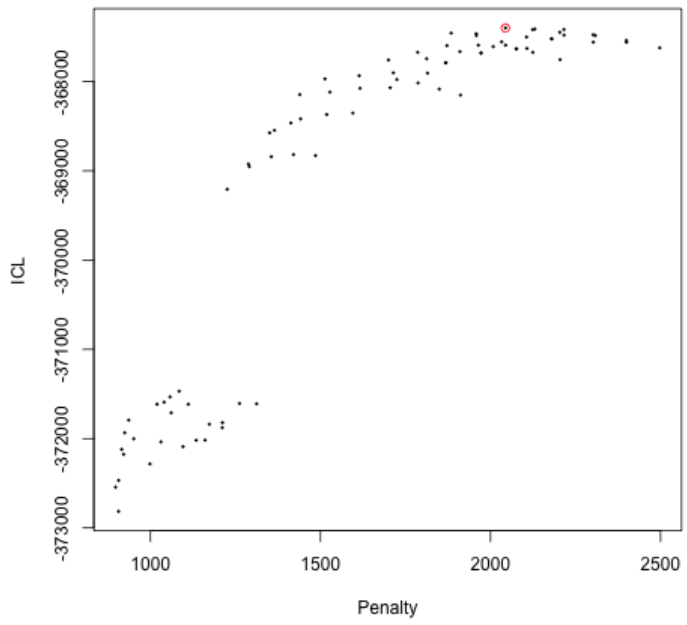


FIGURE 4.2 – MetaRhizo dataset. Plot of the ICL criterion (y-axis) according to the penalty term (x-axis)

proportional abundances of each species. We plot this index according to the group of environmental samples (see Figure 4.3 bottom left). Biological samples seem to be gathered according to their level of diversity. Figure 4.3 bottom right represents the matrix of interaction terms ordered by the absolute value of the Shannon diversity index of the column groups (Figure 4.3 bottom left) and the mean abundance level of bacteria in the row groups (Figure 4.3 top right). The interpretation of such matrix is the following : all bacteria in the same group in rows are associated in the same way with the groups in columns. The sign of $\log\alpha_{kg}$ indicates wether the association between row group k and column group g is positive or negative. The figure 4.3 bottom right suggests that the most abundant bacteria are overrepresented in the plants with a low diversity index and that rare species are more abundant in samples with a higher Shannon index.

4.5.2 GlobalPatterns dataset

The GlobalPatterns dataset published in Caporaso et al. [2011] and available in the phyloseq R package [McMurdie and Holmes, 2013] includes 26 biological samples. We collapsed all taxa OTUs (Operational Taxonomic Units) to the genus level. All genera accounting for less than 0.5% of the sequences for all the samples were filtered out. 215 genera were then analyzed. We performed our LBM with a fixed dispersion parameter ($a = 5$). We selected the (K = 22, G = 16) model according to our selection model criterion

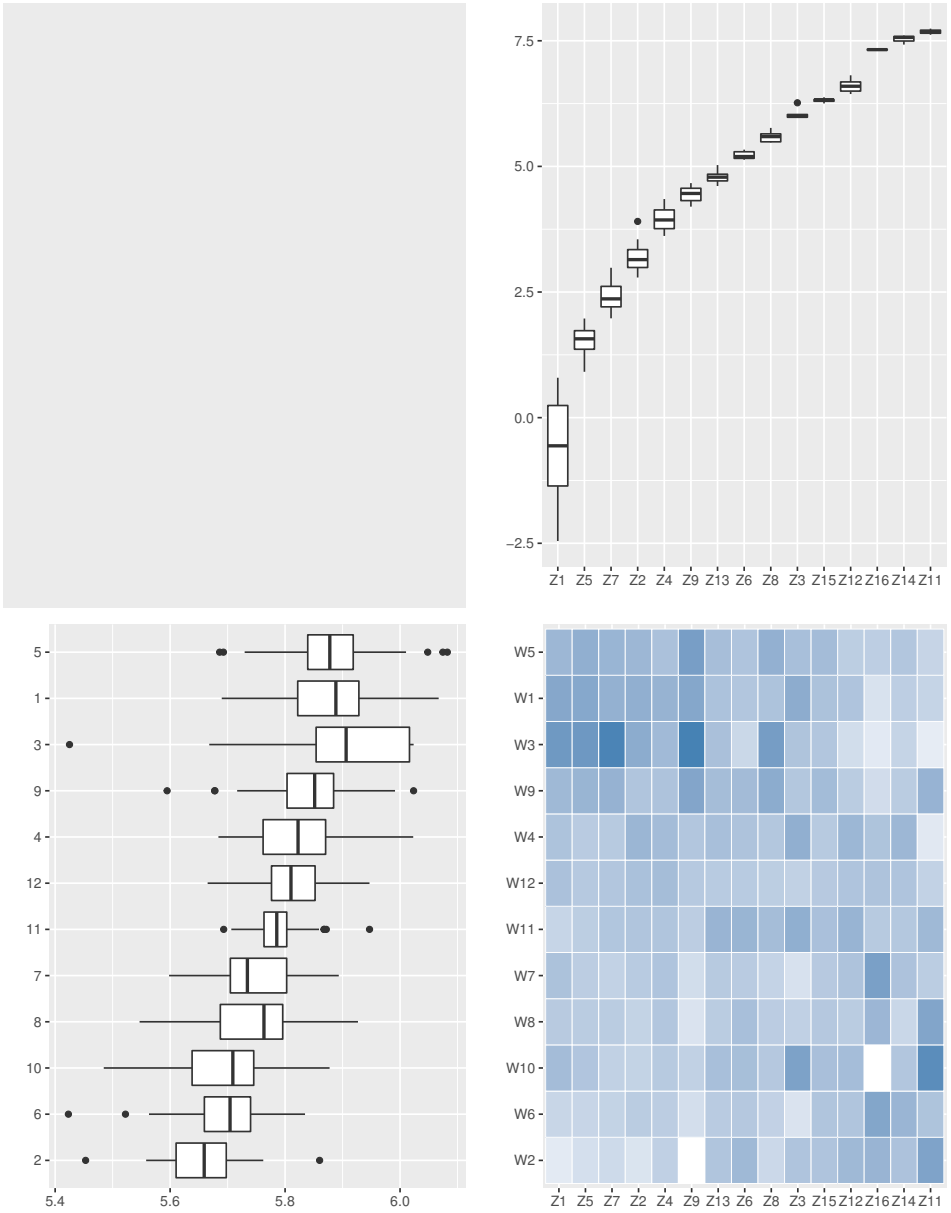


FIGURE 4.3 – MetaRhizo dataset. Top right : boxplot of the μ_i within each group in row Z_k. Bottom left : plot of the absolute value of the Shannon diversity index in x-axis in function of the group of environmental samples W_g (y-axis). Bottom right : Heatmap of the $\log\alpha_{kg}$ interaction terms.

(see Figure 4.4 top-right). We can see on Table 4.1 that biological samples belonging to a specific type (Ocean, Soil, Human body, ⋯) tend to be gathered in same groups and have similar interactions terms (see Figure 4.4 bottom-right). We observe that one of the two tongue sample is clustered with two skin ones. These results are consistent with previous studies [Caporaso et al., 2011; Grice and Segre, 2011; Ren et al., 2016]. The LBM enables to highlight the groups of OTUs structuring the clustering. We see for example on Figure 4.4 that several groups of OTUs are nearly totally absent in mock and feces samples (W4, W5, W15 and W16).

Sample ID	Sample Type	Description	W
AQC1cm	Freshwater (creek)	Allequash Creek, 0-1cm depth	11
AQC4cm	Freshwater (creek)	Allequash Creek, 3-4 cm depth	11
AQC7cm	Freshwater (creek)	Allequash Creek, 6-7 cm depth	11
CC1	Soil	Cedar Creek Minnesota, grassland, pH 6.1	2
CL3	Soil	Calhoun South Carolina Pine soil, pH 4.9	1
Even1	Mock	Even1	16
Even2	Mock	Even2	16
Even3	Mock	Even3	16
F21Plmr	Skin	F1, Day 1, right palm, whole body study	7
LMEpi24M	Freshwater	Lake Mendota Minnesota, 24 meter epilimnion	9
M11Fcsw	Feces	M1, Day 1, fecal swab, whole body study	5
M11Plmr	Skin	M1, Day 1, right palm, whole body study	7
M11Tong	Tongue	M1, Day 1, tongue, whole body study	7
M31Fcsw	Feces	M3, Day 1, fecal swab, whole body study	4
M31Plmr	Skin	M3, Day 1, right palm, whole body study	6
M31Tong	Tongue	M3, Day 1, tongue, whole body study	8
NP2	Ocean	Newport Pier, CA surface water, Time 1	12
NP3	Ocean	Newport Pier, CA surface water, Time 2	13
NP5	Ocean	Newport Pier, CA surface water, Time 3	12
SLEpi20M	Freshwater	Sparkling Lake Wisconsin, 20 meter eplimnion	10
SV1	Soil	Sevilleta new Mexico, desert scrub, pH 8.3	3
TRRsed1	Sediment (estuary)	Tijuana River Reserve, depth 1	14
TRRsed2	Sediment (estuary)	Tijuana River Reserve, depth 2	14
TRRsed3	Sediment (estuary)	Tijuana River Reserve, depth 2	14
TS28	Feces	Twin #1	15
TS29	Feces	Twin #2	15

TABLE 4.1 – Description of environmental samples

4.5.3 *Erysiphe alphitoides* pathobiome dataset

This dataset presented in Jakuschkin et al. [2016] was produced in order to study the pathobiome of the fungus *Erysiphe alphitoides*, the causal agent of oak powdery mildew, one of the most common diseases in European forests. The concept of pathobiome introduced by Vayssier-Taussat et al. [2014] considers the pathogenic agent inside its biotic environment. This dataset includes 116 leaves from three different trees characterized

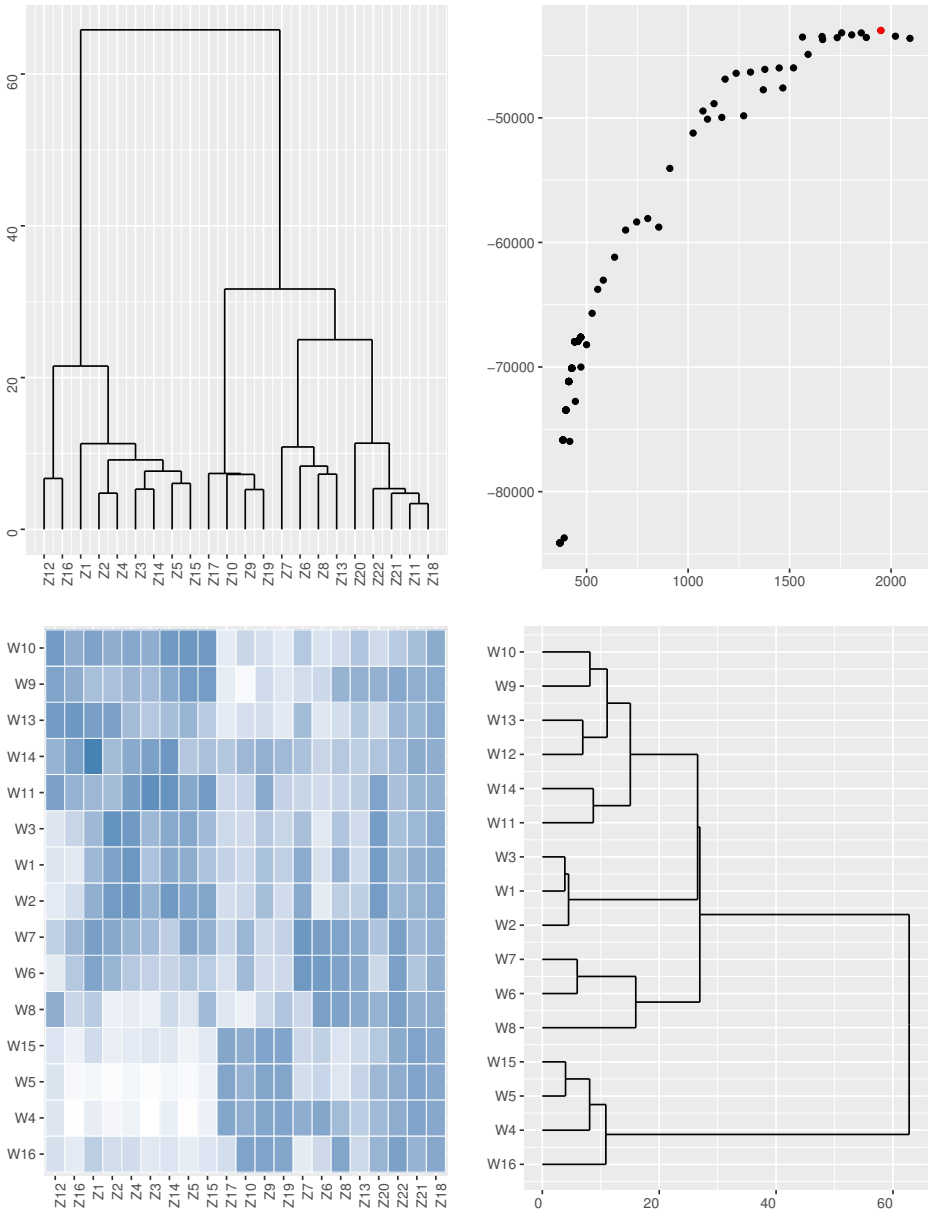


FIGURE 4.4 – GlobalPatterns Data. Top left : Dendrogram of a hierarchical clustering of the groups of OTUs. Top right : Plot of the ICL criterion (y-axis) according the penalty term (x-axis). Bottom left : Heatmap of the $\log\alpha_{kg}$ interaction terms. Bottom right : a hierarchical clustering of the groups of biological samples. Hierarchical clusterings are constructed from the $(\log\alpha_{\mathbf{kg}})_{kg}$ matrix using the euclidian distance and the Ward criterion.

as highly susceptible, intermediately resistant and strongly resistant to powdery mildew, and 114 operational taxonomic units (OTU), among them 48 fungal OTUs including *E. alphitoides* and 66 bacterial OTUs. As the fungal and bacterial OTUs are produced by two different sequencing experiments, we proposed to used two different ν_j per biological sample, one for fungal OTUs and one for bacterial OTUs. We first performed our LBM described in section 4.2 with different fixed values of a ($a \in \{0.1; 0.4; 1; 10\}$). Surprisingly, we selected the model with K = 1 and G = 1 and $a = 0.1$ that means the model with the maximal dispersion and no different group both in rows and columns. Actually, we noticed that using a unique a for all co-clusters is certainly too strong as shown on figure 4.7 : for high values of K and G, we indeed computed a posteriori for each (k, g) the gamma scale parameter and observed very different values. Since our model can be easily extended to a model with a gamma scale parameter specific to each co-cluster a_{kg} , we then performed a LBM with a specific a_{kg} parameter for co-cluster, for different values of (K, G). We select according to the ICL criterion the model K = 18 and G = 2. The two groups defined in columns differ on the basis of their level of infection (Figure 4.5). As we have two groups of samples, we can calculate the log-ratio of the interaction terms $\log \frac{\alpha_{k1}}{\alpha_{k2}}$ for $k = 1, \cdots, 18$. The group with the highest positive log-ratio value (Figure 4.6) includes the pathogenic fungus along with 5 bacteria.

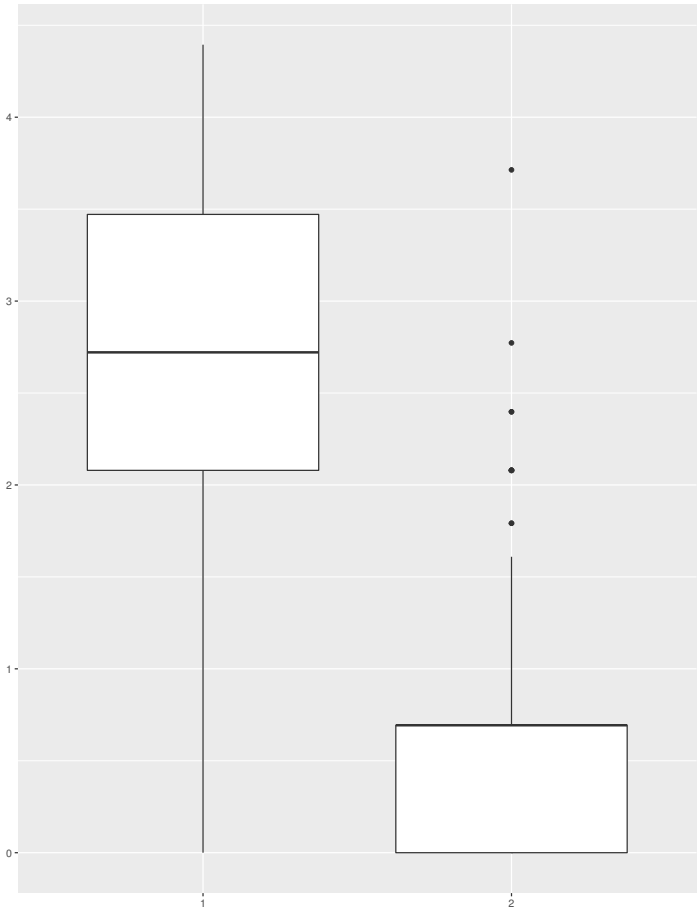


FIGURE 4.5 – *Erysiphe alphitoides* pathobiome dataset. Boxplot of level of infection in log-scale for the two groups of biological samples.

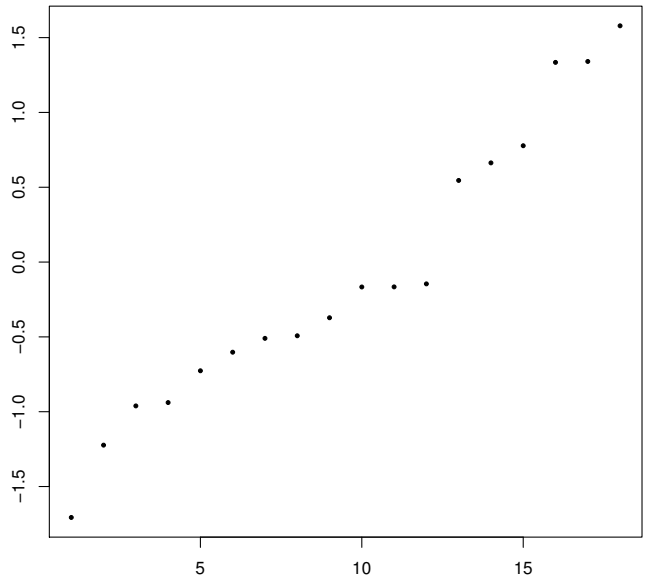


FIGURE 4.6 – *Erysiphe alphitoides* pathobiome dataset. Plot of the ordered $\log(\frac{\alpha_{k1}}{\alpha_{k2}})$ for $k = 1, \ldots, 18$.

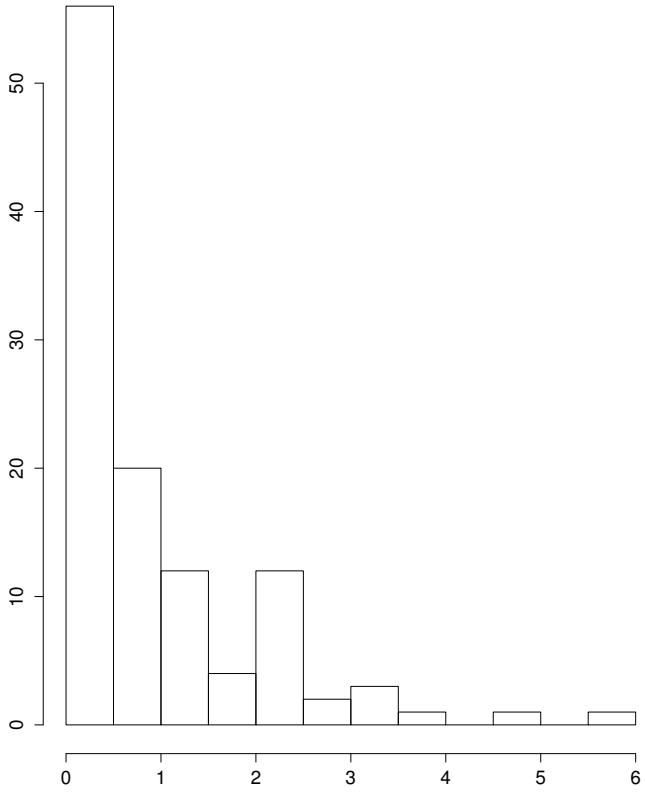


FIGURE 4.7 – *Erysiphe alphitoides* pathobiome dataset. Histogram of the a_{kg} values.

4.6 Discussion

We proposed a model-based biclustering approach combined with a selection criterion adapted to count data. Applications on several datasets were discussed. These analyses suggested that the proposed methodology is an attractive one. On different datasets, we show the flexibility of the model and the ability to help in data interpretation. The ν_j term enables us to take into account a specific sample effect, such as a difference in sampling effort. Any scaling factors available in the litterature for normalization purpose [Dillies et al., 2012] may be plugged-in. In the *Erysiphe alphitoides* example, we use a ν_j specific for each j and for each nature of microorganism (fungi or bacteria). This specific choice enables the combination of different datasets.
As shown before, the estimation of the gamma parameter a is not trivial. We proposed two strategies : (i) a common value which can be either estimated neither selected among a grid in case of erratic behaviour with the proposed estimator of a, or (ii) an estimated parameter specific to each co-cluster (see 4.5.3).

Another possibility as done for differential gene expression analysis [Anders and Huber, 2010; Zhou et al., 2014] may be to estimate a trended parameter depending on the microorganism abundance for example.
We proposed different criteria for model selection. All the criteria often lead to the same selected model. Interestingly, the selection differs in case of very large over-dispersion (*Erysiphe alphitoides* example), suggesting a large value for $\mathcal{H}[\widehat{q}_3(\mathbf{U})]$. In this situation, the use of the second variational ICL $v\text{ICL}_2(\mathcal{M}_{\text{K,G}})$ which penalizes for the entropy of all hidden variables is more adapted. While this paper focuses on 16S or 18S rRNA amplicon-based data, the range of possible applications is broader including shotgun metagenomics and all other over-dispersed count data.

4.7 Appendix

4.7.1 Optimization of $q(\mathrm{U})$

The optimization of the lower bound wrt each $q(\mathrm{U}_{ijr})$ produces a gamma distribution with the following parameters :

$$\begin{aligned}\widetilde{a}_{ijr} &= a + \sum_{k,g} s_{ik} t_{jg} \mu_i \nu_{jr} \alpha_{k,g} \\ \widetilde{b}_{ijr} &= a + \sum_{k,g} s_{ik} t_{jg} y_{ijr}\end{aligned}$$

Démonstration The optimal distribution $q(\mathbf{U})$ is given by :

$$\log q_3^*(\mathbf{U}) = \mathbb{E}_q[\log \mathrm{P}(\mathrm{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{U})] + cst. \tag{4.6}$$

We now decompose the joint distribution. As we are only interested in the functional dependence on the right-hand side on the variable **U**, any term that does not depend on **U** can be absorbed into the constant, leading to

$$\begin{aligned}\log q_3^*(\mathbf{U}) &= \mathbb{E}_q[\log \mathrm{P}(\mathrm{Y}|\mathbf{Z},\mathbf{W},\mathbf{U}) + \log \mathrm{P}(\mathbf{Z}) + \log \mathrm{P}(\mathbf{W}) + \log \mathrm{P}(\mathbf{U})] + cst \\ &= \mathbb{E}_q[\log \mathrm{P}(\mathrm{Y}|\mathbf{Z},\mathbf{W},\mathbf{U}) + \log \mathrm{P}(\mathbf{U})] + cst.\end{aligned}$$

According to the model,

$$\begin{aligned}\log \mathrm{P}(\mathbf{U}) &= \sum_{ijr}(a\log a - \log\Gamma(a) + (a-1)\log u_{ijr} - a u_{ijr}) \\ \log \mathrm{P}(\mathrm{Y}|\mathbf{Z},\mathbf{W},\mathbf{U}) &= \sum_{ijrkg} z_{ik} w_{jg}(-\mu_i \nu_{jr} u_{ijr} + y_{ijr}\log(\mu_i \nu_{jr} \alpha_{\mathrm{Z}_{ik},\mathrm{W}_{jg}} u_{ijr}) - \log y_{ijr}!)\end{aligned}$$

This leads to :

$$\log q_3^*(\mathbf{U}) = \sum_{ijr}(a - 1 + s_{ik} t_{jl} y_{ijr})\mathbb{E}_q[\log \mathrm{U}_{ijr}] + (-a - s_{ik} t_{jl} \mu_i \nu_{jr} \alpha_{k,g})\mathbb{E}_q[\mathrm{U}_{ijr}] + cst$$

We recognize a gamma distribution with the following parameters :

$$\begin{aligned}\widetilde{a}_{ijr} &= a + \sum_{k,g} s_{ik} t_{jg} y_{ijr} \\ \widetilde{b}_{ijr} &= a + \sum_{k,g} s_{ik} t_{jg} \mu_i \nu_{jr} \alpha_{k,g}\end{aligned}$$

4.8 Commentaires et perspectives

4.8.1 Implémentation algorithmique

La méthode proposée dans cette section fera l'objet d'un package R cobiclust qui sera diffusé à la communauté afin que la méthode que nous avons proposée puisse être utilisée.

Pour l'implémentation des algorithmes proposés, nous avons adapté la stratégie proposée par Govaert and Nadif [2008] dans le cas des modèles à blocs latents plus classiques (c'est-à-dire comportant une variable latente en ligne, et une en colonne).

1. Partant de $\mathbf{s}^{(0)}$, $\mathbf{t}^{(0)}$, $\mathbb{E}_q(\mathrm{U}_{ijr})^{(0)}$, $\mathbb{E}_q(log\mathrm{U}_{ijr})^{(0)}$, $\theta^{(0)}$.
2. Calcul de $\mathbf{s}^{(h+1)}$, $\pi^{(h+1)}$, $\alpha^{(h+\frac{1}{3})}$, $\mathbb{E}_q(\mathrm{U}_{ijr})^{(h+\frac{1}{3})}$, $\mathbb{E}_q(log\mathrm{U}_{ijr})^{(h+\frac{1}{3})}$ en utilisant un algorithme VEM partant de $\mathbf{s}^{(h)}$,$\pi^{(h)}$,$\alpha^{(h)}$,$\mathbb{E}_q(\mathrm{U}_{ijr})^{(h)}$,$\mathbb{E}_q(log\mathrm{U}_{ijr})^{(h)}$.
3. Calcul de $\mathbf{t}^{(h+1)}$, $\rho^{(h+1)}$, $\alpha^{(h+\frac{2}{3})}$, $\mathbb{E}_q(\mathrm{U}_{ijr})^{(h+\frac{2}{3})}$, $\mathbb{E}_q(log\mathrm{U}_{ijr})^{(h+\frac{2}{3})}$ en utilisant un algorithme VEM partant de $\mathbf{t}^{(h)}$,$\rho^{(h)}$,$\alpha^{(h+\frac{1}{3})}$,$\mathbb{E}_q(\mathrm{U}_{ijr})^{(h+\frac{1}{3})}$,$\mathbb{E}_q(log\mathrm{U}_{ijr})^{(h+\frac{1}{3})}$.
4. Calcul de $a^{(h+1)}$, $\mu^{(h+1)}$, $\alpha^{(h+1)}$,$\mathbb{E}_q(\mathrm{U}_{ijr})^{(h+1)}$, $\mathbb{E}_q(log\mathrm{U}_{ijr})^{(h+1)}$.

5. Itération des étapes (2), (3) et (4) jusqu'à convergence.

Comme proposé par Govaert and Nadif [2008], dans le but d'accélerer la convergence de l'algorithme, nous ne faisons qu'une itération aux niveaux locaux ((2),(3)) et un plus grand nombre d'itérations au niveau global ((4),(5)). Nous déclarons que la convergence globale est atteinte quand

$$\frac{|(\mathscr{J}(\mathrm{Y},\theta^{h+1},q)-\mathscr{J}(\mathrm{Y},\theta^{h},q)|}{\mathscr{J}(\mathrm{Y},\theta^{h},q)}\leq\epsilon.$$

4.8.2 Choix des termes d'interaction intéressants

Les paramètres parmi les plus intéressants du modèle sont les termes α_{kg}. Nous ne disposons pas de tests sur la significativité de ces termes. Dans la pratique, nous nous contentons d'interpréter ceux ayant les valeurs les plus grandes et observons que certaines estimations sont quasi-nulles. Une extension intéressante du modèle, sera de proposer une version régularisée permettant de forcer certains termes d'interactions à être nuls.

4.9 Références

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), 2010. doi : 10.1186/gb-2010-11-10-r106. 66, 80

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :719–725, 2000. 70

C. Bouveyron and C. Brunet. Model-based clustering of high-dimensional data : A review. *Comput. Stat. Data Anal.*, 71 :52–78, 2014. 64

V. Brault and M. Mariadassou. Co-clustering through latent bloc model : a review. *Journal de la Société Française de la Statistique*, 156(3), 2015. 64, 66

A. Cameron and P. Trivedi. *Regression Analysis of Count Data, 2nd edition.* Cambridge University Press, 2013. 69

J. Caporaso, C. Lauber, W. Walters, D. Berg-Lyons, C. Lozupone, P. Turnbaugh, N. Fierer, and R. Knight. Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 2011. doi : 10.1073/pnas.1000080107. 65, 72, 74

E. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32, 2016. 63

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18 :151–171, 2008. 70

P. de Castro, F. de França, H. Ferreira, and F. Von Zuben. Applying biclustering to text mining : an immune-inspired approach. *Artificial Immune Systems*, pages 83–94, 2007. 64

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 :1–38, 1977. 67

M.-A. Dillies, A. Rau, **Aubert, Julie**, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, and F. Jaffrezic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 2012. 80

G. Govaert and M. Nadif. Block clustering with bernoulli mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis*, 52 :3233–3245, 2008. 67, 81, 82

G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39 :416–425, 2010. 65

G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2) :463–473, 2012. 64, 65

E. Grice and J. Segre. The skin microbiome. *Nature Reviews Microbiology*, 9(4) :244–253, 2011. 74

T. Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1) :89–115, Jan. 2004. ISSN 1046-8188. doi : 10.1145/963770.963774. URL http://doi.acm.org/10.1145/963770.963774. 64

T. Jaakkola. *Advanced mean field methods : theory and practice*, chapter Tutorial on variational approximation methods. MIT Press, 2000. 68

B. Jakuschkin, V. Fievet, L. Schwaller, C. Robin, and C. Vacher. Deciphering the pathobiome : intra- and interkingdom interactions involving the pathogen erysiphe alphitoides. *Microbial Ecology*, 2016. 65, 74

V. Jonsson, T. Österlund, O. Nerman, and E. Kristiansson. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics*, 17, 2016. doi : 10.1186/s12864-016-2386-y. 63, 64

M. Jordan, Z. Ghahramani, T. Jaakkola, and K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37 :183–233, 1999. 67

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25 :1201–1216, 2014. 66, 70

C. Law, Y. Chen, W. Shi, and G. Smyth. voom : precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15, 2014. doi : 10.1186/gb-2014-15-2-r29. 62

M. Lindner and B. Renard. Metagenomic profiling of known and unknown microbes with microbegps. *Plos One*, 10, 2015. doi : doi:10.1371/journal.pone.0117711. 64, 66

M. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15 :550, 2014. doi : 10.1186/s13059-014-0550-8. 62

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis : a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1) : 24–45, Jan 2004. ISSN 1545-5963. doi : 10.1109/TCBB.2004.2. 64

P. McMurdie and S. Holmes. phyloseq : An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8(4) :e61217, 2013. URL http://dx.plos.org/10.1371/journal.pone.0061217. 72

P. McMurdie and S. Holmes. Waste not, want not : Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4) :e1003531, 2014. doi : 10.1371/journal.pcbi.1003531. 63, 64

V. Melnykov. Model-based biclustering of clickstream data. *Comput. Stat. Data Anal.*, 93 : 31–45, 2016. 64

S. Nayfach and K. S. Pollard. Toward accurate and quantitative comparative metagenomics. *Cell*, 166 :1103–1116, 2016. doi : 10.1016/j.cell.2016.08.007. 62

A. Oghabian, S. Kilpinen, S. Hautaniemi, and E. Czeizler. Biclustering methods : Biological relevance and application in gene expression analysis. *Plos One*, 2014. 64

J. Paulson, O. Stine, H. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10 :1200–2, 2013. 63

B. Ren, S. Bacallado, S. Favaro, S. Holmes, and L. Trippa. Bayesian Nonparametric Ordination for the Analysis of Microbial Communities. *ArXiv e-prints*, Jan. 2016. 74

M. Robinson, D. McCarthy, and G. Smyth. edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2009. 62, 66

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 03 1978. doi : 10.1214/aos/1176344136. URL http://dx.doi.org/10.1214/aos/1176344136. 70

N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9, 2012. 62

S. SMarek Gierlinski, C. Cole, P. Schofield, N. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, M. Blaxter, and G. Barton. Statistical models for rna-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 31, 2015. doi : 10.1093/bioinformatics/btv425. 62

M. Van de Wiel, M. Neerincx, T. Buffart, D. Sie, and H. Verheul. Shrinkbayes : a versatile r-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinformatics*, 15, 2014. 63

M. Vayssier-Taussat, E. Albina, C. Citti, J. Cosson, M. Jacques, M. Lebrun, Y. Le Loir, M. Ogliastro, M. Petit, P. Roumagnac, and T. Candresse. Shifting the paradigm from pathogens to pathobiome : new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology*, 4(29), 2014. doi : 10.1073/pnas.1000080107. 74

D. Warton, S. Foster, G. Death, J. Stoklosa, and P. Dunstan. Model-based thinking for community ecology. *Journal Plant Ecology*, 216 :669–682, 2015. 64

J. R. White, N. Nagarajan, and M. Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Computational Biology*, 5, 2009. 64, 66

X. Zhang, H. Mallick, and N. Yi. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *Journal of Bioinformatics and Genomics*, 2016. 63

X. Zhou, H. Lindsay, and M. Robinson. Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic Acids Research*, 42(11) :e91, 2014. 80

Chapitre 5

Analyse statistique de données transcriptomiques

Sommaire

Dans le cadre d'études de données transcriptomiques, une des questions principales est la recherche de régions d'intérêt (gènes, séquences etc.) qui s'expriment de façon différente entre plusieurs conditions expérimentales. Pour répondre à une telle question, la démarche représentée dans la figure 5.1 est la même quelle que soit la technologie utilisée pour obtenir des données. Elle consiste tout d'abord à définir un plan expérimental préalablement à l'étape de production des données. Une fois les données produites, elles sont pré-traitées. Ce pré-traitement consiste en une étape de contrôle qualité, une étape d'analyse d'image dans le cas des puces ou d'analyse bioinformatique dans le cas des données de séquençage. A l'issue de ces analyses dites primaires, nous obtenons ce que nous appelons des données brutes. Ces données peuvent nécessiter une étape de normalisation indépendante ou non du modèle statistique qui servira pour l'analyse différentielle. Pour chaque région d'intérêt, on cherche alors à tester si le niveau d'expression est le même dans les différentes conditions expérimentales. On effectue autant de tests que de régions d'intérêt. Ces analyses nécessitent donc de prendre en compte la multiplicité des tests.

Les travaux présentés dans ce chapitre sont dédiés à l'analyse de données de transcriptomique issues des technologies de puce à ADN et de séquençage de l'ARN. Le principe de ces deux technologies ainsi que les notions biologiques nécessaires à la compréhension de ce chapitre sont présentées dans la section introductive biologique 1.4. Ce chapitre présente à la fois de nouvelles méthodes et des comparaisons de méthodes auxquelles j'ai contribuées. La première section est dédiée à la normalisation des données (détection et correction de biais techniques) et présente deux nouvelles méthodes que j'ai proposées avec mes co-auteurs [Martin-Magniette et al., 2005a,b, 2008] et une comparaison de méthodes à laquelle j'ai contribuée [Dillies et al., 2012]. La deuxième section dédiée à la planification expérimentale présente une méthode pour analyser les dispositifs dits en dye-switch pour lesquels il y a autant de puces à deux couleurs comparant un échantillon correspondant à la condition biologique A marqué en Vert à un échantillon de la condition B marqué en Rouge, que de puces comparant un échantillon de la condition biologique A marqué en Rouge à un échantillon de la condition biologique B marqué en Vert (Mary-Huard et al. [2008]).

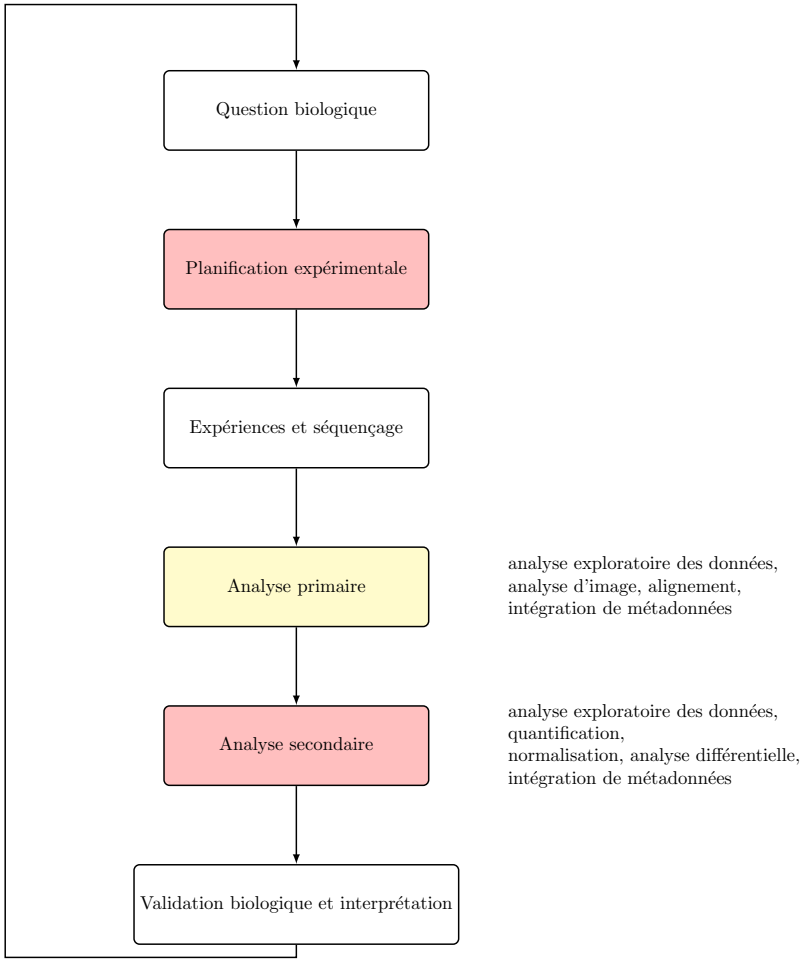


FIGURE 5.1 – Schéma représentant une expérience typique de transcriptomique à partir de données de séquençage de l'ARN

5.1 Normalisation des données

Les données issues des technologies de biologie moléculaire à haut débit sont entachées de bruit technique. Les biais techniques sont inhérents à chaque technologie. Ils sont présents à chaque étape du protocole allant de la production du matériel biologique à l'obtention de la donnée en sortie d'expériences. Ils peuvent parfois être contrôlés par un plan d'expérience adapté ou un bon protocole expérimental mais certains sont systématiques et non contrôlables. Ces derniers nécessitent d'être identifiés puis corrigés par une technique de normalisation adaptée.

L'étape de normalisation est indispensable et a un impact non négligeable sur la suite des analyses. Cette étape nécessite un contact direct avec les données et les producteurs de données et des aller-retours assez fréquents : observation d'un phénomène inattendu (données de mauvaise qualité), recherche de l'origine du biais et modification du protocole expérimental ou mise en place d'une procédure de normalisation. Une fine compréhension du processus de production de données est nécessaire pour comprendre les biais. Les statisticiens ne sont pas les personnes a priori les plus compétentes pour connaître et identifier les biais mais ce sont souvent elles qui se rendent compte que quelque chose sort de l'attendu. Et elles sont alors les plus à même de proposer une modélisation du signal permettant une identification et une correction systématique de ces biais.

5.1.1 Biais de marquage des puces à ADN à deux couleurs

Kerr et al. [2002] ont proposé de modéliser le signal issu des puces à ADN à l'aide d'un modèle d'analyse de variance afin de quantifier les biais les plus importants en comparant les écarts à la moyenne de chaque effet inclu dans le modèle. Soit Y_{ijkg} le logarithme en base 2 de la mesure sur la puce i, avec le fluorochrome j, de l'échantillon d'ARN dans la condition k pour le gène g, le signal peut être modélisé de la façon suivante :

$$Y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + (DG)_{jg} + E_{ijkg} \tag{5.1}$$

avec A_i l'effet de la puce (Array) i, D_j l'effet du fluorochrome (Dye) j, V_k l'effet de la condition (Variety) k, G_g l'effet du gène g. $(AG)_{ig}$,$(VG)_{kg}$,$(DG)_{jg}$ correspondent aux termes d'interactions d'ordre 2. Les erreurs E_{ijkg} sont supposées indépendantes, aléatoires et de moyenne nulle.

Kerr et al. [2002] ont montré que le biais le plus important était le biais de marquage, dû à une efficacité d'incorporation différente des fluorochromes Cy5 (Cyanine 5 marquant en rouge) et Cy3 (Cyanine 3 marquant en vert) utilisés et à une fluorescence naturelle en vert des sondes. La méthode la plus courante pour corriger ce biais a été proposée par Yang et al. [2002] et est présentée la section 2.5. Elle s'effectue puce par puce et consiste pour chaque puce à modéliser les différences d'expression entre l'échantillon marqué en rouge ($j = 1$) et celui marqué en vert (($j = 2$) par une fonction des intensités

moyennes, d'estimer cette fonction par une procédure *lowess* et de retirer l'estimation à la valeur des différences.
Si on suppose que l'échantillon d'ARN $k = 1$ est marqué avec le fluorochrome rouge ($j = 1$), cette différence s'écrit :

$$\Delta_{ig} = (-1)^{i+1}(D_1 - D_2) + (V_1 - V_2) + \delta_g + (-1)^{i+1}\beta_g + F_{ig} \tag{5.2}$$

où pour des raisons de simplification, on note $\Delta_{ig} = Y_{i11g} - Y_{i22g}$, $\delta_g = (VG)_{1g} - (VG)_{2g}$, $\beta_g = (DG)_{1g} - (DG)_{2g}$ et $F_{ig} = E_{i11g} - E_{i22g}$.

La correction proposée est une correction globale et vise à éliminer les premiers termes de l'équation à savoir $(-1)^{i+1}(D_1 - D_2)$ et $(V_1 - V_2)$. Elle est supposée ne pas toucher à la différence d'expression spécifique à un gène entre deux conditions (quantité d'intérêt) δ_g et diminuer $(-1)^{i+1}\beta_g$.

Dans Martin-Magniette et al. [2005a] nous nous intéressons à ce dernier terme correspondant à un biais de marquage gène-spécifique. Nous montrons que ce biais peut être quantifié à l'aide de lames jaunes (lames sur lesquelles le même échantillon est marqué une fois avec un fluorochrome et une deuxième fois avec un second), qu'il peut être important et altérer les conclusions de l'analyse différentielle. Nous proposons un indicateur appelé LBI (Label Bias Index) permettant de l'évaluer.

5.1.2 Biais de marquage gène-spécifique

Le travail présenté ici est le résultat d'une collaboration avec Marie-Laure Martin-Magniette et Jean-Jacques Daudin pour la partie statistique et de personnes de la plateforme transcriptomique de l'Unité de Recherche en Génomique Végétale (maintenant intégrée à l'Institut des Sciences des Plantes de Paris Saclay) et de E. Cabannes qui était à l'époque à l'Institut Pasteur pour les aspects biologiques [Martin-Magniette et al., 2005a,b].

Méthodes

Nous reprenons le modèle (5.1) et nous intéressons au terme d'interaction $(DG)_{jg}$ qui correspond à l'effet de marquage gène-spécifique. Sur chaque lame, nous disposons de deux échantillons d'ARN $k = 1,2$ marqué chacun avec un fluorochrome $j = 1,2$. Si l'on suppose que l'échantillon 1 est marqué avec le fluorochrome 1 (Cy3), la différence d'expression observée entre les deux échantillons déposés sur la puce i pour le gène g après normalisation s'écrit :

$$\Delta_{ig} = \delta_g + (-1)^{i+1}\beta_g + F_{ig} \tag{5.3}$$

où pour des raisons de simplification d'écriture, nous notons $\delta_g = (VG)_{1g} - (VG)_{2g}$, $\beta_g = (DG)_{1g} - (DG)_{2g}$ et $F_{ig} = E_{i11g} - E_{i22g}$. A partir de ce modèle, la différence d'expression entre deux échantillons d'ARN et le biais de marquage gène-spécifique peuvent respectivement être estimés par :

$$\widehat{\delta}_g = \frac{1}{2p}\sum_{i=1}^{2p}\Delta_{ig}$$
$$\widehat{\beta}_g = \frac{1}{2p}\sum_{i=1}^{2p}(-1)^{i+1}\Delta_{ig}$$

avec p le nombre de dye-swaps. Un *dye-swap* est un couple de lames pour lequel sur la première lame, l'échantillon 1 est marqué avec le fluorochrome 1 et l'échantillon 2 avec le fluorochrome 2 et sur la deuxième lame, les mêmes échantillons sont utilisés mais marqués avec une inversion des fluorochromes. Si $p \geq 2$ alors il est également possible d'estimer la variance de F_{ig} de la façon suivante :

$$\widehat{\sigma}_g^2 = \frac{1}{2p-2}\sum_{i=1}^{2p}(\Delta_{ig} - \widehat{\delta}_g - (-1)^{i+1}\widehat{\beta}_g)^2.$$

Nous proposons un index de biais de marquage, appelé **L**abel **B**ias **I**ndex, défini comme la statistique associée au test de l'hypothèse nulle $\{\beta_1 = \cdots = \beta_G = 0\}$ contre l'alternative { au moins un β_g est non nul } :

$$\text{LBI} = \frac{2p\sum_{g=1}^{G}\widehat{\beta}_g^2}{\sum_{g=1}^{G}\widehat{\sigma}_g^2}. \tag{5.4}$$

Sous l'hypothèse nulle et si $\sigma_g^2 = \sigma^2$ alors le LBI est distribué selon une loi de Fisher à $\big((G-1),(2p-2)(G-1)\big)$ degrés de liberté. En pratique, l'hypothèse nulle risque d'être souvent rejetée puisque la puissance du test est grande. Aussi pour décider si un biais de marquage gène spécifique est important, le LBI peut aussi être comparé à l'espérance d'une distribution de Fisher donnée par $1 + \frac{1}{(p-1)(G-1)} - 1$. Quand un seul dye-swap est disponible, le modèle (5.3) est surparamétré et le paramètre β_g est supposé nul.

La seule façon d'estimer l'amplitude du biais de marquage gène spécifique est l'utilisation de lames jaunes. Les lames jaunes sont des lames sur lesquelles le même échantillon est marqué deux fois sur la même lame avec chacun des deux fluorochromes. Elles permettent d'évaluer la qualité des données et l'impact des biais techniques sur l'analyse différentielle. Sur des lames jaunes, le même échantillon est hybridé contre lui-même, ce qui garantit des δ_g nuls. Dans le cas de deux lames jaunes, en dye-swap, le LBI se définit par :

$$\text{LBI} = \frac{2p\sum_{g=1}^{G}(\Delta_{1g} - \Delta_{2g})^2}{\sum_{g=1}^{G}(\Delta_{1g} + \Delta_{2g})^2} \tag{5.5}$$

et est à comparer à la valeur attendue d'une distribution de Fisher, qui est proche de 1.
Le LBI donne un point de vue global sur le biais de marquage gène spécifique. De façon complémentaire, la nullité de β_g peut être testée afin de détecter les gènes affectés par un biais de marquage gène spécifique. Nous proposons d'utiliser la méthode proposée par Delmar et al. [2005] dont l'originalité réside dans l'identification via des modèles de mé-

lange des groupes de gènes de même variance et d'ajuster ensuite pour les tests multiples par la procédure de Bonferroni présentée dans la section 2.4.

Résultats

Le LBI a été calculé sur onze lames jaunes produites à partir d'échantillons humains, d'*Arabidopsis Thaliana*, de drosophiles ou de peuplier. Le LBI est toujours supérieur à 1 quelque soit le type de puces et nous avons montré qu'il était plus important dans les expériences humaines que dans celles sur *Arabidopsis Thaliana* (cf. Table 5.1). Nous n'avons pas recherché les causes techniques du biais de marquage mais avons préconisé l'utilisation de dye-swap quand le biais de marquage gène-spécifique est grand. En concordance avec les valeurs de LBI, le nombre de gènes détectés comme ayant un biais de marquage gène-spécifique est significatif sur les expériences humaines. Les niveaux d'intensité de ces derniers sont très variables, allant de très faiblement exprimés à très fortement exprimés. Ils sont par contre tous classés dans le groupe de plus forte variance. Cela peut suggérer que les gènes du plus fort groupe de variance peuvent ne pas être différentiellement exprimés seulement parce que leur variabilité augmente du fait d'un biais de marquage qui leur est propre.

Organisme / Puce	Jeu de données	LBI	(a)	(b)	Mean LR	Min. LR	Max. LR
humain/1	Wt t1	4.64	0	120	0.87	-1.19	1.58
humain/1	Control t2	5.68	0	153	0.45	-1.46	1.52
humain/1	SDF t3	4.07	0	2	1.97	1.73	2.21
humain/1	Wt t4	4.86	0	113	1.19	-1.16	1.81
humain/1	Control t5	4.64	0	33	0.19	-1.61	1.36
humain/1	SDF t6	6.42	0	189	1.42	-1.26	2.95
humain/ 2	SDF t2	10.29	0	8	-1.11	-1.35	-0.98
humain/ 1	SDF t2	5.15	0	3	-2.05	-2.85	-1.45
At/CATMA	leaf	1.79	0	0	-	-	-
At/CATMA	bud	1.17	0	0	-	-	-
At/CATMA	bud	1.24	0	0	-	-	-

TABLE 5.1 – Wt = Sauvage (Wild type), t=temps, LBI = Label Bias Index, At = *Arabidopsis thaliana* (a) : Nombre de gènes différentiellement exprimés, (b) : Nombre de gènes avec un biais de marquage significatif, LR = Log Ratio des gènes ayant un biais de marquage significatif.

Commentaires

Depuis la publication de cet article, l'origine de ce biais technique a été étudié par Kelley et al. [2008] et Margaritis et al. [2009]. Notre article a été cité 51 fois depuis sa parution, essentiellement pour justifier l'utilisation du dye-swap ou l'usage de répétitions

techniques lors d'expériences de puces à ADN. Il a encore été cité 4 fois ces dernières années pour préciser que la technologie des puces à ADN était entachée de biais et justifier l'usage d'autres technologies, notamment le RNA-seq (cf. p.6).

5.1.3 Normalisation de données issues de puces à plus de deux couleurs

Contexte

La plupart des études de puces à ADN utilise un ou deux fluochromes pour marquer les échantillons et permet ainsi l'hybridation d'un ou deux échantillons sur le même support. Les fluorochromes classiquement utilisés sont les Cyanines 3 (Cy3) et 5 (Cy5). Afin d'augmenter le nombre d'échantillons hybridés simultanément, Forster et al. [2004] ont proposé l'utilisation de deux marqueurs supplémentaires *Alexa 488* et *Alexa 594* et ont démontré que les puces à trois couleurs étaient techniquement possibles. En augmentant le nombre d'échantillons hybridés simultanément, les possibilités de comparaisons directes se multiplient, et les coûts diminuent. Les fluorochromes supplémentaires proposés, *Alexa 488* et *Alexa 594*, font partie de la gamme de fluorochromes *Alexa Fluor* produit par la société *Molecular Probes* dont le fils des fondateurs prénommé Alex a donné son nom. 488 et 594 correspondent à une couleur du spectre d'émission respectivement de cyan-vert, et orange-rouge. L'introduction de fluochrome supplémentaire permet plus de flexibilité dans la construction des plans expérimentaux mais introduit également un biais, causé par le transfert éventuel de signal, nommé *bleeding* d'un canal d'émission à un autre. Ce biais technique peut fausser les conclusions et en fonction du sens de ce transfert soit diminuer la capacité à détecter une différence, soit faire apparaître une différence qui n'est en fait qu'un biais technique. Forster et al. [2004] ont montré que ce phénomène est négligeable entre les deux fluorochromes usuels mais est important entre Alexa594 et Cy3. Pour étudier ce biais et mieux comprendre ce phénomène, nous avons proposé une méthode de normalisation en deux étapes. La première consiste à évaluer le biais de transfert de signal *bleeding* inhérent à l'utilisation d'un troisième fluorochrome et à le corriger si besoin. Et la deuxième étape consiste à une normalisation pour le biais de marquage global via une régression *lowess* généralisée à plus de deux couleurs.

Détection et correction du biais de transfert de signal ou *bleeding*

Forster et al. [2004] appelle canal blanc, un canal pour lequel aucun matériel biologique n'a été hybridé avec le fluorochrome associé. En théorie, ce canal blanc ne devrait produire aucun signal. La perception d'un signal est le signe d'un phénomène d'étalement de couleur, *bleeding*, d'un canal vers un autre et représente une source de biais technique. En effet quand un signal est fort dans un canal, de par ce phénomène il peut augmenter artificiellement le signal d'un autre canal. Afin d'étudier ce biais, nous avons hybridé sur différentes lames un seul échantillon avec un seul fluorochrome et analysé les

Données	Cy5 → Cy3	Cy5 → Alexa	Cy3 → Cy5	Cy3 → Alexa	Alexa → Cy5	Alexa → Cy3
Forster	1 (1)	6 (3)	0.5 (0.5)	26 (14)	2.5 (0.3)	27 (5)
URGV1	1.0 (0.1)	52 (5)	0.0 (0)	26 (2)	5 (0.4)	70(15)

TABLE 5.2 – *Bleeding* : coefficient de régression entre le canal hybridé et les canaux hybridés et blancs pour les jeux de données Forster et URGV1. Moyenne (se) du coefficient de régression (x1000).

données provenant des trois canaux. Nous proposons un modèle pour évaluer et corriger ce biais à l'aide d'expériences dans lesquelles un même échantillon a été marqué avec un seul fluorochome sur une lame. Nous disposons de 6 lames (1 échantillon ×2 répétitions ×3 fluorochromes. Sur chaque lame, le signal est lu dans le canal utilisé effectivement et dans les deux canaux blancs.
Nous observons bien des transferts de signaux [Martin-Magniette et al., 2008]. Le biais dépend cependant du canal. Le biais $Cy3 \rightarrow Cy5$ est négligeable mais le biais $Alexa594 \rightarrow (Cy5, Cy3)$, c'est-à-dire du canal jaune vers les canaux rouge et vert existe.

En notant respectivement G, Y et R le signal vert, jaune et rouge et en indexant par i la sonde nous proposons de quantifier ces transferts de signaux d'un canal à l'autre via les modèles de régression linéaire suivants :

$$\begin{cases} G_i = \alpha_1 + \beta_1 Y_i + \epsilon_i \\ R_i = \alpha_2 + \beta_2 Y_i + \epsilon'_i. \end{cases}$$

Nous estimons les effets par une méthode robuste (implémentée dans la fonction R *rlm*, Huber [1981]) pour limiter les effets des points aberrants. Les résultats (table 5.2) montrent que l'impact de ce transfert est faible sur le signal, l'effet le plus fort est observé entre Alexa594 et Cy3 (0.07). La faible influence quantitative est confirmée par les valeurs de l'erreur standard du signal entre les différents canaux. Les valeurs dans les canaux vides sont 6 à 200 fois plus faibles que dans les canaux hybridés (résultats visibles dans l'article Martin-Magniette et al. [2008] disponible en annexe D). Néanmoins ces conclusions sont valables pour les seuls fluorochromes testés. Il est possible que d'autres marqueurs ou d'autres technologies entrainent un biais plus important. Dans tous les cas, un plan d'expérience dans lequel les fluorochromes sont équilibrés est aussi une solution pour diminuer ce biais de transfert de signal étant donné que les différences d'expression entre les conditions est la différence entre les mesures individuelles prises sur chaque puce.
Nous proposons une procédure pour corriger cet effet quand il est élevé.

Correction du biais de marquage global par une *lowess* généralisée Même si le biais de transfert est un biais nouveau important pour les puces à plus de deux couleurs, le biais le plus important reste le biais de marquage global. Dans le cas de puces deux couleurs,

Source	Avant normalisation	Après normalisation
Puce	1191	1184
Fluorochrome	13269	11
Puce*Fluorochrome	425	43
Gène	310836	309177
Puce*Gène	6362	6378
Fluorochrome*Gène	10595	2739
Condition*Gène	2387	2105
Résidus	24890	23929

TABLE 5.3 – Somme des carrés avant et après normalisation (jeu de données URGV3).

la méthode de normalisation la plus usuelle est la normalisation *lowess* (cf. section 2.5) mais elle est proposée pour deux couleurs et n'est pas directement applicable dans le cas où le nombre de canaux est plus grand. C'est pourquoi nous proposons de la généraliser.

Soit $i = 1, \cdots, n$ l'indice de la sonde et $j = 1, \cdots, p$ l'indice du canal et y_{ij} la mesure d'intensité après log-transformation de la sonde i dans le canal j. Nous modélisons le signal dans un canal blanc de la façon suivante :

$$\mathrm{D}_{ij} = f_j(\overline{\mathrm{Y}_i}) + \mathrm{E}_{ij} \tag{5.6}$$

f_j est estimée par une lowess. $\overline{\mathrm{Y}_i} = \frac{1}{p}\sum_j \mathrm{Y}_{ij}$ est la moyenne des données brutes pour le gène i sur l'ensemble des canaux en échelle log. La valeur dans le canal j après normalisation est définie comme :

$$\widetilde{\mathrm{Y}_{ij}} = \mathrm{Y}_{ij} - f_j(\overline{\mathrm{Y}_i}) = \overline{\mathrm{Y}_i} + \mathrm{E}_{ij}. \tag{5.7}$$

Résultats

Comme précédemment et suivant la stratégie proposée par Kerr et al. [2002], nous avons effectué une analyse de variance en utilisant le même modèle sur les données brutes puis sur les données normalisées. Une méthode de normalisation efficace doit réduire la somme des carrés due aux facteurs techniques sans diminuer celle due au facteur biologique d'intérêt. Une autre façon de vérifier l'impact de la normalisation est de vérifier le nombre de gènes déclarés différentiellement exprimés. Dans une expérience de lames jaunes, aucun gène ne devrait être dans ce cas.

Les transcriptomes des mêmes échantillons ont été produits avec les technologies 2 couleurs (6 lames) et 3 couleurs (3 lames). A l'issue des étapes de normalisation et d'analyse différentielle, plus de gènes ont été déclarés différentiellement exprimés avec la technologie 3 couleurs qu'avec la technologie 2 couleurs.

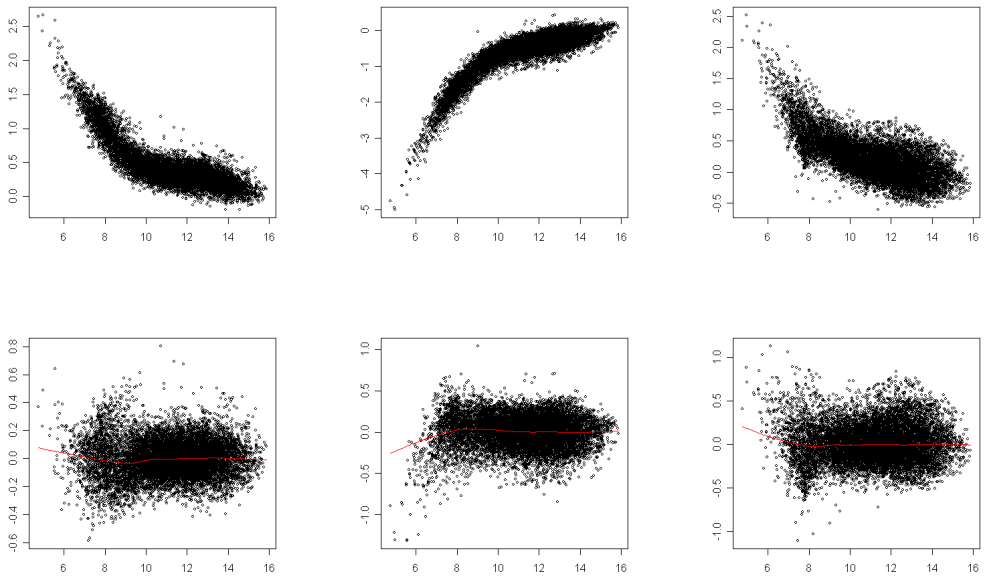


FIGURE 5.2 – MA-plots modifiés, axe des x : intensité moyenne, axe des y : différence entre les intensités du canal considéré et les intensités moyennées sur tous les canaux. Première ligne : données brutes, dernière ligne : données normalisées. Première colonne : Cy5, deuxième colonne : Cy3, troisième colonne : A$lexa$594.

Les technologies à plus de deux couleurs étaient jugées prometteuses dans le milieu des années 2000 et nous avons participé à un projet ANR AgriArray visant à identifier les biais liés à l'intégration de fluorochromes supplémentaires et à leur correction. Dans ce cadre, nous avons à la fois, identifié et proposé une méthode pour corriger le biais de *bleeding* et proposé une généralisation de la méthode *lowess* pour corriger le biais de marquage global. La technologie de séquençage de l'ARN a pris le dessus sur les puces et les puces à plus de deux couleurs n'ont pas vraiment pris leur essor ce qui explique que la méthode de normalisation que nous avons proposée n'a quant à elle été que très peu utilisée.

5.1.4 Comparaison de méthodes de normalisation de données RNA-Seq

Chaque nouvelle technologie qui apparait est de plus en plus performante et semble ne plus nécessiter d'étape de normalisation : "One particularly powerful advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets" [Wang et al., 2009]. Cependant ce n'est qu'une courte illusion, chaque nouvelle technologie apporte aussi son lot de biais techniques.

L'étape de normalisation des données est une étape déterminante, bien qu'ingrate. Le développement d'une méthode de correction de biais techniques peut être assez longue. Elle nécessite de disposer de données, de détecter le biais, de développer la méthodologie pour la détection et la correction de ce biais et surtout de connaitre en partie la vérité

pour juger de la pertinence de la méthode. Les technologies évoluant assez vite, la méthode peut être obsolète avant d'être publiée : c'est une course contre la montre parfois contre-productive. La méthode présentée dans la section précédente a fait l'objet d'une publication scientifique mais n'a jamais été utilisée, la technologie 3 et 4 couleurs n'ayant jamais vraiment percée du fait de l'arrivée d'une nouvelle technologie, le séquençage de l'ARN.

Cette étape de normalisation étant néanmoins très importante, un choix a été fait dans l'unité de ne plus investir dans le développement de nouvelles méthodes mais d'effectuer une veille technologique et bibliographique importante. Suite à l'émergence des technologies de séquençage de l'ARN, les données ont commencé à s'accumuler et quelques méthodes de normalisation à être proposées mais aucun consensus clair n'émergeait sur la méthode la plus appropriée. C'est dans le but de comparer ces méthodes sur leurs capacités à corriger les biais techniques inhérents à l'utilisation d'une technologie de séquençage d'ARN et sur leur impact sur la détection de gènes différentiellement exprimés, que quelques membres du groupe StatOmique (http://www.sfbi.fr/statomique), que je co-anime, ont décidé de mutualiser leurs efforts. Nous avons comparé les méthodes existantes ainsi que certaines développées pour l'ancienne technologie de référence en transcriptomique, soit sept méthodes au total. Afin de représenter le mieux possible la pratique, nous avons comparé ces méthodes à la fois sur des données simulées et sur des données réelles impliquant des espèces différentes, des plans d'expérience différents. Cette étude nous a permis de faire quelques recommandations pratiques sur la méthode de normalisation à utiliser et sur son impact sur l'analyse de données issues du séquençage de l'ARN messager [Dillies et al., 2012].

Méthodes

La source de variation la plus évidente entre échantillons est la différence en taille de banque (*library size*) ou profondeur de séquençage. Plus la profondeur de séquençage (ou nombre total de lectures obtenues par échantillon) sera grande, plus les comptages attribués à un gène seront élevés, et ce, quelque soit le niveau d'expression de ce gène ou sa longueur. On note y_{gk} le nombre de lectures associées au gène g dans l'échantillon k et m le nombre total d'échantillons de l'expérience. La forme la plus simple de normalisation inter-échantillon consiste à normaliser chaque échantillon par un simple facteur échantillon-spécifique noté s_k. Nous avons considéré sept méthodes différentes (Table 5.4) pouvant se regrouper en quatre groupes :

1. Les trois premières méthodes *Total Counts (TC), Upper Quartile (UQ)* [Bullard et al., 2010] et *Médiane (Med)* supposent que les comptages sont proportionnels au niveau d'expression et à la profondeur de séquençage. Elles utilisent respectivement N_k le nombre total de lectures (ou taille de banque) de l'échantillon k, les troisième

(Q_{3k}) et deuxième quartiles (Q_{2k}) de la distribution des comptages dans l'échantillon k dans le calcul de ce facteur de normalisation.

2. La normalisation *Full Quantile (FQ)* [Bolstad et al., 2003; Yang and Thorne, 2003] issue des puces à ADN consiste à rendre similaire les distributions des comptages de chaque échantillon en projetant le vecteur des quantiles de la distribution des comptages de chaque échantillon sur la diagonale.

3. Les normalisations *DESeq* encore appelée *Relative Log Expression (RLE)* [Anders and Huber, 2010] et *Trimmed Mean of M-values (TMM)* [Robinson and Oshlack, 2010] intrinsèques au modèle d'analyse différentielle sont basées sur le concept de taille efficace de banque et repose sur l'hypothèse que la majorité des gènes est invariante entre deux échantillons.
   Ces méthodes définissent un sous-ensemble de gènes invariants ou une référence qui servira de base pour le calcul du facteur de normalisation. La méthode TMM définit ce sous-ensemble G^* comme l'ensemble des transcrits ayant des comptages non nuls, en moyenne ni trop faibles, ni trop forts et n'ayant pas de fortes variations entre les différentes conditions testées. De façon plus précise, une référence est choisie, ici k', pour chaque échantillon k les transcrits ayant les valeurs de $A_{gk}^{(k')} = 0.5 \times [\log_2(\frac{Y_{gk}}{N_k} \times \frac{Y_{gk'}}{N_{k'}})]$ parmi les 5% les plus extrêmes ou les valeurs de log-ratios $M_{gk}^{(k')} = \log_2(\frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}})$ parmi les 30% les plus extrêmes ne sont pas pris en compte dans le calcul du facteur de normalisation. Un facteur est calculé pour chaque échantillon par rapport à la référence comme une moyenne pondérée des $M_{gk}^{(k')}$ sélectionnés après filtre $\log_2(s_k^{(k')} = \frac{\sum_{g\in G^*} w_{gk}^{(k')} M_{gk}^{(k')}}{\sum_{g\in G^*} w_{gk}^{(k')}}$ avec $w_{gk}^{(k')} = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_{k'} - Y_{gk'}}{N_{k'} Y_{gk'}}$; $Y_{gk}, Y_{gk'} \leq 0$.
   Une renormalisation est effectuée pour éviter une dépendance à la référence choisie.

4. La normalisation *Reads per Kilobase per Million mapped reads (RPKM)* [Mortazavi et al., 2008] combine une normalisation inter et intra-échantillon en normalisant à la fois par rapport au nombre total de lectures dans chaque échantillon (en million) et par la taille du gène L_g (en kilobase). A noter que cette méthode permet la comparaison du niveau d'expression des gènes d'un même échantillon. L'estimation du nombre de lectures est non biaisée mais affecte la variance [Oshlack et al., 2010].

Ces sept méthodes de normalisation sont aussi comparées aux données brutes non normalisées, appelées *Raw Counts (RC)*. Toutes les analyses ont été faites avec `R 2.14`. Les scripts sont disponibles en matériel supplémentaire de l'article de Dillies et al. [2012].

Données réelles Les sept méthodes de normalisation décrites précédemment ont été comparées sur 4 jeux de données RNA-Seq réelles impliquant des espèces différentes, des plans d'expériences différents, des caractéristiques différentes en terme de reproduc-

Méthode	$\hat{s}_k$	package R fonction
TC	$\frac{N_k}{\frac{1}{m}\sum_k N_k}$	
UQ	$\frac{Q_{3k}}{\frac{1}{m}\sum_k Q_{3k}\mathbb{1}\{Q_{3k}>0\}}$	
Med	$\frac{Q_{2k}}{\frac{1}{m}\sum_k Q_{2k}\mathbb{1}\{Q_{2k}>0\}}$	
Full Quantile		limma : : normalizeQuantiles
DESeq	$median_g\left(\frac{y_{gk}}{(\pi_{v=1}^{m} y_{gv})^{1/m}}\right)$	DESeq : : estimateSizeFactors
RLE		edgeR : : calcNormFactors(method='RLE')
TMM		edgeR : :calcNormFactors(method='TMM')
RPKM	$\frac{N_k * L_g}{10^3 * 10^6}$	

TABLE 5.4 – Facteur de normalisation ou implémentation associé(e) à chacune des méthodes comparées

tibilité entre répétitions, présence de séquences majoritaires, profondeur de séquençage, importance de la longueur du gène dans l'estimation de l'expression du gène (Table 5.5). La description précise des jeux de données est disponible en annexe E.

Organisme	Type	Nombre de gènes	Répétitions par condition	Taille minimale de banque	Taille maximale de banque	Corrélation entre répétitions	Corrélation entre conditions	% gène le plus exprimé	Type de banque	Séquenceur
H. sapiens	RNA	26437	{3,3}	2.0×10^7	2.8×10^7	(0.98,0.99)	(0.93,0.96)	≈ 1%	SR 54, ND	GaIIx
A. fumigatus	RNA	9248	{2,2}	8.6×10^6	2.9×10^7	(0.92,0.94)	(0.88,0.94)	≈ 1%	SR 50, D	HiSeq2000
E. histolytica	RNA	5277	{3,3}	2.1×10^7	3.3×10^7	(0.85,0.92)	(0.81,0.98)	6.4-16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	{3,2,2}	2.0×10^6	5.9×10^6	(0.95,0.99)	(0.09,0.75)	17.4-51.1%	SR 36, D	GaIIx

TABLE 5.5 – Résumé des jeux de données utilisés pour la comparaison des méthodes de normalisation, incluant le type de banque (SR = single-read or PE = paired-end read, D = directionnel or ND = non-directionnel).

Procédures de comparaison **Caractéristiques qualitatives des données normalisées :** Pour chaque jeu de données, les méthodes ont été comparées sur la base de caractéristiques qualitatives telles que la distribution des comptages, la variabilité entre répétitions. Des boîtes à moustaches des données brutes et normalisées ont été effectuées sur les données transformées $\log_2(\text{comptage}+1)$ afin d'éviter les problèmes associés aux valeurs nulles. La variabilité intra-condition a été mesurée à l'aide d'un coefficient de variation par gène. Sa distribution à travers les échantillons a été représentée sous forme de boîte à moustaches.

Nous avons regardé la variation moyenne d'un sous-ensemble de 30 gènes de ménage humains, en supposant que ces gènes sont de vrais gènes de ménage, c'est-à-dire qu'étant indispensables à la vie de tous les types de cellules, ils sont toujours exprimés, sans mécanisme de régulationet de façon similaire à travers les échantillons. Ces gènes de ménage proviennent de la liste de Eisenberg and Levanon [2003] et présentent une variabilité minimale calculée sur 84 types cellulaires humains des données de GeneAtlas [Su et al., 2004] disponible sur `GEO` (`http://www.ncbi.nlm.nih.gov/geo`) avec le numéro d'accession GSE1133.

Analyse différentielle : Les méthodes ont également été comparées sur la base des résultats de l'analyse différentielle effectuée avec le package Bioconductor `DESeq`, qui a été spécifiquement développé pour la détection de gènes différentiellement exprimés pour les données RNA-Seq data avec peu de répétitions biologiques et présence de surdispersion. Cette méthode est basée sur une distribution Binomiale Négative et une régression locale pour estimer la relation entre moyenne et variance de chaque gène. Le package Bioconductor `DESeq` (version 1.4.0) avec les valeurs par défaut ont été utilisé. Les probabilités critiques ont été ajustées par la procédure de Benjamini-Hochberg (cf. section 2.4). Les gènes avec une probabilité critique ajustée inférieure à 0.05 sont considérés différentiellement exprimés.

Nous avons comparé le nombre de gènes différentiellement exprimés (DE), le nombre de gènes en commun, et pour chaque jeu de données nous avons généré un dendrogramme représentant la similarité entre les listes de gènes DE détectés par chaque méthode, basée sur une distance binaire et un critère d'agrégation de Ward. Les quatre dendrogrammes obtenus ont ensuite été synthétisés en un dendrogramme "consensus" résultant de la moyenne des distances de matrice obtenues pour chaque jeu de données réelles.

Simulations Le modèle de simulation adopté est similaire à celui utilisé par Jeanmougin et al. [2010] et adapté à des comptages. Soit G le nombre de gènes et m le nombre d'échantillons répartis en deux conditions c_1 et c_2. Soit y_{gk} la valeur de l'expression d'un gène donné g dans l'échantillon k. Nous supposons que y_{gk} suit une distribution de Poisson de paramètre $\lambda_g c_k$ dépendant de la condition c_k à laquelle appartient l'échantillon k.

Sous ce modèle, l'hypothèse nulle H_0 de non différence entre les deux conditions est équivalente à $\lambda_g c_2 = \lambda_g c_1$; l'hypothèse alternative H_1 d'expression différentielle entre les deux conditions est équivalente à $\lambda_g c_2 = \tau \times \lambda_g c_1$, où τ correspond à la magnitude de l'expression différentielle entre les deux conditions. Soit π_0 (π_1) les proportions de gènes générés respectivement sous H_0 (H_1) parmi les G gènes.

Les données ont été simulées avec G = 15000, $m = 20$ (dix échantillons par condition) et π_1 allant de 0% à 30%. Afin de générer des données réalistes, le paramètre $\lambda_g c_1$ utilisé pour échantillonner le gène g à partir d'une distribution de Poisson pour la condition c_1 correspond à la moyenne d'expression observée pour chaque gène estimé à partir des données d'un des jeux de données analysés dans l'étude (ici le jeu de données *M. musculus*) ; le paramètre $\lambda_g c_2$ utilisé pour échantillonner le gène g à partir d'une distribution de Poisson pour la condition c_2 est égal à $\lambda_g c_1$ sous H_0 et à $\tau \times \lambda_g c_1$ sous H_1, avec $\tau = \pm 0.2$.

Pour évaluer l'impact de tailles de banques non équivalentes, nous ajoutons la possibilité de multiplier toutes les valeurs d'expression de gène y_{gk} pour un échantillon donné k par une constante K_k égale à $|\epsilon|$, où ϵ est tiré selon une loi $\mathcal{N}(1,1)$. De plus, le jeu de données *M. musculus* contient un ensemble de gènes hautement exprimés contribuant à la majorité des comptages totaux, ce qui permet d'évaluer l'impact de tels gènes majoritaires dans les données simulées.

Taux de faux positifs et puissance : Pour chaque jeu de données simulé, le taux de faux positifs (respectivement la puissance) peut être estimé sur la base des gènes simulés sous H_0 (respectivement H_1). Nous avons considéré trois cas différents :

1. des tailles de banque équivalentes et pas de gènes majoritaires
2. des tailles de banque non équivalentes et pas de gènes majoritaires
3. des tailles de banque équivalentes et la présence de gènes majoritaires.

Pour chaque scénario et proportion de H_1 testée, le taux de faux positifs et la puissance ont été moyennés sur 10 jeux de données simulés pour garantir une précision raisonnable.

Résultats

Sur les données réelles, ne connaissant pas la vérité, les méthodes ont été comparées d'après les différents critères cités précédemment et celles ne se comportant pas de façon correcte ont été jugées non pertinentes, ce qui ne veut pas forcément dire à ce stade que les autres sont adéquates.

Les boîtes à moustaches (boxplots) des données normalisées (Figure 5.3 (a)) montrent de faibles différences entre les méthodes similaires par nature. Cependant pour le jeu de données *M. musculus* qui comporte des transcrits majoritaires (qui captent la majorité des séquences), on remarque que les méthodes utilisant le nombre total de séquences

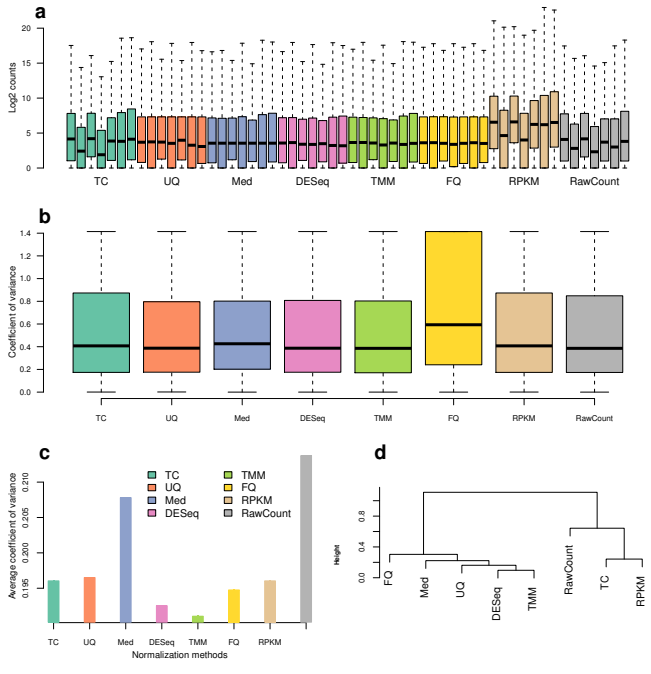


FIGURE 5.3 – Comparaison des méthodes de normalisation sur les données réelles. (a) Boxplots des $\log_2$(comptages + 1) pour toutes les conditions et répétitions du jeu de données *M. musculus*, par méthode de normalisation. (b) Boxplots de la variance intra-groupe pour une des conditions du jeu de données *M. musculus*, par méthode de normalisation. (c) Analyse des gènes de ménage pour le jeu de données *H. sapiens*. (d) Dendrogramme consensus des résultats de l'analyse différentielle, utilisant le package Bioconductor DESeq, pour toutes les méthodes de normalisation et pour les quatre jeux de données considérés.

TC et RPKM sont inefficaces et que les boîtes à moustaches sont similaires à celle réalisée avec les données brutes. La processus de normalisation visant à éliminer une source de variation, il est attendu que la variance intra-condition soit plutôt diminuée. Pour la plupart des jeux de données, nous observons peu de différences entre les différentes méthodes (Annexe E Figure S2). Là encore, seul le jeu de *M. musculus* permet de mettre en évidence un comportement inattendu de la méthode UQ qui augmente la variance intra-condition (Figure 5.3 (b)). Cela se comprend en regardant plus attentivement les distributions des comptages pour les sept échantillons de l'expérience. Un échantillon a une distribution très différente des autres (plus de faibles et forts comptages, et moins de comptages intermédiaires). La normalisation UQ en forçant l'ensemble des distributions à correspondre à la distribution moyenne, a surcorrigé cet échantillon et ainsi augmenté la variabilité intra-condition. La figure 5.3 (c) représentent le coefficient de variation calculé sur le jeu de donnés à partir d'un ensemble de 30 gènes de ménage supposés avoir une expression constante quelque soit la condition étudiée. Cette figure montre que les méthodes DESeq et TMM sont celles qui aboutissent au plus faible coefficient de variation. La figure 5.3 (d) confirme que le comportement de méthodes TC et RPKM est similaire à la méthode qui consiste à ne pas normaliser (données brutes) et qu'elles sont donc inefficaces et à abandonner dans le cadre de l'analyse différence à partir de données issues de séquençage de l'ARN.

Dans le cas de données simulées, la vérité est connue. Nous pouvons faire varier les caractéristiques des jeux de données et ainsi faire apparaître de façon plus marquée les différences entre les différentes méthodes. Dans le cas où la normalisation n'est pas nécessaire (pas de gène majoritaire et des tailles de banque similaires), toutes les méthodes donnent des résultats similaires à ceux obtenus sur les données brutes en terme de faux positifs et puissance, ce qui est attendu la normalisation n'étant pas nécessaire dans ce cas (cf. matériel supplémentaire de l'article en annexe E Figure S5a). Dans le cas de tailles de banque différentes, le taux de faux positifs n'est pas maintenu et la puissance est plus faible pour les données brutes, mais l'ensemble des méthodes donnent des résultats similaires. Le cas le plus discriminant (Figure 5.4) où les tailles de banque sont différentes et quelques gènes majoritaires sont présents, montrent que la présence de ces gènes à fort comptage conduit à un taux de faux positifs plus grand qu'attendu pour cinq des sept méthodes (TC, UQ, Med, Q et RPKM) et que seules DESeq et TMM sont capables de contrôler le taux de faux positifs tout en maintenant la puissance de détection de gènes différentiellement exprimés.

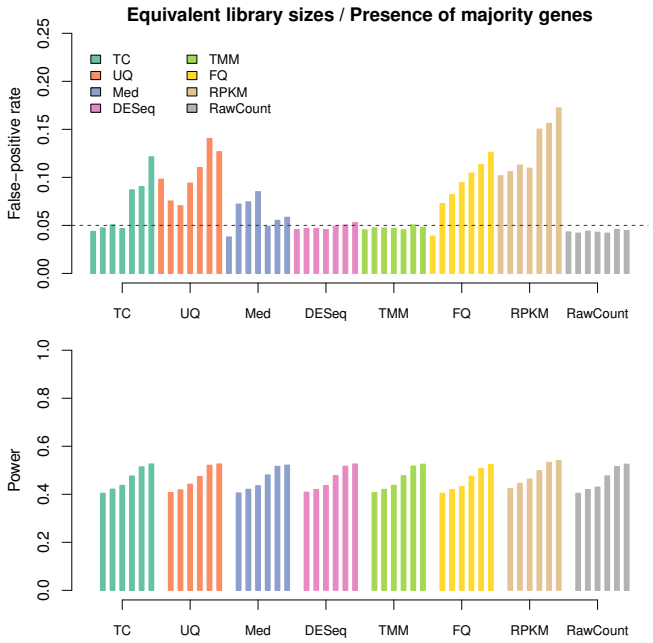


FIGURE 5.4 – Comparaison des méthodes de normalisation pour les données simulées avec des tailles de banque égales et des présences de gènes à fort comptages. Respectivement le taux de faux positif (en haut) et la puissance (en bas) moyennés sur 10 jeux de données indépendants simulés avec des proportions de gènes différentiellement exprimés variant de 0% à 30% pour chaque méthode de normalisation.

En conclusion, les différences entre les méthodes apparaissent lors d'une forte variation de tailles de librairie ou de séquences majoritaires. Les méthodes TMM et DESeq apparaissent comme des méthodes performantes et robustes dans un cadre d'analyse différentielle au niveau du gène.

Commentaires

Le travail de comparaison de méthodes bien que fort utile en pratique est très délicat. Pour convenir qu'une méthode est meilleure, il faut connaître la vérité : cela est possible sur des données simulées mais ces dernières n'ont que rarement l'ensemble des caractéristiques et notamment des biais qui entachent les données réelles. Et la connaissance de la vérité sur des données réelles nécessite de nombreuses expériences ou un énorme travail de validation biologique. Ce travail de comparaison a été possible grâce à un travail collectif et une mise en commun des jeux de données et connaissances de chacun. Il a permis de corriger quelques bugs présents dans les packages. Il a également donné une certaine visibilité au groupe StatOmique et a permis entre autre le financement par l'Institut Pasteur d'une journée d'animation au cours de laquelle nous avons pu faire venir des spécialistes du domaine. En résumé c'est un travail très utile, l'article a à ce jour été cité 165 fois (Source WebOfSciences), tweeté 66 fois et est présent sur 2 pages wikipédia.

5.2 Planifier pour avoir de meilleurs résultats

Intervenir dès le plan d'expérience est la situation idéale pour tout projet et permet à la fois d'éviter des biais malencontreux et de garantir un minimum de puissance statistique. Un dialogue est donc nécessaire entre tous les acteurs du projet bien en amont et quelques règles souvent de bon sens peuvent être appliquées (cf. annexe F). Une fois les données produites, elles sont parfois analysées en routine à l'aide de différents 'pipeline d'analyse'. Ces différents enchainements de traitement ne permettent pas toujours une analyse correcte des données issues de certains plans particuliers. C'était notamment le cas du plan en dye-switch utilisé dans le cadre d'expériences de puces à ADN à deux couleurs. Nous présenterons dans la deuxième partie de ce chapitre une méthode développée dans ce cadre.

5.2.1 Quelques règles de bonnes pratiques pour planifier une expérience de séquençage de l'ARN

Afin de donner quelques règles de bonnes pratiques pour planifier des expériences de séquençage de l'ARN, un travail de revue a été fait avec quelques personnes au sein d'un réseau national INRA de praticiens des données omiques (PEPI Ingéniérie Bio-Informatique et Statistique des donnés à haut débit). Ce travail a permis de faire une bibliographie commune au sein de différents champs disciplinaires (bioinformatique, statistique, biologie) et a abouti à un poster intitulé 'How to Design a good RNA-Seq experiment in an interdisciplinary context ?' présenté en annexe F. Les règles principales :

1. Clarifier la question biologique.

2. Préférer les répétitions biologiques aux répétitions techniques.
3. Multiplexer.
4. Le compromis optimal entre nombre de répétitions et profondeur de séquençage dépend de la question posée.
5. Appliquer les principes de Fisher (randomisation, répétition et contrôle local) à chaque fois que c'est possible.

Ces règles sont reprises au sein des plateformes de bioinformatiques et énoncées lors des formations faites autour de l'analyse de données RNA-Seq.

5.2.2 Dye-Switch

Le travail présenté dans cette section est le résultat d'une collaboration avec Tristan Mary-Huard et Jean-Jacques Daudin pour la partie statistique. J'ai participé à l'implémentation de la méthode et à l'application sur des données issues d'une collaboration avec N. Mansouri et O. Sandra de l'Unité Mixte de Recherche Biologie de la Reproduction de l'INRA de Jouy-en-Josas.

Contexte

Dans le cadre d'expériences de puces à deux couleurs, les plans équilibrés avec inversion des fluorochromes peuvent être divisés en 3 classes (cf. Table 5.6) selon la nature des contraintes d'équilibre :

1. Le plan est équilibré pour l'utilisation des fluorochromes de façon globale, c'est-à-dire au niveau de la condition biologique : chaque échantillon biologique est hybridé une seule fois et donc présent avec un seul fluorochrome, sur une seule puce.
2. Le plan est équilibré individuellement. Chaque échantillon est divisé en deux, une partie est hybridée avec Cy3 sur une lame et l'autre avec Cy5 sur une autre lame. Chaque échantillon biologique est hybridé exactement deux fois.
3. Le plan est équilibré pour les couples d'échantillons biologiques. Un même couple d'échantillons (issus de deux conditions biologiques à comparer) est hybridé sur deux lames avec inversion des fluorochromes sur la deuxième lame.

Ce troisième type de plan est ce qu'on appelle un plan en dye-swap, le plus fréquemment utilisé quand l'erreur technique est plus importante que la variabilité biologique ou quand un biais de marquage gène-spécifique (cf. section 5.1.2) est supposé. Cependant l'utilisation de répétitions techniques se fait souvent au détriment des répétitions biologiques. Les premier et deuxième types de plan sont tous les deux appelés plan en dye-switch. L'analyse statistique issue de ces différents plans peut être très différente. Les cas 1 et 3 sont assez directs et bien décrits, les unités expérimentales étant indépendantes. Le dye-switch cas 2, corrige effectivement du biais de marquage mais peut introduire une

dépendance dans les mesures en $\log_2$. Dans Mary-Huard et al. [2008], nous avons proposé deux procédures adaptées à ce type de plan.

1	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B5	A3	B9	A5	B6	A7	B10	A9	B9
	Cy3	B3	A2	B8	A4	B2	A6	B1	A8	B4	A10
2	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
	Cy3	B1	A2	B2	A3	B3	A4	B4	A5	B5	A1
3	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
	Cy3	B1	A1	B2	A2	B3	A3	B4	A4	B5	A5

TABLE 5.6 – Trois différents types de plans avec inversion de fluorochromes pour la comparaison de deux traitements (A et B), avec un nombre égal de lames. A_i correspond au $i^{\text{ème}}$ échantillon biologique dans la condition A. (1) Plan globalement équilibré, avec 10 échantillons biologiques par condition. (2) Plan équilibré individuellement avec 5 échantillons biologiques par condition. (3) Plan en dye-swap avec 5 échantillons biologiques par condition.

Ces procédures sont comparées aux procédures de type maximum de vraisemblance ou REML (restricted maximum likelihood) sur des données réelles et des données simulées.

Méthodes

Les données, une fois normalisées, peuvent se modéliser de la façon suivante :

$$\begin{aligned} Y_{Ai} &= \mu_A + \delta_{l(i)} + B_{j(i)} + M_i + T_i \\ Y_{Bi} &= \mu_B + \delta_{l'(i)} + B_{j'(i)} + M_i + T'_i \ , \end{aligned} \tag{5.8}$$

où

- μ_A et μ_B sont les mesures d'expression des conditions A et B.
- $\delta_{l(i)}$ est un effet fixe à deux niveaux correspondant à l'effet fluorochrome. $\delta_{l(i)} = \delta_1$ (resp. δ_2) pour tous les échantillons marquées avec Cy5 (resp. Cy3). Ce terme tient compte du *biais de marquage gène-specifique*.
- $B_{j(i)}$ représente un terme aléatoire gaussien de moyenne nulle et de variance σ^2_B, correspondant à un effet aléatoire de l'échantillon $j(i)$. Cette variable est spécifique de l'échantillon biologique et est appelée *erreur biologique*, reflétant la variabilité du matériel biologique au sein de chaque population A et B.
- M_i représente un vecteur de terme aléatoire gaussien de moyenne nulle et de variance σ^2_M. M_i est l'effet du spot associé au gène concerné sur la puce i et a la même valeur pour deux échantillons hybridés sur la puce i. Ce terme d'erreur prend en compte l'hétérogénéité spatiale sur la puce qui affecte les mesures des deux fluorochromes.

— T_i représente un terme aléatoire gaussien de moyenne nulle et de variance σ^2_T, correspondant à la variabilité technique, incluant les étapes de marquage, d'hybridation et mesure de l'intensité de la fluorescence. Cette variable a une valeur spécifique pour chaque combinaison gène×fluorochrome×échantillon, même si les échantillons sont hybridés sur une même puce à un même spot, de telle sorte que T_i et T'_i sont des variables aléatoires indépendantes. T_i et M_i sont les deux composantes de *l'erreur technique*.

Modèles sur la différence d'expression Soit $D_i = Y_{Ai} - Y_{Bi}$, $i = 1,...,2n$. En utilisant (5.8) on obtient :

$$D_i = \mu_A - \mu_B + B_{j(i)} - B_{j'(i)} + \delta_{l(i)} - \delta_{l'(i)} + T_i - T'_i \tag{5.9}$$

qui peut aussi s'écrire

$$D_i = \mu + BD_i + (-1)^{i+1}\delta + TD_i \tag{5.10}$$

où

— $\mu = \mu_A - \mu_B$ est la vraie différence d'expression entre les conditions A et B pour le gène concerné,
— $BD_i = B_{j(i)} - B_{j'(i)}$ est une variable aléatoire de moyenne nulle et de variance $\sqrt{2}\sigma^2_B$,
— $TD_i = T_i - T'_i$ est une variable aléatoire indépendante de moyenne nulle et de variance $\sqrt{2}\sigma^2_T$,
— $\delta = \delta_1 - \delta_2$ est la vraie différence entre les effets des fluorochromes Cy3 et Cy5. Ce terme tient compte du biais de marquage gène spécifique.

Chaque variable D_i suit une loi normale de moyenne $\mathbb{E}(D_i) = \mu + (-1)^{i+1}\delta$ et de variance $\mathbb{V}(D_i) = 2\sigma^2_B + 2\sigma^2_T$. Tous les termes de covariances $\text{cov}(D_i, D_j)$ sont nuls sauf les suivants :

$$\text{cov}(D_i, D_{i+1}) = \sigma^2_B$$

avec par convention $2n+1 = 1$.

Dans cette étude, nous présentons et comparons différentes stratégies pour l'analyse statistique de plans d'expérience équilibrés sur les individus.

Procédures testées Nous avons proposé deux procédures de test adaptées aux dispositifs en dye-switch, la méthode *Unbaised Paired Method (UP)* adaptée au cas apparié et la méthode *Unbaised Unpaired Method (UU)* adaptée au cas non apparié. L'efficacité de ces procédures a été comparée à celle de trois procédures plus classiques pour l'analyse différentielle (Table 5.6). Les procédures sont les suivantes (voir la section Méthodes de l'article [Mary-Huard et al., 2008] pour plus de détails) :

— Méthode naïve (NM) : pour chaque gène, la statistique de test naïve

$$T_N = \sqrt{2n}\frac{\overline{D}}{\sqrt{S^2}}$$

est calculée, avec $\overline{D} = \frac{1}{2n}\sum_i D_i$ et $S^2 = \frac{1}{2n-2}\sum_i (D_i - (\overline{D})_{(i)})^2$.

— Méthode non biaisée appariée (UP) : pour chaque gène, la statistique T_{UP}

$$T_{UP} = \sqrt{2n}\frac{\overline{D}}{\sqrt{(S^2 + 2C)}}$$

est calculée. D'un point de vue théorique la valeur de C doit être positive. Dans la pratique, la valeur estimée peut être négative. Dans de tels cas, C est tronqué à 0.

— Méthode non biaisée non appariée (UU) : pour chaque gène, la statistique T_{UU}

$$T_{UU} = \sqrt{n}\frac{\overline{Y_A} - \overline{Y_B}}{\sqrt{S^2_{Y_A} + S^2_{Y_B} - 2C_{Y_A Y_B}}}$$

est calculée. D'un point de vue théorique la valeur de $_{Y_A Y_B}$ doit être positive. Dans la pratique, la valeur estimée peut être négative. Dans de tels cas, $_{Y_A Y_B}$ est tronqué à 0. De plus l'estimateur sans biais de la variance est $S^2_{Y_A} + S^2_{Y_B} - 2C_{Y_A Y_B}$. En cas d'estimateur négatif, la variance est fixée à un seuil donné (ici 0.001).

— Estimation par maximum de vraisemblance dans un modèle mixte (ML) : pour chaque gène, le modèle (5.8) est ajusté avec un algorithme du maximum de vraisemblance.

— Estimation par maximum de vraisemblance restreint dans un modèle mixte (REML) : pour chaque gène, le modèle (5.8) est ajusté avec un algorithme du maximum de vraisemblance restreint.

Il est important de considérer à la fois l'algorithme ML et REML pour le modèle mixte, chacun ayant ses avantages. Tandis que le ML fournit des estimateurs biaisés des composants de la variance, les calculs sont plus rapides et les algorithmes convergent. REML donne des estimateurs non biaisés des paramètres mais peut ne pas converger si le nombre d'observations est faible. Les calculs ML et REML sont faits en utilisant le package R Maanova Kerr et al. [2000].

Ces cinq procédures ont été comparées à la fois sur des données réelles et simulées.

Résultats

Les termes aléatoires tenant compte des effets puce et échantillon doivent être inclus dans le modèle statistique au niveau gène pour les expériences en *dye-switch*. Nous avons montré sur des simulations que le test de Student apparié 'naïf' conduit à plus de faux positifs qu'attendus, surtout quand l'effet échantillon biologique est grand. Ce test peut

		$\mu = 1$			$\mu = 3$		
Nb d'échantillons	σ_B^2	UU	UP	REML	UU	UP	REML
5	0.5	5.6	13.6	10.6	55.5	92.1	86.75
5	2	2.8	5.0	12.95	17.9	29.4	34.75
10	0.5	13.2	39.3	33.97	77.8	100.0	99.64
10	2	3.5	7.8	9.06	45.0	63.5	63.06
20	0.5	35.0	80.1	78.13	98.8	100.0	100.0
20	2	7.3	14.5	13.93	82.6	94.8	94.53
30	0.5	51.9	95.5	95.05	100.0	100.0	100.0
30	2	12.1	22.5	21.74	96.2	99.6	99.53

TABLE 5.7 – Puissance (probabilité de rejeter H_0 $\times 100$) des différentes procédures de tests de détection d'une faible ($\mu = 1$, gauche) ou forte ($\mu = 3$, droite) expression différentielle.

n	UP CPU	REML CPU	No REML CV
5	2.3	787	56.9
10	2.6	212	5
20	2.8	467	0
30	3.2	1046	0.16

TABLE 5.8 – Temps CPU utilisateur des procédures (UP) et (REML), pour $\sigma^2 = 0.5$ et différentes tailles n d'échantillons. La dernière colonne fournit le nombre moyen de gènes pour lesquels la procédure REML ne converge pas.

être utilisé de façon fiable quand la variance biologique est plus faible que la variance technique. L'estimateur REML pour le modèle mixte fournit quand à lui un nombre correct de fausses découvertes au prix d'une forte complexité de calcul. Il lui arrive de ne pas converger pour des tailles d'échantillons faibles ou moyennes et parfois il donne des résultats faux. Au contraire, la méthode *UP* que nous proposons est facile à implémenter et ne nécessite pas de fortes ressources computationnelles. La méthode conduit à des résultats plus robustes et une analyse plus puissante que la méthode REML quand la variabilité biologique est grande et le nombre d'échantillons faibles, ce qui est couramment le cas dans les expériences de puces à ADN.
Pour les expériences avec peu de répétitions biologiques, il est conseillé d'utiliser des méthodes régularisées [Delmar et al., 2005; Kerr and Churchill, 2001; Smyth, 2004]. Ces stratégies sont basées sur des méthodes statistiques qui prennent en entrée la variance individuelle de chaque gène et donne une version régularisée de la variance pour chaque gène. La procédure *UP* donne une estimation pour la variance de l'expression différentielle pour chaque gène et peut donc être couplée à des méthodes régularisées.

5.3 Références

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(R106) :R106, 2010. 98

B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinf.*, 19 : 185–193, 2003. 98

J. Bullard, E. Purdom, K. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinf.*, 11 (94), 2010. 97

P. Delmar, S. Robin, and J. Daudin. Varmixt : efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, 21(4) :502–508, 2005. doi : 10.1093/bioinformatics/bti023. 91, 110

M.-A. Dillies, A. Rau, **Aubert, Julie**, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, and F. Jaffrezic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 2012. 87, 97, 98

E. Eisenberg and E. Levanon. Human housekeeping genes are compact. *Trends Genet.*, 19 (7) :362–365, Jul 2003. 101

T. Forster, Y. Costa, D. Roy, H. Cooke, and K. Maratou. Triple-target microarray experiments : a novel experimental strategy. *BMC Genomics*, 5(1) :13, 2004. ISSN 1471-2164. doi : 10.1186/1471-2164-5-13. URL http://www.biomedcentral.com/1471-2164/5/13. 93

P. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981. 94

M. Jeanmougin, A. de Reynies, L. Marisa, C. Paccard, G. Nuel, and M. Guedj. Should we abandon the t-test in the analysis of gene expression microarray data : A comparison of variance modeling strategies. *PLoS ONE*, 5(e12336), 2010. 101

R. Kelley, H. Feizi, and T. Ideker. Correcting for gene-specific dye bias in dna microarrays using the method of maximum likelihood. *Bioinformatics*, 24(1) :71–77, 2008. doi : 10.1093/bioinformatics/btm347. URL http://bioinformatics.oxfordjournals.org/content/24/1/71.abstract. 92

K. Kerr and G. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet Res*, 77 :123–128, 2001. doi : 10.1017/S0016672301005055. 110

K. Kerr, M. Martin, and G. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6) :819–837, 2000. doi : 10.1017/S0016672301005055. 109

M. K. Kerr, C. A. Afshari, L. Bennett, P. Bushel, J. Martinez, N. J. Walker, and G. A. Churchill. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12(1) :203–218, 2002. 89, 95

T. Margaritis, P. Lijnzaad, D. van Leenen, D. Bouwmeester, P. Kemmeren, S. van Hooff, and F. Holstege. Adaptable gene-specific dye bias correction for two-channel dna microarrays. *Mol Syst Biol.*, 5(266), 2009. 92

M.-L. Martin-Magniette, **Aubert, J.**, E. Cabannes, and J.-J. Daudin. Evaluation of the gene-specific dye bias in cdna microarray experiments. *Bioinformatics*, 21(9) :1995–2000, 2005a. 87, 90

M.-L. Martin-Magniette, **Aubert, J.**, E. Cabannes, and J.-J. Daudin. Answer to the comments of k. dobbin, j. shih and r. simon on the paper 'evaluation of the gene-specific dye-bias in cdna microarray experiments'. *Bioinformatics*, 21(14) :3065–3065, 2005b. 87, 90

M.-L. Martin-Magniette, **Aubert, J.**, A. Bar-Hen, S. Elftieh, F. Magniette, J.-P. Renou, and J.-J. Daudin. Normalization for triple-target microarray experiments. *BMC Bioinformatics*, 9(216), 2008. 87, 94

T. Mary-Huard, **Aubert, J.**, N. Mansouri-Attia, O. Sandra, and J.-J. Daudin. Statistical methodology for the analysis of dye-switch microarray experiments. *BMC Bioinformatics*, 9(98), 2008. 87, 107, 108

A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, 5 :621–628, 2008. 98

A. Oshlack, M. Robinson, and M. Young. From rna-seq reads to differential expression results. *Genome Biology*, 11(220), 2010. 98

M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(R25), 2010. 98

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 :1–25, 2004. doi : 10.2202/1544-6115.1027. 110

A. Su, T. Wiltshire, S. Batalov, H. Lapp, K. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. Cooke, J. Walker, and J. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101(16) :6062–6067, Apr 2004. doi : 10.1073/pnas.0400782101. URL http://dx.doi.org/10.1073/pnas.0400782101. 101

Z. Wang, M. Gerstein, and M. Snyder. Rna-seq : a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10 :57–63, 2009. 96

Y. Yang and N. Thorne. *Science and Statistics : A Festschrift for Terry Speed*, volume 40, chapter Normalization for two-color cDNA microarray data, pages 403–418. IMS Lecture Notes - Monograph Series, 2003. 98

Y. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cdna microarray data : a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30 :e15, 2002. doi : 10.1093/nar/30.4.e15. 89

Chapitre 6

Apports statistiques aux collaborations avec des biologistes

Sommaire

De par mon positionnement, j'ai participé à de nombreux projets de génomique et notamment analysé un grand nombre d'expériences de transcriptome avec des puces à ADN [Abdelkarim et al., 2011; Ait Yahya-Graison et al., 2007; Chaouat et al., 2011; Dalmasso et al., 2012a,b; Faucon et al., 2009; Forquin et al., 2011; Frey et al., 2007; Hébert et al., 2011, 2012; Lédée et al., 2009, 2011; Mansouri-Attia et al., 2009a,b; Vilg et al., 2014]. L'objectif de ce chapitre est de présenter à travers deux exemples de collaboration la façon dont les compétences statistiques sont mobilisées et la plus-value apportée par les statistiques aux projets de génomique. Le premier exemple concerne la planification expérimentale, la normalisation et la recherche de gènes différentiellement exprimés à partir de données issues de puces à ADN. Le deuxième concerne la recherche de gène différentiellement exprimés dans le cadre d'analyse du traductome chez l'oursin à partir de données de séquençage d'ARN.

6.1 Exemple de planification d'expériences de puces à ADN deux couleurs

Le syndrome de Down (SD) causé par une trisomie 21 est une des causes génétiques les plus courantes de retard mental. Les variations de l'expression de gènes du chromosome 21 ont été étudiées à partir de cellules lymphoblastoïdes de patients contrôles et de patients porteurs de la trisomie à l'aide d'une puce à ADN dédiée. Ce travail a donné lieu à une publication Ait Yahya-Graison et al. [2007]. Nous présenterons dans cette section la démarche qui a conduit à la définition du plan d'expériences puis les analyses qui ont été faites. Les notations utilisées sont similaires à celles utilisées dans le chapitre précédent.

6.1.1 Planification expérimentale

Objectifs et contraintes

L'objectif principal des expériences effectuées à l'aide de puces à deux couleurs était de détecter les gènes différentiellement exprimés entre des patients atteints de trisomie 21 (TS21) et des patients témoins tout en prenant en compte la variabilité existante entre les individus. Un objectif secondaire était de détecter les gènes affectés par un effet du sexe. L'étape de planification expérimentale consistait donc à proposer un plan adapté à ces objectifs et prenant en compte les contraintes expérimentales.

Les contraintes expérimentales étaient de deux types : matérielles et budgétaires. Nous disposions de 10 échantillons TS21 (7 hommes et 3 femmes), 11 échantillons contrôles (4 hommes et 11 femmes) et devions utiliser un maximum de 45 puces.

Modélisation dans un canal

Pour chaque gène, le modèle linéaire mixte suivant a été utilisé :

$$y_{ijkln} = \mu + V_k + S_l + A_i + D_j + VS_{kl} + VD_{kj} + SD_{lj} + I(VS)_{kln} + \epsilon_{ijkln}$$

où y_{ijkln} est l'expression normalisée du gène en $\log_2$ pour la maladie de type k (TS21 ou témoin), le sexe l (homme ou femme), l'individu n ($n = 1, \ldots, 21$), marqué avec le fluorochrome j (rouge pour Cy5 ou vert pour Cy3) sur la puce i. Les symboles V, S, A, D, I(VS) représentent respectivement les effets fixes dus à la maladie, au sexe, à la puce, au fluorochrome et l'effet aléatoire individu-emboité-dans-la-maladie et sexe. Nous supposons que les $I(VS)_{kln}$ sont indépendants et suivent une loi $\mathcal{N}(0, s_g^2)$, avec l'indice g pour le gène. Les ϵ_{ijkln} sont supposés indépendants et suivent une loi $\mathcal{N}(0, \sigma_g^2)$.

Le modèle peut se réécrire sous la forme matricielle suivante :

$$Y = X\theta + ZU + \epsilon$$

où θ est le vecteur des effets fixes (V, S, A, D, VS, VD, SD), U est le vecteur des $I(VS)_{ijk}$ et ϵ le vecteur des ϵ_{ijkln}. $Y \sim \mathcal{N}(X\theta, \Sigma)$ où $\Sigma = 2\sigma_g^2 Id + s_g^2 ZZ^T$. Y comprend N (nombre total d'individus) lignes, X est la matrice décrivant l'individu (quelle maladie, quel sexe) et Z dans ce cas particulier est égal à la matrice identité.

Modélisation deux canaux

Nous nous intéressons à la différence d'expression sur une puce (signal marqué en rouge − signal marqué en vert) pour un gène

$$\begin{aligned} y_{ijj'kk'll'mm'} &= y_{ijklm} - y_{ij'k'l'm'} \\ &= (V_k - V_{k'}) + (S_l - S_{l'}) + (VS_{kl} - VS_{k'l'}) \\ &+ (D_j - D_{j'}) + (VD_{kj} - VD_{k'j'}) + (SD_{lj} - SD_{l'j'}) \\ &+ (I(VS)_{ksn} - I(VS)_{k's'n'}) + (\epsilon_{ijklm} - \epsilon_{ij'k'l'm'}). \end{aligned}$$

Notons Δ la matrice décrivant la comparison faite sur chaque puce. Cette matrice comporte I (nombre de lames) lignes et N (nombre d'individus) colonnes. La $i^{\text{ème}}$ ligne de Δ comporte des zéros partout sauf pour deux colonnes pour lesquelles les valeurs sont soit 1 si l'individu est marqué en rouge, soit -1 si il est marqué en vert. Le modèle s'écrit sous la forme matricielle suivante :

$$\Delta Y = \Delta X\theta + \Delta ZU + \Delta\epsilon$$

$\Delta Y \sim \mathcal{N}(\Delta X\theta, \Sigma_A)$ où $\Sigma_A = 2\sigma_g^2 \Delta\Delta^T + s_g^2 \Delta ZZ^T \Delta^T$.

Comment choisir Δ ?

Les matrices X et Z sont données. Choisir un plan d'expérience revient donc à choisir une matrice Δ.
Les effets techniques tels que l'effet puce, l'effet fluorochrome ou les interactions du fluorochrome avec les autres effets ne nous intéressent pas.
L'expression différentielle entre deux individus sur une puce élimine l'effet puce et la constante. Si l'on souhaite éliminer les effets VD_{kj} et SD_{sj}, il faut proposer un plan équilibré pour l'effet fluorochrome. Nous pouvons omettre D_j car les données seront normalisées pour corriger de l'effet fluorochrome.

Le plan doit de plus respecter les contraintes expérimentales et objectifs présentés précédemment.

Nous choisirons un plan
— qui a toutes les propriétés précédentes,
— qui minimise la variance des effets maladie et garantie une variance pas trop grande pour les autres effets fixes selon un ratio σ^2/s^2 donné a priori, ie qui minimise le déterminant de la matrice de variance covariance.

Selon la théorie du modèle linéaire, la matrice de variance pour les effets fixes est égale à :

$$V(\hat{\theta}_A) = (\Delta^T X^T (\Delta \Sigma \Delta^T)^{-1} \Delta X)^{-1}$$

Nous calculons la variance de l'effet maladie et le déterminant de la matrice de variance covariance pour un certain nombre de plans respectant les propriétés et selon un ratio σ^2/s^2 donné a priori (ici = 2).
Un plan en boucle prenant en compte la variabilité biologique a été proposé. Il comprend 40 puces. Chaque patient apparaît dans 2 à 8 expériences (puces), et les échantillons d'un même patient sont marqués autant de fois avec chaque fluorochrome (Cy5 ou Cy3). Sur chaque lame, un patient contrôle est comparé à un patient SD afin d'avoir une meilleure puissance pour détecter un effet maladie. 10 lames comparent un homme atteint d'un syndrome de Down (TS21) avec un homme contrôle, 10 lames une femme atteinte d'un syndrome de Down à une femme contrôle, 10 lames un homme TS21 à une femme contrôle, 10 lames une femme TS21 à un homme contrôle. Le plan d'expérience est décrit dans la table 6.1.

6.1.2 Analyse différentielle

Afin de trouver les gènes différentiellement exprimés pour les effets Maladie, Sexe et Maladie*Sexe, une analyse de variance en adéquation avec le plan d'expérience a été ef-

	Hommes avec une TS21							Femmes avec une TS21		
Contrôles	1	2	3	4	5	6	7	8	9	10
Hommes										
11	-1	1		-1		1			-1	1
12				1	-1			1		-1
13		-1	1			-1		1	1	-1
14					1			-1	-1	1
Femmes										
15		1						-1		
16					-1				1	
17					1		-1	1		-1
18			-1			1		-1	1	
19	1								-1	
20		-1							1	
21						-1	1		-1	1

TABLE 6.1 – Expériences effectuées. Chaque '1' indique une puce. '+1' indique que les échantillons TS21 et contrôles ont respectivement été marqués en Cy5 et Cy3. −1' indique que les échantillons TS21 et contrôles ont respectivement été marqués en Cy3 et Cy5

fectuée. Le modèle mixte permet de distinguer la variabilité patient ou individuelle de la variabilité due au bruit technique ou expérimental. La procédure Mixed du logiciel SAS® avec la méthode REML a été utilisée.

Après les étapes d'analyse d'image, de filtre et normalisation, le nombre d'observations par spot varie de 8 à 40 et est suffisante pour calculer une variance par gène.

Le premier modèle testé est le modèle complet. Comme aucun gène n'est significatif pour l'interaction maladie*sexe ni pour l'effet sexe, ces effets ont été enlevés du modèle. Nous analysons finalement un modèle simplifié

$$\begin{aligned} y_{ijj'kk'll'mm'} &= y_{ijklm} - y_{ij'k'l'm'} \\ &= (V_k - V_{k'}) + (I(VS)_{ksn} - I(VS)_{k's'n'}) + (\epsilon_{ijklm} - \epsilon_{ij'k'l'm'}). \end{aligned}$$

Les probabilités critiques brutes ont été ajustées avec la procédure de Benjamini and Hochberg [1995] qui contrôle le FDR. Les gènes avec une probabilité critique ajustée inférieure à 5% ont été déclarés comme étant différentiellement exprimés de façon significative entre la condition trisomique et la condition contrôle. Une analyse par rapport à un seuil à 1.5 correspondant au ratio du nombre de copies chez les patients trisomiques par rapport aux contrôles a également été faite.

6.1.3 Résultats

Les analyses ont permis de détecter des gènes différentiellement exprimés entre les contrôles et les patients atteints du Syndrome de Down. Des variations différentes de $\frac{3}{2}$

(ratio du nombre de copies TS21 par rapport aux contrôles). Les résultats ont été validés par PCR quantitative. 29% des transcrits exprimés du chromosome 21 sont surexprimés chez les patients atteints du syndrome de Down et correspondent à des gènes ou à des cadres ouverts de lectures. Parmi ces derniers, 22% augmentent proportionnellement à l'effet gène-dosage (dû aux nombres de copies différentes) et 7% sont amplifiés. Les autres sont soit compensés, soit très variables entre les individus. La plupart des transcrits du chromosome 21 sont compensés pour l'effet gène-dosage. Les gènes surexprimés semblent impliqués dans le phénotype lié au syndrome de Down contrairement aux gènes compensés. Les gènes très variables pourraient tenir compte des variations phénotypiques observés chez les patients.

6.2 Etude du traductome chez l'oursin - Analyse de données à haut débit du polysome

Ce travail est issu d'une collaboration avec Julia Moralès et Hélène Chassé (UMR8225 CNRS-UPMC Laboratoire de Biologie Intégrative des Modèles Marins) pour la partie biologique et Erwan Corre et Gildas Le Corguillé pour la partie bioinformatique (plateforme ABiMS, CNRS-UPMC Station biologique de Roscoff). Plus de détails sur la question biologique, ainsi que sur la production des données ou l'interprétation des résultats sont disponibles en annexe I dans l'article en préparation.

6.2.1 Contexte

La fécondation chez l'oursin provoque la reprise de l'activité de traduction et des divisions cellulaires du développement précoce. De plus, la régulation de l'expression des gènes se fait au niveau post-transcriptionnel durant les premières étapes du développement, grâce à l'utilisation des ARN messagers maternels stockés dans l'oeuf. L'objectif de la thèse d'Héloïse Chassé [Chassé, 2015] visait à analyser le traductome en réponse à la fécondation afin de connaître :

1. les fonctions activées par la traduction via l'identification des ARNm différentiellement traduits,
2. les mécanismes de ce recrutement polysomal en recherchant des motifs communs enrichis dans les ARNm traduits.

Une seule méthode ou modélisation avait été proposée pour analyser ce type de données au moment de l'obtention des données de traductome. Il s'agissait de la méthode ANOTA [Larsson et al., 2011] implémentée dans le package Bioconductor tRanslatome

[Tebaldi et al., 2014]. Cette méthode ne pouvait pas s'appliquer directement sur les données présentées ici du fait d'une normalisation supplémentaire nécessaire par le nombre de gradients (cf. sous-section 6.2.3). De plus, les hypothèses du modèle (notamment de normalité des résidus) n'étaient pas vérifiées du fait de la nature des données issues du séquençage de l'ARN. Les collègues avaient tout de même essayé cette méthode avant de me contacter ; mais les résultats obtenus ne correspondaient pas du tout à leurs connaissances du phénomène.

6.2.2 Données

Afin de répondre à ces objectifs, 24 banques ont été faites (environ 30M de lectures par banque, paired-end 2x100pb). Les données ont été produites selon un plan factoriel complet comprenant 3 (nombre de femelles) ×2 (avec puromucine (puro) ou sans puromycine) ×2 (localisation de l'ARNm : cytoplasme ou polysome) ×2 (statut des oeufs de la femelle : fécondé (F) ou non fécondé (NF)) = 24 essais. La localisation des ARN messagers en fonction de l'état de la traduction est schématisée sur la figure 6.1. Pour appréhender le traductome (population des ARNm en cours de traduction) (voir sous-section 1.3.2), les ARNm associés aux polysomes et donc traduits activement ont été isolés sur gradient de sucrose, et ensuite séquencés. Bien que chaque condition soit réalisée à nombre de cellules constant, la quantité d'ARNm obtenue par gradient étant faible, il a parfois été nécessaire de pooler les ARNs provenant de plusieurs gradients pour atteindre la quantité d'ARN nécessaire pour le séquençage. Les lectures sont ensuite alignées sur un transcriptome do novo spécifiquement associé à l'étude en cours et annoté. Les niveaux d'ARNm dans les polysomes étant affectés par des mécanismes de transcription et post-transcription qui influent également sur les niveaux des ARNm dans le cytoplasme [Larsson et al., 2013], les ARNm dans le cytoplasme ont également été séquencés. Afin d'étudier les ARNm effectivement traduits, les échantillons ont été traités ou non avec de la puromycine. La puromycine est un antibiotique provoquant la terminaison prématurée de la traduction en imitant un ARN de transfert que le ribosome va attacher à l'extrémité d'un polypeptide naissant (1.3.2). Si un ARN est traduit, la fraction qui est dans les polysomes est sensible à la puromycine et l'ARN ne sera donc plus présent dans l'échantillon traité (polysome+puromycine).

6.2.3 Modélisation proposée

Soit Y_{ijg} le nombre de lectures alignées sur l'ARN messager g de la femelle $i (i = 1,2,3)$ dans les polysomes, dans la condition $j (j = \texttt{NFpuro}, \texttt{NF}, \texttt{Fpuro}, \texttt{F})$. On suppose que $Y_{ijg} \sim \text{NB}(\mu_{ijg}, \phi_g)$ avec la paramétrisation de la loi binomiale négative telle que présentée dans la section 2.3, ϕ_g est un paramètre de dispersion propre à chaque ARN messager ou gène.

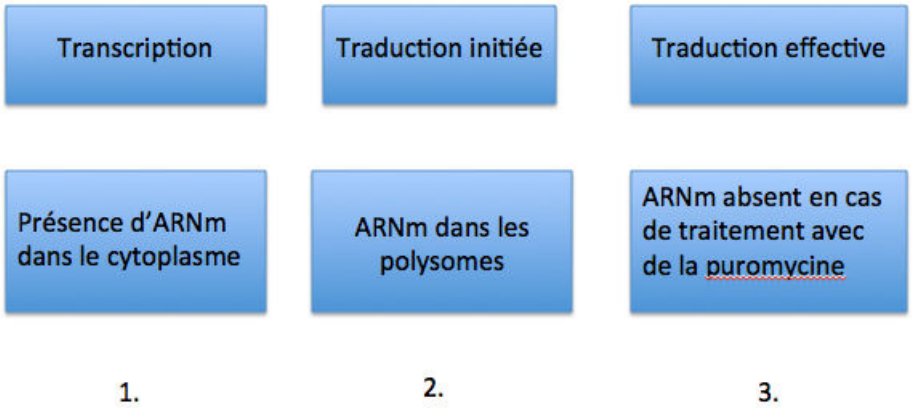


FIGURE 6.1 – Localisation des ARN messagers en fonction des états de transcription/traduction.

On pose le modèle suivant :

$$\log(\mu_{ijg}) = \kappa_{ijg} + (\mathrm{F})_i + (\mathrm{G})_j + \epsilon_{ijg}$$

où

- $(\mathrm{G})_j$ est l'effet principal de la condition j ;
- $(\mathrm{F})_i$ est l'effet principal de la femelle i ;
- ϵ_{ijg} est un terme d'erreur qui est supposé indépendant entre les observations.

La particularité du modèle repose dans la spécificité du terme d'offset défini de la façon suivante :

$$\kappa_{ijg} = \log(s_{ij}\nu_{ij}\pi_{ijg})$$

avec

- s_{ij} est un facteur de normalisation pour la taille de la banque de l'échantillon (ou profondeur de séquençage) correspondant aux ARNm dans les polysomes pour la femelle i dans la condition j ;
- ν_{ij} est le nombre de gradients de sucrose préalablement poolés lors de la préparation de l'échantillon d'ARN dans les polysomes de l'échantillon ij ;
- π_{ijg} est la fraction originaire de l'ARNm g associée aux cytoplasmes parmi toutes les fractions d'ADNc de l'échantillon correspondant à la femelle i dans la condition j.

En effet, les niveaux d'ARNm dans les polysomes sont affectés par des mécanismes de transcription et post-transcription qui influent également sur les niveaux des ARNm dans le cytoplasme [Larsson et al., 2013]. π_{ijg} n'est pas directement disponible. Nous l'estimons par $\frac{y^c_{ijg}}{s^c_{ij}\nu^c_{ij}}$, où s^c_{ij} correspond à un facteur normalisant pour la taille de la banque de l'échantillon d'ARN cytoplasmique de la femelle i dans la condition j et ν^c_{ij} est le nombre de gradients de sucrose qui ont été poolés dans l'échantillon ij d'ARN cytoplasmique.

Toutes les analyses statistiques ont été effectuées avec le logiciel R [R Core Team, 2016] en utilisant des packages Bioconductor [Gentleman et al., 2004] . Les facteurs de norma-

lisation s_{ij} et s^c_{ij} sont calculés par la méthode TMM (Trimmed Mean of M-values, Robinson and Oshlack [2010]). Nous utilisons ensuite le package Bioconductor edgeR [Robinson MD and GK, 2010] avec une matrice d'offsets pour ajuster un modèle binomial négatif par gène avec une dispersion gène-spécifique calculé comme dans McCarthy et al. [2012] et Chen et al. [2014]. La normalisation, comme mis en avant dans notre article Dillies et al. [2012], a un impact important sur toute la suite de l'analyse. La matrice d'offsets est supposée tenir compte de toutes les normalisations nécessaires (ici, profondeur de séquençage, nombre de gradients, et valeurs basales dans le cytoplasme). Des tests de rapport de vraisemblance sont ensuite effectués pour :

— comparer les ARNs affectés par la fécondation (F vs NF) ;
— comparer les ARNs effectivement traduits pour les fécondés (F vs Fpuro) ;
— comparer les ARNs effectivement traduits et enrichis dans les fécondés versus les non fécondés (Fpuro vs NFpuro). Cela nous donne les ARNs recrutés (ou dé-recrutés) à la fécondation.

Les probabilités critiques brutes ainsi obtenues sont ajustées pour les tests multiples par la procédure de Benjamini and Hochberg [1995] qui contrôle le taux de fausses découvertes. Les ARN messagers avec une probabilité critique ajustée inférieure à 5% sont considérés comme significatifs.

6.2.4 Résultats et discussion

Les résultats obtenus sont détaillés dans l'article en préparation disponible en annexe I. Ils ont été en partie validés par des analyses en RT-PCR quantitatives. Ces analyses ont permis de montrer [Chassé, 2015] que les ARNm maternels ne sont pas uniformément recrutés dans les polysomes et traduits, mais qu'il existe une importante sélectivité, contrairement à ce qui est classiquement admis. Les catégories de transcrits préférentiellement recrutés dans les polysomes à la fécondation sont celles du 'cycle cellulaire', de la 'signalisation', et des 'protéines de liaison à l'ARN', ou alors des transcrits codant des protéines régulatrices.

6.3 Références

M. Abdelkarim, N. Vintonenko, A. Starzec, A. Robles, **Aubert, Julie**, M.-L. Martin, S. Mourah, M.-P. Podgorniak, S. Rodrigues-Ferreira, C. Nahmias, P.-O. Couraud, C. Doliger, O. Sainte-Catherine, N. Peyri, L. Chen, J. Mariau, M. Etienne, G.-Y. Perret, M. Crepin, J.-L. Poyet, A.-M. Khatib, and M. Di Benedetto. Invading basement membrane matrix is sufficient for mda-mb-231 breast cancer cells to develop a stable in vivo metastatic phenotype. *PLoS ONE*, 6(8), 08 2011. 115

E. Ait Yahya-Graison, **Aubert, J**, L. Dauphinot, I. Rivals, M. Prieur, G. Golfier, J. Rossier, L. Personnaz, N. Créau, H. Bléhaut, S. Robin, J. Delabar, and M. Potier. Classification of human chromosome 21 gene-expression variations in down syndrome : Impact on disease phenotypes. *American Journal of Human Genetics*, 81(3) :475–491, 2007. 115

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *JRSSB*, 57,1 :289–300, 1995. 118, 122

G. Chaouat, N. Rodde, M. Petitbarat, R. Bulla, M. Rahmati, S. Dubanchet, S. Zourbas, I. Bataillon, N. Coque, B. Hennuy, J. Martal, C. Munaut, **Aubert, J.**, V. Serazin, T. Steffen, J. Jensenius, J. Foidart, O. Sandra, F. Tedesco, and N. Ledee. An insight into normal and pathological pregnancies using large-scale microarrays : lessons from microarrays. *Journal of Reproductive Immunology*, 89(2) :163–72, may 2011. 115

H. Chassé. *Regulations traductionnelles de l'embryon precoce d'oursin*. PhD thesis, Ecole Doctorale 515 - Complexite du Vivant, 2015. 119, 122

Y. Chen, A. T. L. Lun, and G. K. Smyth. Differential expression analysis of complex rna-seq experiments using edger, 2014. 122

M. Dalmasso, **Aubert, J.**, V. Briard-Bion, V. Chuat, S.-M. Deutsch, S. Even, H. Falentin, G. Jan, J. Jardin, M.-B. Maillard, M. Parayre, M. Piot, J. Tanskanen, and A. Thierry. A temporal -omic study of propionibacterium freudenreichii cirm-bia1t adaptation strategies in conditions mimicking cheese ripening in the cold. *PLoS ONE*, 7(1), 2012a. 115

M. Dalmasso, **Aubert, J.**, S. Even, H. Falentin, M.-B. Maillard, M. Parayre, V. Loux, J. Tanskanen, and A. Thierry. Propionibacterium freudenreichii accumulates intracellular glycogen and trehalose in conditions mimicking cheese ripening in the cold. *Applied and Environmental Microbiology*, 2012b. 115

M.-A. Dillies, A. Rau, **Aubert, Julie**, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, and F. Jaffrezic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 2012. 122

F. Faucon, E. Rebours, C. Bevilacqua, J.-C. Helbling, **Julie Aubert**, S. Makhzami, S. Dhorne-Pollet, S. Robin, and P. Martin. Terminal differentiation of goat mammary tissue during pregnancy requires the expression of genes involved in immune functions. *Physiological Genomics*, 40 :61–82, 2009. 115

M. Forquin, A. Hébert, A. Roux, **Aubert, J.**, C. Proux, J. Heilier, S. Landaud, C. Junot, P. Bonnarme, and I. Martin-Verstraete. Global regulation in response to sulfur availability in the cheese-related bacterium, brevibacterium aurantiacum. *Applied and Environmental Microbiology*, 2011. 115

I. Frey, I. Rubio-Aliaga, A. Siewert, D. Sailer, A. Drobyshev, J. Beckers, M. Hrabe de Angelis, **Aubert, J**, A. Bar-Hen, O. Fiehn, H. Eichinger, and H. Daniel. Profiling at mrna, protein and metabolite level reveals alterations in renal amino acid handling and glutathione metabolism in kidney tissue of pept2 mice. *Physiological Genomics*, 28(3) :301–310, 2007. 115

R. Gentleman, V. Carey, D. 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. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and Z. J. Bioconductor : open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), 2004. doi : 10.1186/gb-2004-5-10-r80. 121

A. Hébert, M.-P. Forquin-Gomez, A. Roux, **Aubert, J.**, C. Junot, V. Loux, J.-F. Heilier, P. Bonnarme, J.-M. Beckerich, and S. Landaud. Exploration of sulfur metabolism in the yeast kluyveromyces lactis. *Applied Microbiology and Biotechnology*, 91 :1409–1423, 2011. 115

A. Hébert, M.-P. Forquin-Gomez, A. Roux, **Aubert, J.**, C. Junot, J.-F. Heilier, S. Landaud, P. Bonnarme, and J.-M. Beckerich. Study of yarrowia lipolytica reveals new insights into sulfur metabolism in yeasts. *Applied and Environmental Microbiology*, 2012. 115

O. Larsson, N. Sonenberg, and R. Nadon. anota : Analysis of differential translation in genome-wide studies. *Bioinformatics*, 27, 2011. doi : doi:10.1093/bioinformatics/btr146. 119

O. Larsson, B. Tian, and N. Sonenberg. Toward a genome-wide landscape of translational control. *Cold Spring Harbor Perspectives in Biology*, 5, 2013. doi : doi:10.1101/cshperspect.a012302. 120, 121

N. Lédée, **Aubert, J.**, B. Hennuy, C. Munaut, V. Serazin, S. Petitbarat, S. Dubanchet, G. Chaouat, and O. Sandra. The preconceptional endometrial environment affects both implantation and gestation in human. *Journal of Reproductive Immunology*, page 119, 2009. 115

N. Lédée, C. Munaut, **Aubert, J.**, V. Serazin, M. Rahmati, G. Chaouat, O. Sandra, and J.-M. Foidart. Specific and extensive endometrial deregulation is present before conception in ivf/icsi repeated implantation failures (if) or recurrent miscarriages. *The Journal of Pathology*, 2011. 115

N. Mansouri-Attia, O. Sandra, **Julie Aubert**, S. Degrelle, R. E. Everts, C. Giraud-Delville, Y. Heyman, L. Galio, I. Hue, X. Yang, X. C. Tian, H. A. Lewin, , and J.-P. Renard. Endometrium as an early sensor of in vitro embryo manipulation technologies. *PNAS*, 2009a. 115

N. Mansouri-Attia, **Julie Aubert**, P. Teinaud, C. Giraud-Delville, G. Tagouhti, L. Galio, R. E. Everts, S. Degrelle, C. Richard, I. Hue, X. Yang, C. Tian, H. A. Lewin, J.-P. Renard, and O. Sandra. Gene expression profiles of bovine caruncular and intercaruncular endometrium at implantation. *Physiological Genomics*, 39 :14–27, 2009b. 115

D. McCarthy, Y. Chen, and G. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40 :4288–4297, 2012. 122

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. 121

M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(R25), 2010. 122

M. D. Robinson MD and S. GK. edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26 :139–140, 2010. 122

T. Tebaldi, E. Dassi, G. Kostoska, G. Viero, and A. Quattrone. translatome : an r/bioconductor package to portray translational control. *Bioinformatics*, 30(2) :289–291, 2014. doi : 10.1093/bioinformatics/btt634. 120

J. V. Vilg, N. V. Kumar, E. Maciaszczyk-Dziubinska, E. Sloma, D. Onesime, **Aubert, Julie**, M. Mogicka, R. Wysocki, and M. J. Tamas. Elucidating the response of kluyveromyces lactis to arsenite and peroxide stress and the role of the transcription factor klyap8. *BBA - Gene Regulatory Mechanisms*, 2014. 115

Annexe A
Article sur le fdr local présenté dans la section 3.1

BMC Bioinformatics

Research article

Determination of the differentially expressed genes in microarray experiments using local FDR

J Aubert, A Bar-Hen, J-J Daudin* and S Robin

Address: UMR INAPG/INRA/ENGREF 518, 16, rue C. Bernard, 75231 Paris Cedex 05, France

Email: J Aubert - aubert@inapg.fr; A Bar-Hen - avner@inapg.fr; J-J Daudin* - daudin@inapg.fr; S Robin - robin@inapg.fr

* Corresponding author

Published: 06 September 2004

BMC Bioinformatics 2004, **5**:125 doi:10.1186/1471-2105-5-125

Received: 27 May 2004

Accepted: 06 September 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/125

Abstract

Background: Thousands of genes in a genomewide data set are tested against some null hypothesis, for detecting differentially expressed genes in microarray experiments. The expected proportion of false positive genes in a set of genes, called the False Discovery Rate (FDR), has been proposed to measure the statistical significance of this set. Various procedures exist for controlling the FDR. However the threshold (generally 5%) is arbitrary and a specific measure associated with each gene would be worthwhile.

Results: Using process intensity estimation methods, we define and give estimates of the local FDR, which may be considered as the probability for a gene to be a false positive. After a global assessment rule controlling the false positive error, the local FDR is a valuable guideline for deciding wether a gene is differentially expressed. The interest of the method is illustrated on three well known data sets. A R routine for computing local FDR estimates from *p*-values is available at http://www.inapg.fr/ens_rech/mathinfo/recherche/mathematique/outil.html.

Conclusions: The local FDR associated with each gene measures the probability that it is a false positive. It gives the opportunity to compute the FDR of any given group of clones (of the same gene) or genes pertaining to the same regulation network or the same chromosomic region.

Background

Microarrays are part of a new class of biotechnologies that allow the monitoring of the expression level of thousands of genes simultaneously. Among the applications of microarrays, an important task is the identification of differentially expressed genes, i.e genes whose expressions are associated with the status of the patient (treatment/control for example).

The biological question of the identification of differentially expressed genes can be restated as a one (for paired data) or two-sample (for unpaired data) hypothesis testing procedure: is the gene differentially expressed between the two situations? However, when thousands of genes in a microarray data set are evaluated simultaneously by fold changes or significance tests approach, multiple testing problems immediately arise and lead to many false positive genes. In this 'one-by-one gene' approach the probability of detecting false positives rises sharply.

The False Discovery Rate (FDR), is defined as the expected fraction of false rejections among those hypotheses rejected. In their seminal paper Benjamini & Hochberg [1] provided a distribution free procedure (BH) for choosing a threshold on *p*-values that guarantees that the FDR is less than a target level α. The same paper demonstrated that

the BH procedure is more powerful than the Bonferroni method that controls the familywise error rate.

The FDR gives an idea of the expected number of false positive hypotheses that a practitioner can expect if the experiment is done an infinite number of time. As usual with expectation, it gives very little information about the number of false discovery hypotheses in a given experiment.

Motivation

The value of 1, 5 or 10% for the FDR, which determines the threshold t, is arbitrary. Storey and Tibshirani [2] stressed the importance of assessing to each feature its own measure of significance. They proposed to use the q-value,

$$\frac{\hat{m}_0 P_i}{R_i},$$

where P_i is the p-value of the ordered gene i, R_i is the total number of rejected genes whose p-values are less than the threshold $t = P_i$ and $\widehat{m_0}$ is an estimate of the total number of non differentially expressed genes, m_0.

The q-value is appealing because it gives a measure of significance that can be attached to each gene, but it must be stressed that it is not an estimate of the probability for the gene to be a false positive. The q-value is generally lower than the latter because it is computed using all the genes that are more significant than gene i. Obviously a gene whose p-value is near to the threshold t does not have the same probability to be differentially expressed than a gene whose p-value is close to zero. Therefore the q-value gives a too optimistic view of the probability for the gene to be a false positive.

Therefore it is interesting to obtain an estimate of the FDR attached to each gene, called local FDR, from an inferential point of view and without any assumption about the distribution of the p-values under H_1.

Results

Let

$H_0(i) = \{\text{gene } i \text{ is not differentially expressed}\}$.

Let the local FDR be the probability that a given gene is not differentially expressed. More specifically, $FDR(i)$ is the probability that a gene, whose p-value is P_i, is not differentially expressed, taking into account the whole set of tests. A raw local FDR estimate is defined in a first step. In a second step the local FDR estimate is defined as a smoothed value based on the raw values.

Let $P_1 < \ldots < P_m$ denote the ordered p-values for testing $H_0(i)$. The raw local FDR estimate for gene i is:

$$\widehat{\mathrm{FDR}}(i,\lambda) = \begin{cases} m_0(\lambda)(P_i - P_{i-1}) & \text{if } i > 1 \\ m_0(\lambda)P_1 & \text{if } i = 1 \end{cases}$$

where

$$\hat{m}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)}$$

where λ is a tuning parameter and $W(\lambda) = \#\{i, P_i > \lambda\}$, see Storey [3].

Assume that the p-values for the non-differentially expressed genes are independent. The raw local FDR estimate has the following properties:

• Under $H_0(i)$ and $H_0(i - 1)$ and if $E(\hat{m}_0) = m_0$, $\widehat{FDR}(i, \lambda)$ is unbiased with mean 1.

• Let $\widehat{FDR}(i, m_0) = m_0(P_i - P_{i-1})$. Under $H_0(i)$ and $H_0(i - 1)$ and if m_0 is known, $V(\widehat{FDR}(i, m_0)) = m_0^3/[(m_0 + 1)^2(m_0 + 2)] \approx 1$, for m_0 large enough. This value is a lower bound for $V(\widehat{FDR}(i, \lambda))$ when m_0 is unknown.

• The variance of the raw local FDR under H_1 is generally much smaller than under H_0.

• $\frac{1}{j}\sum_{i \le j} \widehat{\mathrm{FDR}}(i,\lambda) = q_j$ where q_j is the q-value of gene j.

The q-value may thus be viewed as the mean of the local FDR of the genes with p-values lower than P_j.

$\widehat{FDR}(i, \lambda)$ is generally a very variable estimator. Moreover the local FDR should increase with the p-value. This is not the case for the raw local FDR. Therefore it is necessary to use a smoothed estimate.

The smoothed local FDR(i) is

$$\widehat{\mathrm{FDR}}_s(i,\lambda) = f_i(\widehat{\mathrm{FDR}}(j,\lambda), j = 1, m)$$

where f_i is a smoothing function of the $\widehat{FDR}(j, \lambda)$ for $j = 1, m$, computed at position P_i.

$\widehat{\mathrm{FDR}}_s(i, \lambda)$ gives a very valuable guideline for the choice of a threshold. One may consider the curve of the local FDR versus the index of the gene ordered by their p-values: a good candidate for the threshold should be a point with

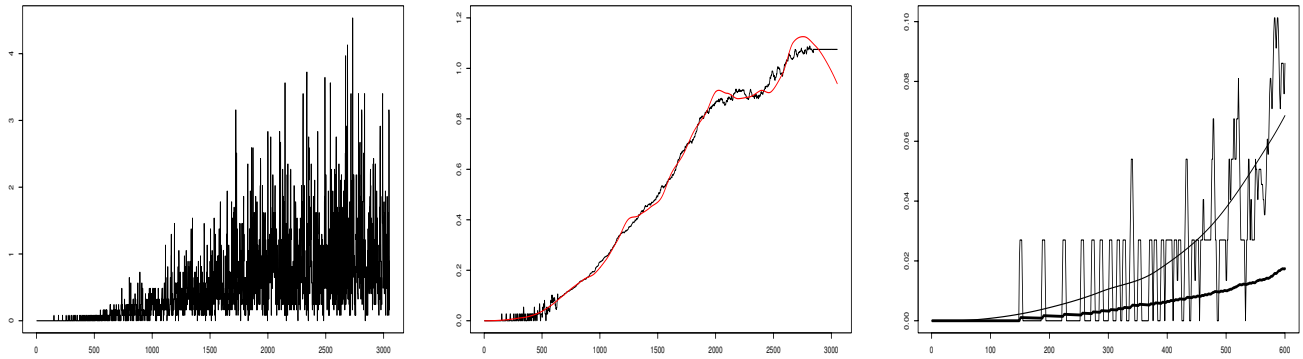


Figure 1
Plots of the local FDR estimate for Golub data *x*-axis: index of genes ordered along their *p*-values, *y*-axis: local FDR estimate. (a): raw values, (b): smooth estimates: moving average (discrete jumps), lowess (smooth curve), (c): zoom on the first 600 genes of (b): moving average (discrete jumps), lowess (upper smooth curve), *q*-value (lower thick smooth curve).

a high second order derivative, which corresponds to an abrupt change in the slope of the curve (see the examples of the following section). The second order derivative of the smoothed local FDR can be computed numerically using finite differences.

As an interesting application of the local FDR, it is possible to compute the FDR associated with a class of genes or clones by summing up the local FDR estimate of each clone or gene: one may consider for example clones corresponding to the same gene, genes known involved in a given regulatory network, or gene from the same chromosomic region, and associate a FDR with the whole class. These genes do not need to have consecutive *p*-values. The following sections demonstrate how the local FDR can be useful using the data of well known experiments.

Local FDR on Golub data set

Golub [4] were interested in identifying genes that are differentially expressed in patients with two types of leukemias (ALL, AML). Gene expression levels were measured using Affymetrix high-density chips containing 6817 human genes. The learning set comprises 27 ALL cases and 11 AML cases.

Data are available in the R multtest package. We used the preprocessing proposed by the authors and the *p*-values based on random permutations of the ALL/AML labels on Welch *t*-statistics for each gene, Dudoit [5], on the 3051 remaining genes. m_0 is estimated with bootstrap method as suggested by Storey and Tibshirani and implemented in the library GeneTS of software R.

Figure 1(a) presents the $\widehat{FDR}(i)$ for ordered genes and 1(b) presents the smooth curves obtained using lowess with a span of 0.2 and an adaptative moving average method.

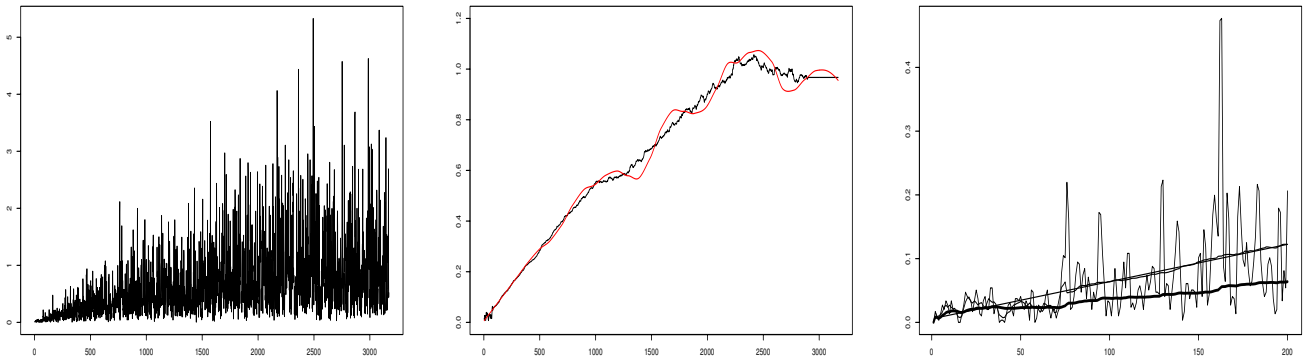
We can see that there is an abrupt change of the smoothed local FDR around gene number 500 which corresponds to a threshold $t = 0.15$ for the *p*-value. This may be an indication about the threshold. The Figure 1(c) presents a zoom of the Figure 1(b) for the first 600 *p*-values. We can see in Figure 1(c) that if we select the 438 (14%) top genes, we obtain a *q*-value equal to 0.0078 while the 438^{th} gene has a local FDR equal to 0.027. It must be noticed that there is a big difference between the two measures of FDR because the numerous regulated genes with very small *p*-values have a great influence on the *q*-value, which is not the case of the local FDR (see Figure 1(c)).

The *p*-values have been obtained using random permutations. Therefore the *p*-values are discrete with several genes possessing the same *p*-value. Therefore the values of $\widehat{FDR}(i, \lambda)$ may be equal to 0 because the difference between two successive *p*-values is 0. The discrete structure of the *p*-values implies a departure from the theoretical continuous uniform distribution. This explains why the moving average smoothing creates discrete jumps which appear in Figure 1(c).

If the distribution of the statistics under H_0 is correct, the *p*-values are distributed as a uniform distribution over [0, 1]. The empirical distribution of the high observed *p*-val-

Table 1: *p*-value, *q*-value and local FDR estimates for three genes in Hedenfalk data.

gene	*p*-value	rank	*q*-value	raw local FDR	smoothed local FDR
MSH2	0.00005	8	0.013	0.013	0.010
PDCD5	0.00048	47	0.022	0.013	0.033
CTGF	0.0036	159	0.049	0.176	0.098



(a) (b) (c)

Figure 2
Plots of the local FDR estimate for Hedenfalk data *x*-axis: index of genes ordered along their *p*-values, *y*-axis: local FDR estimate. (a): raw values, (b): smooth estimates: moving average (discrete jumps), lowess (smooth curve), (c): zoom on the first 200 genes of (b): raw values (discrete jumps), moving average and lowess (smooth curves), *q*-value (lower thick smooth curve).

ues (say above 0.5) is far from the uniform distribution. There are several non-exclusive possibilities to explain this: more than 50% of the genes are differentially expressed, the gene results for non-differentially expressed are correlated or there is a technical problem in the random permutations of the Welch *t*-statistics.

Local FDR on Breast Cancer data set

Storey and Tibshirani [2], have analysed in detail data from Hedenfalk [6] on 15 microarrays on breast cancer. Using the same *p*-values, we have computed local FDR estimates. The three genes which have been analysed in detail by Storey and Tibshirani [2] are presented in Table 1.

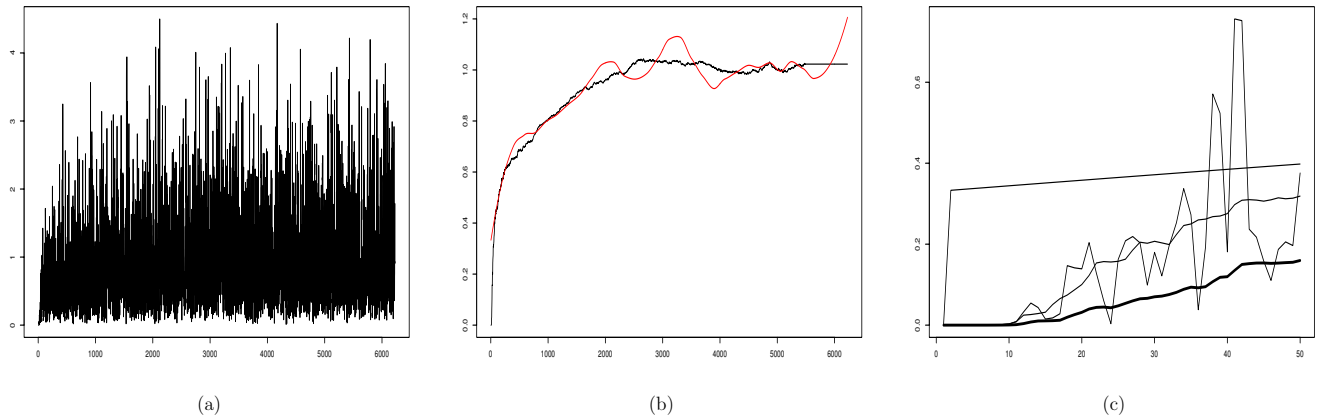
One can see that the smooth local FDR estimate is generally greater than the *q*-value and gives a better idea of the probability that a gene is a false positive. For example, at the level of 5%, CTGF will be considered as differentially expressed on the basis of the *q*-value while it will be considered as non differentially expressed using the local FDR.

Figure 2(a) presents the $\widehat{FDR}(i)$ for ordered genes and 2(b) presents the smooth curves obtained using lowess with a span of 0.2 and moving average methods. The two smoothing methods give similar results.

Setting $\lambda = 0.5$, Storey and Tibshirani [2] estimate that 67% of the 3170 genes in the data are not differentially expressed. The asymptote near 1 of the smooth curve supports this estimation.

Local FDR on ApoAi data

The goal of the study is to identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels compared to inbred control mice. The mouse model is the apolipoprotein AI (ApoAI) knock-out mice. ApoAI is a gene known to play a pivotal role in HDL metabolism. The statistical analysis is



(a) (b) (c)

Figure 3
Plots of the local FDR estimate for Apo-AI data *x*-axis: index of clones ordered along their *p*-values, *y*-axis: local FDR estimate. (a): raw values, (b): smooth estimates: moving average (small discrete jumps), lowess (smooth curve), (c): zoom on the 50 first genes of (b): raw values (discrete jumps), moving average (smooth curve) lowess (upper rectangular curve), *q*-value (lower thick smooth curve).

described in Dudoit [7]. Height clones are expected to be differentially expressed between the control and the knock-out mices because they are clones of the ApoAI gene or of genes coregulated with ApoAI. The height clones are actually the 8 top clones detected by the statistical tests. However there are other following clones which seem statistically significant if we consider the *q*-value. We can see on the Figure 3(c) that the local FDR values are much higher than the *q*-values.

Figure 3(a) presents the $\widehat{FDR}(i)$ for ordered clones and Figure 3(b) presents the smooth curves obtained using lowess with a span of 0.2 and moving average methods. The two smoothing methods give different results at the two ends of the [0, 1] interval. The moving average method which uses a special adaptative algorithm for the ends gives a better smoothing. This is particularly important for the clones with a small *p*-value for which it is crucial to obtain good estimates of the probability of being false positives. The lowess smoothing does not work well for the 50 first clones. In this particular case the default smoothing parameter $f = 0.2$ is not well suited and should be lower. However if it is chosen too low, the smoothing will not fit well the rest of the curve.

There are two clones of the gene Apo-AI. If we want to estimate the FDR of these two clones taken in a whole, we compute the mean of the smoothed local FDR of the two clones (the first and the height top clones) and obtain a local FDR for the gene Apo-AI, which is equal to $\frac{0 + 0.00048}{2} = 0.00024$. This example shows that it is possible to estimate the local FDR of any group of clones. This opportunity provided by the local FDR is certainly one of its major advantage with many potential applications.

Discussion

The curve of the smoothed local FDR is an efficient tool to summarize the information about the number and the statistical significance of differentially expressed genes, and may also be used to give an indication about the validity of the statistical assumptions. Moreover it is a valuable tool to choose the threshold for separating the differentially expressed genes from the non-differentially expressed one: one can choose a value of *t* maximizing the second derivative. Alternatively one can use a cost function and choose the threshold that minimizes the mean cost for a given cost function: using cost of the experiment, cost of false positive gene validation and the profit of discovering a differentially expressed gene, it is direct to compute the optimal strategy for choosing the threshold.

Note that a decision rule based on the local FDR would lead to a different set of selected genes than the usual one obtained by controlling the FDR. Consider the set of tests for which the local FDR is below 0.05, say. This set is not identical to the set identified by the standard criterion that $FDR < 0.05$. The local FDR is higher than the *q*-value. Therefore the first set is strictly included in the second

one. The local FDR rule is therefore more conservative than the usual FDR one.

Conclusions

The *p*-value gives the probability that a non differentially expressed gene would be as or more extreme than the gene under concern. The *q*-value indicates the estimated proportion of genes as or more extreme than the gene under concern that are a false positive. The local FDR gives the estimated proportion of genes around the gene under concern which are false positive. The latter may be used as the probability that the gene under concern is a false positive, taking into account the multiplicity of the test. One of the major interest of the local FDR is that it gives the opportunity to compute the FDR of any given group of clones (of the same gene) or genes pertaining to the same regulatory network or the same chromosome.

Methods

Model

Basically, the various procedures proposed in the literature aim to test the null hypothesis

$H_0(i) = \{\text{gene } i \text{ is not differentially expressed}\}$.

Let consider a particular experiment. We observed the differential expression of the genes and compute the associated ordered *p*-values P_i. In the following we will use the classical property: the *p*-values corresponding to non differentially expressed genes are uniformly distributed over [0, 1]. Furthermore, we will assume, as often, that these *p*-values are independent. However, the independence of the *p*-values of differentially expressed genes is not required. Consider a multiple testing situation in which m tests are being performed. Let m_0 be the number of non differentially expressed genes. Let $I(t)$ be the set of the genes having a *p*-value lower than t: $I(t) = \{i : P_i \le t\}$ and $R(t) = \#I(t)$, its cardinal. Let

$V(t) = \#[I(t) \cap (i \in H_0)]$

and

$S(t) = \#[I(t) \cap (i \in H_1)]$.

Using a threshold t, the m genes can be classified according to the following 2 × 2 table 2:

The Family Wise Error Rate (FWER) is defined to be

$\text{FWER} = P[V(t) \ge 1]$.

A classical way to control FWER is given by the Bonferroni inequality. This quantity corresponds to the most direct extension from a test hypothesis procedure but can be very restrictive in a multiple testing procedure.

The status of the gene associated with the P_i is an unobserved value. It is the same framework as point process (see for example [8]). In fact we observe $R(t) = V(t) + S(t)$ the sum of two counting processes. The first one $V(t)$ is a counting process associated with non differentially expressed gene. Since the *p*-values under H_0 are uniformly distributed, $V(t)$ has a binomial distribution with parameter m_0 and t. The intensity of $V(t)$ is constant and proportional to m_0. $S(t)$ is the counting process associated with gene under H_1 and very few can be said about its distribution. One may expect the intensity of $S(t)$ to be decreasing with t. The false discovery rate is defined as:

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

It corresponds to the expected proportion of rejections that are incorrect.

The BH procedure works as follows. Let $P_1 < \ldots < P_m$ denote the ordered *p*-values. Calculate $k = \max_i\{P_i \le \alpha i/m\}$. The procedure rejects all null hypotheses for which $P_i \le P_k$. If the tests are independent, this procedure ensures that

$$\text{FDR} \le \frac{m_0}{m}\alpha \le \alpha.$$

Let $\text{FDR}(t)$ be the FDR when rejecting all null hypotheses with $P_i \le t$. Because the *p*-values of non-differentially expressed genes are uniformly distributed over [0, 1], a natural estimate of $\text{FDR}(t)$ is

$$\widehat{\text{FDR}}(t) = \frac{m_0 t}{R(t)}.$$

Therefore the problem is to estimate m_0. Storey [3], proposed to estimate m_0 with

$$\hat{m}_0(\lambda) = \frac{W(\lambda)}{(1-\lambda)}$$

where λ is a tuning parameter. In particular the case λ = 0 leads to $\hat{m}_0 = m$. This is the most conservative case and corresponds to the BH procedure. Since the practical implementation of Storey method gives reasonably good results, we used it in the examples.

FDR is defined as the expectation of the ratio of two counting processes $V(t)$ and $R(t)$: $\text{FDR}(t) = E[V(t)/\max(R(t), 1)]$. The expectation of $V(t)$ is $m_0 t$ and $R(t)$ is observed. Therefore, Storey [3] propose to use the following estimate:

Table 2: Classification of m genes using threshold

	H_0 accepted	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	m_0
H_0 false	$T(t)$	$S(t)$	m_1
Total	$W(t)$	$R(t)$	m

$$\widehat{\mathrm{FDR}}(t,\lambda) = \frac{m_0(\lambda)t}{R(t)}.$$

The ratio of the expectations differs from the expectation ratio but Storey [3] proved that $E(\widehat{FDR}(t, \lambda)) \geq \mathrm{FDR}(t)$ using a convexity argument.

Definition and Estimation of the Local FDR

As stated before, $V(t)$ and $R(t)$ are counting (i.e. cumulative) processes. It would be very interesting to estimate the ratio of the local intensities of the two processes at point t. The intensity of process $V(t)$ is equal to m_0 and thus is known, provided that we know m_0. The intensity of process $R(t)$ is unknown, but $R(t)$ is observed. Therefore, using point process methods it is possible to estimate its intensity at each point t.

We first define the cumulative processes from t_1 to t_2:

Let $0 \leq t_1 < t_2$, $I(t_1, t_2) = \{i : t_1 < P_i \leq t_2\}$,

$R(t_1, t_2) = \#I(t_1, t_2)$,

$V(t_1, t_2) = \#[I(t_1, t_2) \cap (i \in H_0)]$

and

$S(t_1, t_2) = \#[I(t_1, t_2) \cap (i \in H_1)]$.

$\mathrm{FDR}(t_1, t_2)$ is defined as the expected ratio of $V(t_1, t_2)$ and $R(t_1, t_2)$:

$$\mathrm{FDR}(t_1,t_2) = E\left[\frac{V(t_1,t_2)}{max(R(t_1,t_2),1)}\right].$$

It is a generalization of the usual FDR: if $t_1 = 0$ and $t_2 = t$ then $\mathrm{FDR}(t_1, t_2) = \mathrm{FDR}(t)$. So, the natural estimate of $\mathrm{FDR}(t_1, t_2)$ is:

$$\widehat{\mathrm{FDR}}(t_1,t_2,\lambda) = \frac{m_0(\lambda)(t_2 - t_1)}{R(t_1,t_2)}.$$

The substitution of 0 by t_1 does not change the proof, so using the same convexity argument as Storey [3], we obtain the following property:

$E(\widehat{FDR}(t_1, t_2, \lambda)) \geq \mathrm{FDR}(t_1, t_2)$.

The local FDR is the $\mathrm{FDR}(t_1, t_2)$ for small intervals $[t_1, t_2]$. If we want to estimate the local FDR around the p-value of the gene i, the question can be restated as how to estimate the ratio of the intensities of two processes around a given point P_i.

The intensity of process $R(t)$ has to be estimated at each value of t. It is possible to consider small windows of size h, or alternatively, to consider windows of different sizes corresponding to a fixed count for $R(t)$. We have chosen the latter solution, for windows of variable size seem more appealing in the particular context.

Let $\mathrm{FDR}(i)$ be the local FDR around P_i. To estimate $\mathrm{FDR}(i)$ we need to define a neighborhood around P_i. Let $V_i = V(P_{i-1}, P_i)$. Remarking that $R(P_{i-1}, P_i) = 1$, we have $\mathrm{FDR}(i) = E(V_i)$. Furthermore

$E(V_i) = P(V_i = 1)$

since V_i is a binary variable. Thus $\mathrm{FDR}(i)$ provides an unbiased estimation of $P(V_i = 1)$, the probability for gene i to be a false positive.

The raw local FDR estimate for gene i is:

$$\widehat{\mathrm{FDR}}(i,\lambda) = \begin{cases} m_0(\lambda)(P_i - P_{i-1}) & \text{if } i > 1 \\ m_0(\lambda)P_1 & \text{if } i = 1 \end{cases} \quad (1)$$

Assume that $H_0(i)$ and $H_0(i - 1)$ are true and $E(\hat{m}_0) = m_0$. Therefore this estimate is unbiased with mean 1.

Using definition (1), it is direct to obtain:

$$\frac{1}{j}\sum_{i\le j}\widehat{\mathrm{FDR}}(i,\lambda) = \widehat{\mathrm{FDR}}(P_j,\lambda)$$

which equals the *q*-value of gene *j*. The *q*-value may thus be viewed as the mean of the raw local FDR of the genes with *p*-values lower than P_j.

Under the hypothesis H_0, it is known that the differences between successive ordered values of independent realizations of the uniform([0,1]) distribution have a Beta distribution with parameters 1 and m_0 (see Johnson [9] Chap. 26). Therefore the variance of the raw local FDR estimate for non-differentially expressed genes when m_0 is known is equal to $m_0^3/[(m_0 + 1)^2 (m_0 + 2)] \approx 1$, for m_0 large enough.

The variance of estimates (1) under H_1 is generally much smaller than under H_0 (see Figures 1(a), 2(a) and 3(a) for an illustration). However, one may see on these Figures that $\widehat{FDR}(i, \lambda)$ is a very variable estimator.

This fact is well known in point process literature, [8]. Moreover, the interval $]P_{i-1}, P_i[$ is not symmetric. If we consider the neighborhood interval around P_i defined by $t_1 = (P_{i-1} + P_i)/2$, $t_2 = (P_{i+1} + P_i)/2$ then we obtain another estimate of the local FDR:

$$\widehat{FDR}(\mathrm{i},\lambda) = \frac{\widehat{m_0}(\lambda)\,(\mathrm{P_{i+1}} - \mathrm{P_{i-1}})}{2}$$

Note that (2) is a moving average of order 2 of (1). It is well known that estimates provided by moving average (or kernel estimators) are more stable, see [8].

This smoothing is generally not enough to obtain usable results and we can consider any kind of smoothing. We propose to estimate FDR(*i*) by

$$\widehat{\mathrm{FDR}}_s(i,\lambda) = f_i(\widehat{\mathrm{FDR}}(j,\lambda), j = 1,m)$$

where f_i is a smoothing function of the $\widehat{FDR}(j, \lambda)$ for $j = 1, m$, computed at position P_i.

The smoothing method must be suited to the properties of the raw FDR:

• its variance is low for low *p*-values corresponding to highly differentially expressed genes

• its variance is very high for *p*-values corresponding to non differentially expressed genes

Therefore the window of smoothing should be short for low *p*-values and large for *p*-values corresponding high *p*-values. The lowess smoothing method has a fixed number of neighbor points. Therefore its window size depends of the density of points around the *p*-value under concern. The density of points is higher for low *p*-values which in turn implies a shorter window size, which is a good property. However the adaptation of the window size is not sufficient in some cases such as in the Apo-AI example. Moreover the smoothed FDR should be an increasing function of the *p*-values, a property which is not satisfied by the lowess smoothing. Therefore we prefer to use an *ad hoc* moving average smoothing using the following algorithm for computing $\widehat{\mathrm{FDR}}_s(i, \lambda)$: let $0 < t_1 < t_2 < t_3$ be three pre-definite thresholds and $m_1 < m_2 < m_3 < m_4$ four pre-definite integers.

• if $\max_{j\le i} \widehat{FDR}(j, \lambda) < t_1$ use a moving average of order $min(2i - 1, m_1)$

• if $t_1 < \max_{j\le i} \widehat{FDR}(j, \lambda) < t_2$ use a moving average of order $min(2i - 1, m_2)$

• if $t_2 < \max_{j\le i} \widehat{FDR}(j, \lambda) < t_3$ use a moving average of order $min(2i - 1, m_3)$.

• if $\max_{j\le i} \widehat{FDR}(j, \lambda) > t_3$ use a moving average of order $min(2i - 1, m_4)$.

We have obtained good empirical results on many data sets with $t_1 = 0.01$, $t_2 = 0.05$, $t_3 = 0.2$, $m_1 = 3$, $m_2 = 5$, $m_3 = 15$ and $\widehat{FDR}(\mathrm{i},\lambda) = \frac{\widehat{m_0}(\lambda)\,(\mathrm{P_{i+1}} - \mathrm{P_{i-1}})}{2}$ with the constraint that $\widehat{\mathrm{FDR}}_s(i, \lambda)$ is not decreasing. This adaptative moving average method is quite empirical. This topic deserve some more work to build a well assessed smoothing method. This is one of our ongoing research project.

Authors' contributions

Avner Bar-Hen, Jean-Jacques Daudin and Stephane Robin equally contributed to the statistical work and the redaction task. Julie Aubert coded the R-program and analyzed the three data sets.

References

1. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *JRSSB* 1995, **57,1:**289-300.
2. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *PNAS* 2003, **100,16:**9440-9445.
3. Storey JD, Taylor JE, Siegmund D: **Strong control, conservative point estimation, and simultaneous conservative consistency**

of false discovery rates: A unified approach. *JRSSB* 2004, **66:**187-205.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov M, Coller JP, Loh M, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer:class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
5. Dudoit S, Shaffer CBJ: **Multiple hypothesis testing in microarray experiments.** *Statistical Science* 2003, **18,1:**71-103.
6. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP: *N Engl J Med* 2001, **344:**539-548.
7. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments.** *Statistica Sinica* 2002, **12:**1.
8. Cressie N: *Statistics for Spatial Data* New York: Wiley; 1993.
9. Johnson NL, Kotz S, Balakrishnan N: *Continuous Univariate Distributions* New York: Wiley; 1995.

Annexe B
Rapport technique présentant la méthode *MixThres* présentée dans la section 3.2

MixThres: mixture models to define a hybridization threshold in DNA microarray experiments

by

F. Picard, M.-L. Martin-Magniette, S. Gagnot, V. Brunaud, J. Aubert, A.-V. Gendrel, S. Robin, M. Caboche, A. Lecharny, V. Colot.

Research Report No. 20
October 2008

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

MixThres: mixture models to define a hybridization threshold in DNA microarray experiments

F. Picard[⋆,†,◦], M.-L. Martin-Magniette[†,‡,◦], S. Gagnot[‡], V. Brunaud[‡], J. Aubert[†], A.-V. Gendrel[†,⋆], S. Robin[†], M. Caboche[‡], A. Lecharny[‡], V. Colot[‡,+].

[⋆] *Laboratoire Biométrie et Biologie Evolutive,*
UMR CNRS 5558-Univ. Lyon 1, F-69622, Villeurbanne, France,
[†] *UMR AgroParisTech/INRA MIA 518,*
16 rue C. Bernard, 75231 Paris Cedex 05, France.
[‡] *URGV UMR INRA 1165-CNRS 8114-UEVE,*
2 rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France.
[⋆] *MRC Clinical Sciences Centre, Faculty of Medicine,*
Imperial College London, United Kingdom.
[+] *CNRS-ENS UMR 8186, Paris, France.*
[◦] *both authors contributed equally to this work.*

Contact:
`mlmartin@agroparistech.fr`

Abstract

Even if one of the major applications of two-color DNA microarray hybridizations is to detect differentially expressed genes using intensity log-ratios, single channel signals provide also useful information as absolute value measurements which allow the description of gene expression patterns. In this context, it becomes crucial to determine the set of probes that hybridize, that is for which the intensity signal is greater than a hybridization threshold to be fixed. Existing procedures are either an arbitrary thresholding or require the knowledge of a population of non-hybridized probes. In this work we present the MixThres method to determine an adaptive hybridization threshold from intensity levels of the complete set of probes hybridized on a chip. We define a hybridization threshold based on the histogram of the probe intensity values. Our procedure is divided into two steps. First the intensity distribution is estimated using mixture models. Second a hybridization threshold is defined from the components of the mixture. We validate our method on DNA tiling array and expression array data. We show that our method has a good reproducibility, its specificity is greater than 97 % and its precision of 88 %. The R package MixThres is available at `http://www.agroparistech.fr/mia/outil.html`

Introduction

It is well known that microarray data are corrupted by different sources of noise, one of them being the autofluorescence of the probes. When the goal of the experiment is to study differential expression between two conditions, no distinction needs to be made between non-differentially expressed and non-hybridized probes since the resulting intensity log-ratios are close to zero on average. Nevertheless, when the purpose of the experiment is to detect transcripts produced within one condition, the identification of hybridized probes becomes crucial. Throughout the paper a non-hybridized probe is a probe that has an intensity signal lower than the hybridization threshold. For these probes, it means that there is not enough signal resulting from hybridization with specific target, and not enough identity between the probe and the other targets to allow hybridization.

To the best of our knowledge, existing procedures are usually based on an estimation of a local background per spot from image acquisition softwares. In this context, an arbitrary threshold is defined for each individual probe. For example a probe can be declared above the background if red and green intensities are more than two standard deviations above background (1). Consequently, these methods are strongly dependent on the estimation of the background, which may be a poor measure of nonspecific fluorescence (2). Alternatives are proposed (3; 4), but their procedures require the knowledge of either positive and negative controls or a population of non-hybridized probes. In this work we propose to develop a statistical method to determine an adaptive hybridization threshold from intensity levels of the complete set of probes present on a chip. Our objective is also to develop a method which can be applied to any type of DNA microarray experiments.

We define a hybridization threshold based on the histogram of the logarithm (base 2) of the median intensities of the probes. Our procedure is divided into two steps. The first step consists in the estimation of the intensity distribution using mixture models. The second step is to define the threshold from the estimated density based on the components of the mixture. When mixture models are used, one needs to define the distribution of the components of the mixture as well as their number. Intensity histograms under study are defined on a finite space since the intensity signal varies between a lower bound defined as the value of the DNA autofluorescence and an upper bound defined as the saturation value. Moreover an important number of probes have a signal that is close to the lower bound which leads to a positively skewed histogram. We propose to use a Gaussian mixture model complemented with the introduction of a truncation parameter to model undirectly the dissymmetrical form. The number of components of the mixture is chosen using the BIC. We also propose to compare different models with or without truncation parameters. Finally in the second step of the procedure, the hybridization threshold is defined using conditional probabilities of membership to the different mixture components.

Two approaches are used to validate the method. The first one consists in using expression data obtained with a DNA tiling microarray that covers the whole chromosome 4 of *Arabidopsis thaliana*. The probes of this array cover genic as well as intergenic regions and when hybridizing labeled mRNAs only, we expect that probes corresponding to intergenic regions should not hybridize. We show with these dataset that our method has a good reproducibility and a specificity of 98%. The second approach consists in applying the method

to transcriptome data produced on a DNA microarray and to confirm significant signal by RT-PCR approaches. With this second approach, we estimate the method precision at 88%.

The objectif of this paper is to provide both a detailed description (histogram modelling, parameter estimation, threshold definition) and a validation of the methodology. A R package named MixThres is also available on our web site. Our methodology has already been applied in three different biological projects published in biological journals, where the methodology cannot be detailed. The first application concerned ChIP-chip data that were analyzed with truncated Gaussian mixture models to determine enriched probes (5). The authors have studied the chromatin factor TERMINAL FLOWER 2/LIKE HETEROCHROMATIN PROTEIN 1 (TFL2/LHP1) and have shown that TFL2/LHP1 associates with hundreds of small domains, almost all of which correspond to genes located within euchromatin. The aim of the second application was to improve gene annotation of *Arabidopsis thaliana* at the structural and functional levels by studying 522 samples hybridized on CATMA microarrays (6). The hybridization threshold of these 522 hybridized samples allowed them to identify 465 novel genes. In the last application the authors built a repertoire of approximately 20 000 *Arabidopsis thaliana* promoter regions, amplified by PCR and printed on glass arrays to get a promoter microarray, used then for ChIP-chip experiment (7). Truncated Gaussian mixture models was used to model the immunoprecipited sample (IP sample) of the histone acetyltransferase GCN5. The authors have shown that GCN5 associated with 40 % of the tested promoters.

Materials and Methods

Data

In this article, the signal under study comes from a two-color microarray experiment which is used to perform single-channel analysis. In this context, we define the signal as the logarithm (base 2) of the intensity of one of the two channels. As it is the case in differential analysis, a normalization procedure is required to remove technical biases. To our knowledge the only single-channel normalization procedure consists in a within-array correction followed by a possible between-array correction (8). The within-array normalization is a redistribution of the lowess correction on each channel, and the possible between-array normalization is used to force signals coming from different arrays to share a common distribution. We use their within-array correction to define our data: let $R_{ig}^{\texttt{raw}}$ and $G_{ig}^{\texttt{raw}}$ denote the raw logarithms (base 2) of the red and green channel median intensities for array i and probe g ($g = 1, \dots G$). After the within-array normalization, the normalized signals R_{ig} and G_{ig} are defined by

$$R_{ig} = R_{ig}^{\texttt{raw}} - \frac{1}{2}c(A_{ig}), \qquad G_{ig} = G_{ig}^{\texttt{raw}} + \frac{1}{2}c(A_{ig}), \tag{1}$$

where $c(A_{ig})$ is the lowess correction (9). We could work from these data, but we prefer to work with a dye-swap to control dye biases. Consequently Y_g the signal intensity of gene g is defined such that: $Y_g = (R_{1g}+G_{2g})/2$ or $Y_g = (G_{1g}+R_{2g})/2$ depending on the experimental design. Since we assume that the distributions of the intensity signal of the two arrays of a dye-swap are close by definition, we do not perform a between-array normalization.

The range of intensities measured by the Genepix scanner for one probe on one array is between 0 and 16. Since probes present some autofluorescence, the signal is generally

greater than the lower bound, say ℓ and lower than an upper bound, say u. In practice the lower bound ℓ is close to 5. As for the upper bound u, it is about 16 and may be greater due to the reconstruction of the normalized signal using Equations (1). It does not mean there was saturation as it is the case for the raw data. A second major characteristics of the signal intensity is that an important number of probes have a signal which is close to the lower bound. For the intensity histogram, it leads to a dissymmetrical form with a left peak as shown in Figure 1.

Our objective is to define a hybridization threshold from the intensity histogram. For this purpose, we propose to estimate the distribution of the intensity signal by using a mixture model and to define a hybridization threshold based on conditional probabilities of belonging to each component of the mixture for each probe.

Modelling the intensity signal distribution

The use of mixture of distributions appears natural to model the intensity histogram. Each component of the mixture can be interpreted in terms of clusters of probes with similar signal intensities and the component of interest is the one with the highest mean intensity, corresponding to hybridized probes. When using mixture models, the choice of the form of the distributions is crucial. As a preliminary analysis, many usual distributions were tested including non-symmetrical distributions such as Lognormal, Gamma or Weibull distributions, but none systematically fitted the left peak of the empirical distribution (data not shown). In order to model this asymmetry, we introduce truncation parameters $\ell = \min_g(y_g)$ and $u = \max_g(y_g)$, and we build mixture models for truncated Gaussian distributions. The expression of a Gaussian density truncated at ℓ and u is easily derived from the density of a non-truncated Gaussian. It equals

$$g_\ell^u(y;\theta) = \frac{f(y;\theta)}{F(u;\theta) - F(\ell;\theta)} \mathbb{I}\{\ell \leq y \leq u\},$$

where $f(\bullet;\theta), F(\bullet;\theta)$ represent the density and cumulative distribution functions of a non-truncated Gaussian of parameter $\theta = (\mu, \sigma^2)$. The introduction of truncation parameters allows us to re-weight the densities on the support $[\ell, u]$. Now if we consider a mixture model of K truncated Gaussians and we denote p_k the proportion of the k-th component in the mixture, the density of the data is defined by:

$$g_\ell^u(y;p,\theta) = \sum_{k=1}^{K} p_k g_\ell^u(y;\theta_k).$$

The vectors of parameters of the mixture model are $\theta = (\theta_1, \ldots, \theta_K)$ and $p = (p_1, \ldots, p_K)$ with $\sum_{k=1}^K p_k = 1$. The key element of this model is that weights $(F(u;\theta_k) - F(\ell;\theta_k))$ will be more important for components with a mean close to the truncation bounds. This strategy is used to model the left peak observed on real intensity histograms.

An EM algorithm to estimate the parameters

When using truncated Gaussian distributions, it appears that the empirical estimators of the mean and of the variance are biased due to truncation. Then a fixed-point algorithm

can be used for correction (10), using the fact that the maximum likelihood estimators are equal to the moment estimators when ℓ and u are known. In the context of a mixture of truncated Gaussian distributions, it is also crucial to perform this correction. We propose thus to modify the traditional EM algorithm (11) by including a fixed-point algorithm in the M-step.

Let Z_{kg} be a hidden random variable equal to 1 if gene g belongs to component k and 0 otherwise. By definition conditional probabilities τ_{kg} equal $\Pr\{Z_{kg} = 1|Y_g = y_g\}$ and are computed in the E-step of the algorithm. Let $p^{(h)}$ and $\theta^{(h)}$ the values of the parameters at iteration (h), then at iteration $(h+1)$ of the E-step conditional probabilities are equal to :

$$\tau_{kg}^{(h+1)} = \frac{p_k^{(h)} g_\ell^u(y_g; \theta_k^{(h)})}{\sum_{l=1}^K p_l^{(h)} g_\ell^u(y_g; \theta_l^{(h)})}.$$

The M-step is a maximisation step where the new values of parameter p and θ are computed. In our version the M-step is divided into two steps: in the M_1-step, the proportions of the components in the mixture and the empirical estimators of the mean and the variance are computed and then in the M_2-step these estimators are corrected, leading to the unbiased estimators of θ. More precisely at iteration $(h+1)$ in the M_1-step, we compute $m_k^{(h+1)}, (s_k^2)^{(h+1)}$ and $p_k^{(h+1)}$ defined by:

$$\begin{aligned}
m_k^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)} y_g}{\sum_{g=1}^G \tau_{kg}^{(h+1)}}, \\
(s_k^2)^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)} (y_g - m_k^{(h+1)})^2}{\sum_{g=1}^G \tau_{kg}^{(h+1)}}, \\
p_k^{(h+1)} &= \frac{\sum_{g=1}^G \tau_{kg}^{(h+1)}}{G}.
\end{aligned}$$

The estimators of m_k and s_k^2 are biased estimators of μ_k and σ_k^2, which are corrected in the M_2-step using a fixed-point algorithm. Denoting (j) the j^{th} iteration of the fixed-point algorithm, and setting $(\sigma_k^2)^{(0)} = (s_k^2)^{(h+1)}$, $\mu_k^{(0)} = m_k^{(h+1)}$, and $\theta_k^{(j)} = (\mu_k^{(j)}, (\sigma_k^2)^{(j)})$, we get the following algorithm:

$$\begin{aligned}
\mu_k^{(j+1)} &= m_k^{(h+1)} - A_k^{(j)} (\sigma_k^2)^{(j)}, \\
(\sigma_k^2)^{(j+1)} &= (s_k^2)^{(h+1)} \times \left\{ 1 + B_k^{(j)} + (\sigma_k^2)^{(j)} (A_k^2)^{(j)} \right\}^{-1}.
\end{aligned}$$

with

$$\begin{aligned}
A_k^{(j)} &= \frac{f(\ell; \theta_k^{(j)}) - f(u; \theta_k^{(j)})}{F(u; \theta_k^{(j)}) - F(\ell; \theta_k^{(j)})}, \\
B_k^{(j)} &= \frac{(\ell - \mu_k^{(j)}) f(\ell; \theta_k^{(j)}) - (u - \mu_k^{(j)}) f(u; \theta_k^{(j)})}{F(u; \theta_k^{(j)}) - F(\ell; \theta_k^{(j)})}.
\end{aligned}$$

In MixThres package we use the ε-accelerated version of the EM algorithm (12). The theoretical convergence of this EM algorithm is proved in (13).

Model choice

The modified EM algorithm allows us to estimate the parameters of the mixture model for a given number of components and for given truncation parameters. To fit best the histogram, we consider a collection of mixture models of untruncated, left, right and left-right truncated Gaussian distributions for which the number of components varies between 1 and K_{max}. Then we choose the best model which minimizes the BIC. Let us denote $m_K\{\ell, u\}$ a mixture model with K components with left-right truncation bounds ℓ and u respectively, the BIC is defined such that:

$$BIC(m_K\{\ell, u\}) = -2\log \mathcal{L}(Y; p, \theta | m_K\{\ell, u\}) + \log(G) \times (3K - 1),$$

where $\mathcal{L}(Y; \theta | m_K\{\ell, u\})$ denotes the likelihood of the mixture model $m_K\{\ell, u\}$.

Definition of the hybridization threshold

An ideal situation would be to select a mixture model with two components, one component for the non-hybridized population and one for the hybridized population. But in practice this situation does not occur because the reality is more complex leading to a signal distribution which is not bimodal. So, once the best model has been selected using the BIC, the components are ordered according to their mean value and the number of components is denoted $\widehat{K}$. Then our aim is to define a hybridization threshold to distinguish non-significant hybridization signal from significant hybridization signals. What we know is that the component $\widehat{K}$ with the highest mean is composed of hybridized probes, but nothing can be inferred for other components, since there is still an ambiguity between truly hybridized probes with low intensity and non-hybridized probes. One classical method to cluster probes is the *maximum a posteriori* rule (MAP): probe g belongs to component k_g^* if $k_g^* = \underset{s}{\text{Argmax}} \{\hat{\tau}_{sg}\}$. This procedure defines a natural threshold

$$T_{MAP} = \min \left\{ y_g \mid k_g^* = \widehat{K} \right\}.$$

Nevertheless, this procedure is very conservative in practice. This is why we propose the following threshold, $T(\varepsilon)$ above which a probe is declared as being hybridized:

$$T(\varepsilon) = \max \left\{ y_g \mid k_g^* < \widehat{K} \; \exists s \in \{1, \ldots, k_g^* - 1\}, \hat{\tau}_{sg} \geq \varepsilon \right\}.$$

In other words, intensity values are ranked by descending order, for each probe g k_g^* is determined using the MAP rule. Then if k_g^* differs from $\widehat{K}$, we calculate the conditional probability of belonging to each component with lower mean than $\mu_{k_g^*}$. The threshold $T(\varepsilon)$ is then defined as the first intensity value such that one of the calculated conditional probabilities is greater than ε. The performance of this procedure is assessed in the following with $\varepsilon = 10^{-4}$.

Results

We apply this method on two datasets. The first one consists of data obtained using a DNA tiling microarray of the entire known sequence of *Arabidopsis thaliana* chromosome

4 (19 Mb). The probes of this array cover genic as well as intergenic regions. The advantage of this array is that when hybridizing labeled mRNAs only, we expect that probes corresponding to intergenic regions should not hybridize. This dataset, named tiling array dataset, is used to estimate the specificity of our method. It is available on the ftp site of CATdb (`ftp://urgv.evry.inra.fr/CATdb/`). The second dataset consists of transcriptome experiments on CATMA microarray, available in the database CATdb (14). This second dataset, names CATMA dataset, has already been used to determine genes missed by the official annotation (6). This second dataset allows us to estimate the method precision. For both microarrays, all information relative to the probe sets is available on the integrative database FLAGdb++ (15).

Materials

The tiling array dataset contains results from 8 hybridizations corresponding to four biological samples in dye-swap. For each sample, polyA+ RNA was extracted from flower buds and open flowers harvested a fixed time over a one-week period from approximately 200 Arabidopsis wild-type plants of ecotype Columbia. Plants were grown under long-day conditions (16hrs white light, 22 degrees Celsius/ 8 hrs darkness, 18 degrees Celsius). Reverse transcription and cDNA labeling were performed as previously described (16). The DNA tiling microarray is described in detail elsewhere (5) and its accession number in ArrayExpress is A-MEXP-602. Briefly, this array contains ∼21000 sequential ∼1kb fragments that cover the 19 Mb Arabidopsis chromosome 4 sequence as well as a few regions located on the other four Arabidopsis chromosomes.

The CATMA dataset contains microarray data were extracted from the CATdb database developped (14). All samples were hybridized on a same array type, named CATMA for Complete Arabidopsis Thaliana MicroArray. CATMA microarray is a generic array which contains 24 576 Gene Specific Tag, small ORF and also probes designed in the chloroplastic et mitochondrial genomes. The CATMA dataset has already been analyzed with our method to identify hundreds of novel functional genes in the *Arabidopsis* genome (6). The interest of this analysis is an experimental validation allowing us an estimation of the method precision.

Data pre-processing

For both datasets, the raw data comprise the logarithm base 2 of median feature pixel intensities at wavelength 635 nm (red) and 532 nm (green). No background was subtracted. The array-by-array normalization is performed to remove systematic biases. First, we exclude spots that are considered badly formed features. Then we perform a global intensity-dependent normalization using the lowess procedure (9). Finally, for each block, the log-ratio median calculated over the values for the entire block is subtracted from each individual log-ratio value.

For the tiling array dataset, as explained in the experimental design and the Materials section, we focus on the expression data of wild-type plants of Columbia ecotype since the *Arabidopsis thaliana* annotation and the microarray probes are based on this ecotype. For each biological sample, we define the intensity signal as the average on the dye-swap, since the correlations between the two normalized signals obtained from a dye-swap varies

between 0.92 and 0.98. This indicates a high reproducibility of the microarray data between technical replicates. Informations about the data are given in Table 1.

For both datasets, to estimate the intensity histogram, a collection of mixture model of Gaussian untruncated, left-truncated, right-truncated or left-right truncated is considered. For each family of Gaussian the number of components of the mixture model varies between 1 and K_{max}=5. Consequently the model collection consists in 20 models (4 families of Gaussian times 5 different number of components). The hybridization threshold is calculated following the procedure described in Methods. The results of the tiling-array dataset are summarized in Table 2. Figures 1, 2, 3 show an example of an intensity signal histogram and its estimation, as well as the components of the selected mixture model and the estimated hybridization threshold. For the CATMA dataset, we refer to the paper of (6) for the results and their interpretation.

Reproducibility of MixThres

For each biological sample and each probe of the tiling array dataset, we calculate a hybridization index which equals 1 if the signal intensity is higher than the hybridization threshold and 0 otherwise. Since four biological samples are available, we obtain for each probe a hybridization score between 0 and 4. Out of the 21602 probes, 4681 are declared hybridized four times (score=4) and 13681 are never declared hybridized (score=0), thus 85 % of the results are coherent between the four biological samples. Moreover, the threshold as well as the percentage of probes declared hybridized vary between the four biological samples reasonably(Table 2). This shows that our method provides reproducible results. Consequently from now, we present results based on the hybridization score. A probe is declared hybridized if the hybridization score is at least 3. We choose this definition since a score of 3 means that the probe is declared for the two biological replicates and for one of the technical replicates. According to this rule, 27,3 % among the 21602 probes are declared hybridized: 4681 have a score of 4 and 1218 a score of 3. At first sight this percentage of 27,3 % can be considered low, nevertheless since targets are labeled mRNA it must be interpreted with respect to the percentage of tiles covering genes, which is about 73 %.

Specificity of the method estimated with the first dataset

By definition the specificity is the probability to declare a probe not hybridized rightly. To estimate it we focus on the set of 5724 probes which cover intergenic regions.

Among the 5724 probes, 143 are declared hybridized. Consequently the specificity is estimated at at least 97,5 %. We analyze the 143 potential false positives in detail to find an explanation of the hybridization signal. To do so we perform a bioinformatic analysis. We first perform a blast of the sequences of the 143 probes against the whole genome (e-value $< 10^{-10}$) to find hits where at least 85 nucleotides are identical. The number of hits is greater than 5 for 67 probes among the 143, indicating possible cross-hybridization. Secondly we investigate the quality of the microarray probes, which were built from PCR products. During a validation step of the microarray, a PCR product quality index, available in the FLAGdb++ database, was attributed to each probe and for 10 probes this index indicates that either the size of the PCR product associated to the probe is wrong or it gives a multiple band product. So the nature of these 10 probes is unknown and they should have

been removed from the set of probes which cover inter-genomic regions. Third thanks to the database FLAGdb++, we find that 13 probes cover at least an Expressed Sequence Tag (EST), and 4 probes have associated Massively Parallel Signature Sequencing data (MPSS data). Briefly, Ests are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene and MPSS data are short sequence signatures from a defined position within an mRNA. The output of MPSS is similar to SAGE but the method of obtaining the data is different. We refer to the data help of FLAGdb++ for more complete explanations. A similar expertise was performed for the 5724 probes and it allows us to refine the set of truly non-hybridized probes from 5664 to 3701 probes, which gives a new estimation of the specificity at 98.7 % (49 false positives among 3701). Similar estimate of the specificity is found when the procedure is applied to a seedling experiment (data not shown).

Precision of the method estimated with the second dataset

Another quantity of interest is the sensitivity defined by the probability to declare rightly a probe hybridized. Unfortunately it is impossible to determine a set of probes which must hybridize and this is why we do not carry out this study. In contrast we are able to estimate the precision which is the proportion of probes declared righly hybridized amongst all probes declared hybridized with the CATMA dataset. Indeed in the analysis of the 522 hybridized samples, for all probes declared hybridized at least once, an experimental validation was performed to validate this result. Among the 465 new genes found hybridized at least one times in the 522 hybridized samples, the hybridization evidence was confirmed by RT-PCR approaches for 88%, thus the method precision is estimated at 88%.

Discussion

We propose a method to identify hybridized probes using mixture distributions. We provide a definition of a hybridization threshold to identify hybridized probes. From the modelling point of view, considering the truncation leads to a better fit of the left peak of intensity histograms. On the studied datasets we observe that the first smallest value of BIC is associated with a truncated Gaussian mixture model and the second smallest value with the untruncated Gaussian mixture with the same number of components. Interestingly both hybridization thresholds derived from these two models are equal. This indicates that the threshold value is well-defined and does not depend on the truncation parameters. We suggest users to fix K_{max} to 5, this number of component is usually sufficient to model correctly the intensity histogram.

The hybridization threshold depends on parameter ε, which has been set at 10^{-4} in this study. Consequently the specificity and sensitivity depend on ε since a greater value will lead to a decrease (increase) in specificity (sensitivity). In order to assess the role of ε, we have tested several values from 10^{-3} and 10^{-6} and estimations of the specificity and the precision are the same. We emphasize that our method focuses on the identification of hybridized probes. Consequently the user needs to be careful regarding the interpretation of the hybridization threshold. When the signal is higher than the threshold, we showed that the corresponding probe is hybridized (see the section Specificity). However when the signal is lower than the threshold, the corresponding probe is not necessarily non-hybridized.

We apply the method on normalized data from dye-swaps, but it can also be applied on raw data or on normalized data without dye-swap. In these cases the user must be aware that technical biases could exist and alter the results. At present our model does not consider biological replicates. When available, we propose to check for reproducibility using a hybridization score as shown in Reproducibility section. Note that the generalization of a mixture model taking several samples into account is straightforward if the number of components is independent of the sample.

Funding

A.-V. Gendrel was supported by a graduate studentship from the French Ministry of Research. V. Colot was supported by a grant from the EU Network of Excellence The Epigenome.

References

[1] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[2] C.S. Brown, P.C. Goodwin, and P.K. Sorger. Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Science of the United States of America.*, 98(16):8944–8949, 2001.

[3] M Bilban, LK Buehler, S Head, G Desoye, and V Quaranta. Defining signal thresholds in dna microarrays: exemplary application for invasive cancer. *BMC Genomics*, 3(1):19, 2002.

[4] V. Stolc, Z. Gauhar, C. Mason, G. Halasz, MF. van Batenburg, SA. Rifkin, S. Hua, T. Herreman, W. Tongprasit, PE. Barbano, HJ. Bussemaker, and KP. White. A gene expression map for the euchromatic genome of drosophila melanogaster. *Science*, 306:655–660, 2004.

[5] F. Turck, F. Roudier, S. Farrona, M.-L. Martin-Magniette, E. Guillaume, N. Buisine, S. Gagnot, R.A. Martienssen, G. Coupland, and V. Colot. Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet*, 3(6):e86, Jun 2007.

[6] S. Aubourg, M.-L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, PE. Jullien, M. Ingouff, V. Thareau, T. Schiex, A. Lecharny, and JP. Renou. Analysis of catma transcriptome data identifies hundreds of novel functional genes and improves gene models in the arabidopsis genome. *BMC Genomics*, 8:401, 2007.

[7] M. Benhamed, M.-L. Martin-Magniette, L. Taconnat, F. Bitton, C. Servet, R. De Clercq, B. De Meyer, C. Buysschaert, S. Rombauts, R. Villarroel, S. Aubourg, J. Beynon, R.P. Bhalerao, G. Coupland, W. Gruissem, F.L.H. Menke, B. Weisshaar, J.-P. Renou, D.-X. Zhou, and P. Hilson. Genome-scale arabidopsis promoter array identifies targets of the histone acetyltransferase GCN5. *Plant Journal*, 2008.

[8] Y. H. Yang and N. Thorne. Single channel normalisation for cDNA microarray data. *IMS Lecture Notes– Monograph Series*, 40:403–418, 2003.

[9] 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.

[10] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. (2nd Edition)*. Wiley Series in Probability and Statistics. John Wiley & Sons. N.Y., 1994.

[11] A. Dempster, Laird N., and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[12] M. Kuroda and M. Sakakihara. Accelerating the convergence of the em algorithm using the vector ϵ algorithm. *Computational Statistics and Data Analysis*, 51:1546–1561, 2006.

[13] M. Wang, M. Kuroda, M. Sakakihara, and Z. Geng. Acceleration of the em algorithm using the vector epsilon algorithm. *Computational Statistics*, 23:469–486, 2008.

[14] S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J.-P. Renou, A. Lecharny, and V. Brunaud. CATdb: a public access to arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36(Database-Issue):986–990, 2008.

[15] F. Samson, V. Brunaud, S. Duchene, Y. De Oliveira, M. Caboche, A. Lecharny, and S. Aubourg. FLAGdb++: a database for the functional analysis of the arabidopsis genome. *Nucleic Acids Res.*, 32:D347–50, 2004.

[16] Z. Lippman, AV. Gendrel, M. Black, MW. Vaughn, N. Dedhia, WR. McCombie, K. Lavine, V. Mittal, B. May, KD. Kasschau, JC. Carrington, RW. Doerge, V. Colot, and R. Martienssen. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430:471–476, 2004.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 2 about here.]

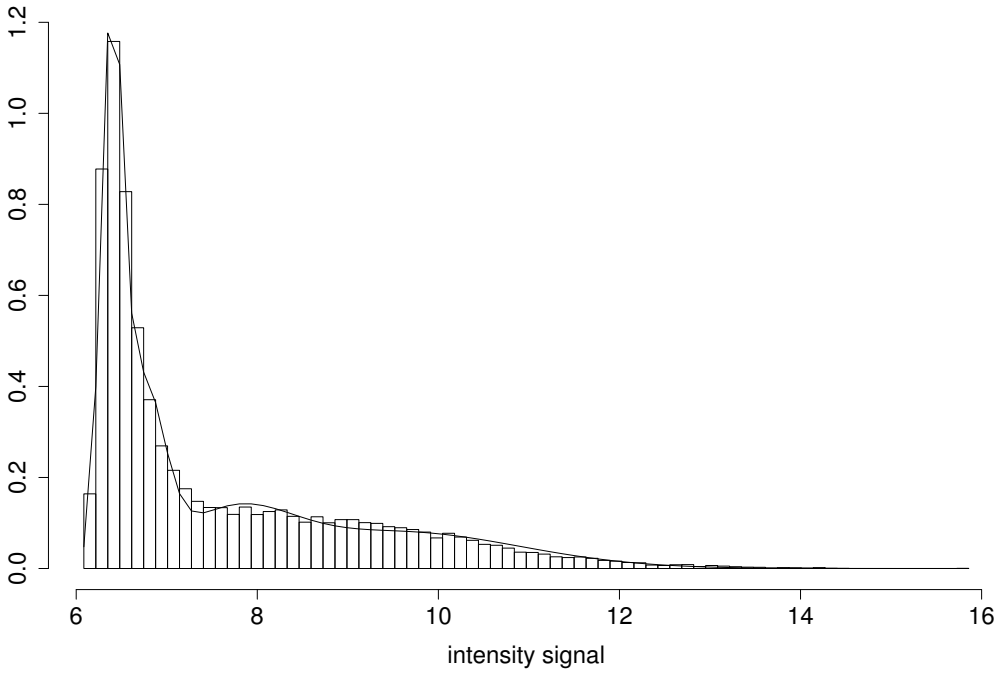


Figure 1: Intensity histogram of the second biological sample with estimated density (mixture of 4 no-truncated Gaussian distributions).

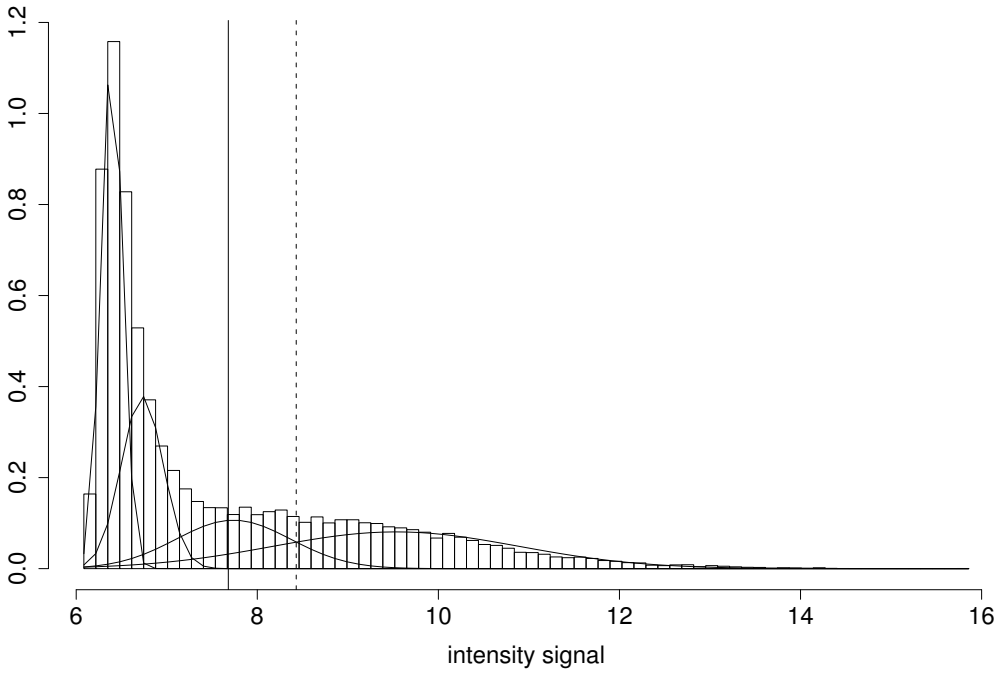


Figure 2: Intensity histogram of the second biological sample with the 4 components of the selected model. Plain vertical line indicates the hybridization threshold $T(\varepsilon) = 7.68$ with $\varepsilon = 10^{-4}$. Dot vertical line indicates the MAP threshold $T_{MAP} = 8.43$.

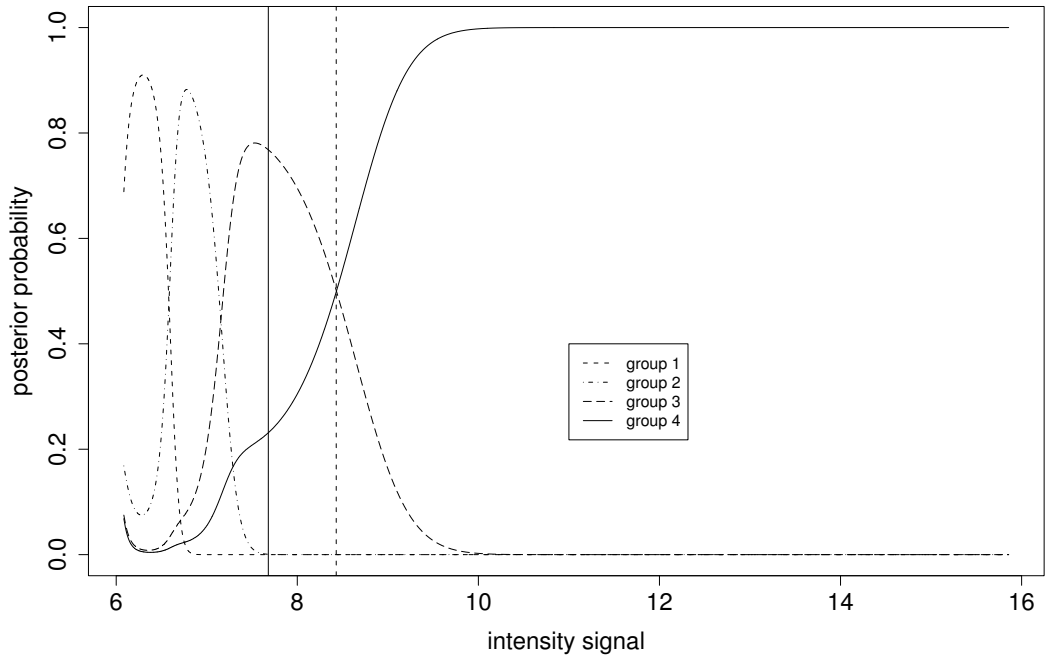


Figure 3: Conditional probabilities for the second biological sample according to the intensity signal. Plain vertical line for the hybridization threshold $T(\varepsilon)$ with $\varepsilon = 10^{-4}$. Dot vertical line for T_{MAP}.

Table 1: Data summary for the *Arabidopsis thaliana* experiment.

Biological sample	Technical sample	Normalized signal range
1	117.Red	[5.50;15.55]
1	116.Green	[5.76;15.91]
2	142.Red	[6.00;16.15]
2	143.Green	[5.86;15.56]
3	118.Red	[6.09;15.09]
3	119.Green	[6.26;15.57]
4	134.Red	[6.20;15.39]
4	135.Green	[6.24;15.56]

Table 2: Results of the hybridization study for each dye-swap. The second column is the correlation between the 2 arrays of each dye-swap.

Biological sample	Corr.	Selected model	Threshold	% of hybridized probes
1	0.92	5 right trunc. Gauss.	8.55	24.2 %
2	0.97	4 no trunc. Gauss.	7.68	35.4%
3	0.96	5 right trunc. Gauss.	9.23	27.1%
4	0.98	5 right trunc. Gauss.	9.47	30.2%

Annexe C
Article sur le biais de marquage gène-spécifique présenté dans la section 5.1.2

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 9 2005, pages 1995–2000
doi:10.1093/bioinformatics/bti302

Gene expression

Evaluation of the gene-specific dye bias in cDNA microarray experiments

Marie-Laure Martin-Magniette[1,2,*], Julie Aubert[2], Eric Cabannes[3] and Jean-Jacques Daudin[2]

[1]URGV UMR INRA 1165—CNRS 8114—UEVE, 2 rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France,
[2]UMR INAPG/ENGREF/INRA MIA 518, 16 rue C. Bernard, 75231 Paris Cedex 05, France and
[3]Laboratoire d'Immunologie Virale, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France

Received on July 7, 2004; revised on January 25, 2005; accepted on January 27, 2005
Advance Access publication February 2, 2005

ABSTRACT

Motivation: In cDNA microarray experiments all samples are labeled with either Cy3 or Cy5. Systematic and gene-specific dye bias effects have been observed in dual-color experiments. In contrast to systematic effects which can be corrected by a normalization method, the gene-specific dye bias is not completely suppressed and may alter the conclusions about the differentially expressed genes.

Methods: The gene-specific dye bias is taken into account using an analysis of variance model. We propose an index, named label bias index, to measure the gene-specific dye bias. It requires at least two self–self hybridization cDNA microarrays.

Results: After lowess normalization we have found that the gene-specific dye bias is the major source of experimental variability between replicates. The ratio (R/G) may exceed 2. As a consequence false positive genes may be found in direct comparison without dye-swap. The stability of this artifact and its consequences on gene variance and on direct or indirect comparisons are addressed.

Availability: http://www.inapg.inra.fr/ens_rech/mathinfo/recherche/mathematique

Contact: mlmartin@inapg.fr

INTRODUCTION

Many experimenters and statisticians (Kerr *et al*., 2002; Churchill, 2002) recommend using dye-swap design in cDNA microarray experiments to correct gene-specific dye bias. This artifact is not suppressed by normalization procedures such as the lowess (Yang *et al*., 2002). For a reference design some experimenters claim that dye-swaps are not necessary (Sterrenburg *et al*., 2002) whereas others use dye-swap design to preclude gene-specific dye bias (Pritchard *et al*., 2001; Brem *et al*., 2002). In direct comparison, even when the labeling artifact is better recognized, its consequences are often minimized. For example Yue *et al*. (2001) wrote 'Any variation observed in differential expression was likely a result of real variations in experimental mRNA levels rather than an artifact of the labeling system.' Tseng *et al*. (2001) described the gene*label interaction but concluded 'Theoretically some degree of gene-label interaction may exist. However this interaction appears to be insignificant in magnitude compared to other sources of variation in the present experiment.'

To our knowledge, few papers have investigated the influence of the gene-specific dye bias: Dombkowski *et al*. (2004) have shown that dye orientation can significantly influence results on differential analysis in a reference design. They have estimated that over 20% of the conclusions of their differential analysis may be inaccurate using an approach with single dye orientation. They did not identify the cause of the bias, but have urged the experimenters to use dye-swap until this artifact is better characterized. Rosenzweig *et al*. (2004) have investigated the nature of the gene-specific dye bias on a direct comparison experiment. Their analysis suggests that this artifact may concern the same probes. They proposed in their paper a new and less expensive design than the dye-swap, which attenuates the gene-specific dye bias but does not completely correct it.

In this paper, we propose an index to evaluate the magnitude of the gene-specific dye bias. The idea of the index comes from an analysis of two self–self hybridization slides. When we analyzed them, we were surprised to obtain many differentially expressed genes. The reason is that the mean log-ratio $\log_2(R_1R_2/G_1G_2)$ was wrongly calculated in place of $\log_2(R_1G_2/G_1R_2)$, where R_i and G_i denote respectively the red and green intensity on the array i. With the mean log-ratio $\log_2(R_1G_2/G_1R_2)$, no differentially expressed genes were obtained, as was expected. We have been amazed by the importance of the effect of a simple reverse of dye. To better understand the phenomenon we have written the corresponding statistical model, and deduced an index to estimate the magnitude of the gene-specific dye bias.

The paper is organized as follows. In the next section we present the statistical model taking gene-specific dye bias into account, and an index [label bias index (LBI)] to evaluate the magnitude of this artifact. Next the LBI is computed on experiments concerning several array types and organisms. We note that it is almost constant for each array type but varies from one to another. One array type seems to have low gene-specific dye bias. We are not able to explain the reasons, but this fact shows that it is possible to control this artifact. Finally we discuss the consequence of the gene-specific dye bias in direct and indirect comparisons, and try to give some insight into the mechanism of this bias.

*To whom correspondence should be addressed.

METHODS

This section is devoted to the statistical model. We underline the importance of keeping the interaction between gene and dye in the model to take gene-specific dye bias into account in the differential analysis, and we evaluate the gene-specific dye bias.

Model allowing for gene-specific dye bias

A dye-swap experiment consists of two replicate microarrays where opposite dye orientations are used. Thus each RNA sample is labeled with each dye. We consider an experiment where p dye-swaps are made. To study the data, we use the analysis of variance. Our notations follow those of Kerr *et al.* (2002). Let Y_{ijkg} be the logarithm base 2 of the measurement for array i, dye j, RNA sample k and gene g. We consider the following model:

$$Y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (DG)_{jg} + E_{ijkg}, \quad (1)$$

where A_i is the i-th array effect, D_j is the j-th dye effect, V_k is the k-th RNA sample effect, G_g is the g-th gene effect, and $(DG)_{jg}$ and $(VG)_{kg}$ are the corresponding interaction terms. The terms E_{ijkg} represent independent random errors with mean 0. If the RNA sample $k = 1$ is labeled with the dye $j = 1$ in the first array $i = 1$, then the observed difference of expression between the two RNA samples on the array i equals

$$Z_{ig} = V_1 - V_2 + (-1)^i(D_1 - D_2) + (VG)_{1g} - (VG)_{2g} + (-1)^i\{(DG)_{1g} - (DG)_{2g}\} + \tilde{E}_{ig},$$

where the errors $\tilde{E}_{ig}$ are independent random variates with mean 0.

To remove systematic biases, we perform an array-by-array normalization using the lowess procedure (Yang *et al.*, 2002). It suppresses the first four constant terms, and is supposed to alleviate the DG terms and not to alter the VG terms. We refer to the work of Kerr *et al.* (2002) for an explanation. After the normalization step, the observed difference of expression between the two RNA samples on the array i equals

$$Z'_{ig} = (VG)_{1g} - (VG)_{2g} + (-1)^i\{(DG)'_{1g} - (DG)'\}_{2g} + F_{ig},$$

where the errors F_{ig} are random variates with mean 0. The normalization step implies that $\sum_{g=1}^{G} Z'_{ig} = 0$; therefore the errors F_{ig} are not independent by construction, and they verify that $\sum_{g=1}^{G} F_{ig} = 0$. It implies a weak structural dependence of order $1/G$ between the F_{ig}. In the following we assume that the F_{ig} are independent. The departure from this assumption is too weak to have any practical importance provided that $G \geq 1000$. The difference $(VG)_{1g} - (VG)_{2g}$ is the true difference of expression between the two RNA samples. It is the difference of interest for identifying differential expressed genes. When it is non-null, it states that the gene is not transcribed in the same manner in the two RNA samples. The difference $(DG)'_{1g} - (DG)'_{2g}$ represents what is called the gene-specific dye bias. When it is non-null, it states that the probe corresponding to the gene g incorporates one of the dyes preferentially. To simplify the notations, we denote the difference $(VG)_{1g} - (VG)_{2g}$ by δ_g and the difference $(DG)'_{1g} - (DG)'_{2g}$ by β_g. The observed difference of the gene g between the two RNA samples in the array i is now re-written:

$$Z'_{ig} = \delta_g + (-1)^i\beta_g + F_{ig}, \quad (2)$$

where δ_g is the gene g differential expression and β_g the specific dye bias of the gene g. From this model we can estimate for each gene the differential expression between the two RNA samples and the gene-specific dye bias by

$$\hat{\delta_g} = \frac{1}{2p}\sum_{i=1}^{2p} Z'_{ig}$$

and

$$\hat{\beta_g} = \frac{1}{2p}\sum_{i=1}^{2p}(-1)^i Z'_{ig}.$$

When at least two dye-swaps are available ($p \geq 2$), we can also estimate the variance of F_{ig}, say σ_g^2, by the empirical estimator defined by

$$\widehat{\sigma_g^2} = \frac{1}{2p-2}\sum_{i=1}^{2p}(Z'_{ig} - \widehat{\delta}_g - (-1)^i\widehat{\beta}_g)^2.$$

It is then possible to perform a differential analysis and also an analysis of the gene-specific dye bias. For the latter purpose, it suffices to test the null hypothesis $\{\beta_1 = \cdots = \beta_G = 0\}$ against the alternative hypothesis {At least one gene is such that $\beta_g \neq 0$}. The associated test statistic can be viewed as a global index to evaluate the gene-specific dye bias. It is easily and quickly computed. We name it the LBI and it is defined by

$$\text{LBI} = \frac{\sum_{g=1}^{G}\widehat{\beta_g^2}}{\sum_{g=1}^{G}\widehat{\sigma_g^2}}. \quad (3)$$

Under the null hypothesis and assuming that $\sum_g^2$, the LBI is distributed as a Fisher distribution with $[G-1, (2p-2)(G-1)]$ degrees of freedom. The null hypothesis is rejected as soon as the test statistic is greater than $F_{G-1,(2p-2)(G-1)}(1-\alpha)$, where $F_{a,b}(\alpha)$ denotes the α-quantile of a Fisher distribution with (a, b) degrees of freedom. Note that in practice, the null hypothesis may often be rejected since the power of the test is high. So to decide if the gene-specific dye bias is important, the LBI can be also compared with the expectation of a Fisher distribution, given by $1 + \{1/[(p-1)(G-1)-1]\}$.

Although it is possible to take into account the gene-specific dye bias, in many studies the authors prefer to neglect it (e.g. Tseng *et al.*, 2001; Comander *et al.*, 2004). This leads to setting $\beta_g = 0$ for $g = 1, \ldots, G$ in the model (2). The variance of F_{ig} is thus estimated by

$$\tilde{\sigma_g^2} = \frac{1}{2p-1}\sum_{i=1}^{2p}(Z'_{ig} - \widehat{\delta}_g)^2.$$

Straightforward calculations show that $\tilde{\sigma_g^2}$ is a biased estimator of σ_g^2 if β_g differs from 0. To be precise, the bias equals $2p\beta_g^2/(2p-1)$. Therefore assuming wrongly that the β_g are null leads to overestimating the variance σ_g^2. Hence the power of the test for detecting a difference of expression will be lower when $\tilde{\sigma_g^2}$ is used in place of $\widehat{\sigma_g^2}$: some differentially expressed genes will not be detected.

When only one dye swap is made, the model (2) is over-parametrized: the number of parameters is larger than the number of observations. It is thus impossible to estimate simultaneously the difference of expression (δ_g), the gene-specific dye bias (β_g) and the variance (σ_g^2). Only two parameters per gene can be estimated. Since the major interest is the differential analysis, the parameter β_g is usually supposed to be null. In the following section, we propose a method to assess this assumption.

Evaluation of the gene-specific dye bias from self–self hybridization slides

As noticed above, when only one dye-swap is available, the statistical model (2) is no longer usable to study the observed difference of expression between two different RNA samples. Nevertheless if we consider self–self hybridization slides where the same RNA sample is hybridized against itself, it guarantees that the true difference of expression is null ($\delta_g = 0$) and thus the model (2) becomes a one-way ANOVA model:

$$Z'_{ig} = (-1)^i\beta_g + F_{ig}.$$

It is thus possible to estimate the magnitude of the gene-specific dye bias. For that purpose we calculate the LBI, defined as previously by the statistic of Fisher to test the null hypothesis $\{\beta_1 = \cdots = \beta_G = 0\}$. If $p \neq 1$, it is defined by:

$$\text{LBI} = \frac{\sum_{g=1}^{G}\hat{\beta_g}^2}{\sum_{g=1}^{G}\hat{\sigma_g}^2}, \quad (4)$$

with $\hat{\delta_g} = 0$, for all $g = 1, \ldots, G$. Under the null hypothesis, the LBI is distributed as a Fisher distribution with $[G-1, (2p-1)(G-1)]$ degrees of

Table 1. LB1 and gene-specific dye bias from 11 self–self hybridization arrays

Organism/array	Dataset	RegSS	RSS	LBI	(a)	(b)	Mean LR	Min. LR	Max. LR
Human/array 1	Wt t1	0.158	0.034	4.64	0	120	0.87	−1.19	1.58
Human/array 1	Control t2	0.156	0.027	5.68	0	153	0.45	−1.46	1.52
Human/array 1	SDF t3	0.221	0.054	4.07	0	2	1.97	1.73	2.21
Human/array 1	Wt t4	0.227	0.047	4.86	0	113	1.19	−1.16	1.81
Human/array 1	Control t5	0.280	0.060	4.64	0	33	0.19	−1.61	1.36
Human/array 1	SDF t6	0.278	0.043	6.42	0	189	1.42	−1.26	2.95
Human/array 2	SDF t2	0.120	0.012	10.29	0	8	−1.11	−1.35	−0.98
Human/array 1	SDF t2	0.080	0.016	5.15	0	3	−2.05	−2.85	−1.45
At/CATMA	leaf	0.028	0.016	1.79	0	0	—	—	—
At/CATMA	bud	0.041	0.035	1.17	0	0	—	—	—
At/CATMA	bud	0.043	0.035	1.24	0	0	—	—	—

Wt, wild type; t, time; RegSS, regression sum of squares; RSS, residuals sum of squares; LBI, label bias index; At, *A.thaliana*; (a), number of genes differentially expressed; (b), number of genes having a significant dye bias; LR, log ratio for genes having a significant dye bias.

freedom. Consequently the null hypothesis is rejected as soon as the LBI is greater than $F_{G-1,(2p-1)(G-1)}(1-\alpha)$. It readily follows that for $p = 1$,

$$\mathrm{LBI} = \frac{\sum_{g=1}^{G}(Z'_{1g} - Z'_{2g})^2}{\sum_{g=1}^{G}(Z'_{1g} + Z'_{2g})^2}. \quad (5)$$

Under the null hypothesis, its distribution is a Fisher with $(G-1, G-1)$ degrees of freedom. The null hypothesis is thus rejected as soon as the LBI is greater than $F_{G-1,G-1}(1-\alpha)$. As previously the null hypothesis is often rejected since the number of degrees of freedom is of the magnitude of G. Consequently to decide if the gene-specific dye bias is important, the LBI can be compared with the expectation of the Fisher distribution, which is equal to $(G-1)/(G-3) \sim 1$.

The LBI gives a global overview of the gene-specific dye bias. It is also interesting to have a gene-by-gene approach. For that purpose we propose to test $\{\beta_g = 0\}$ for each gene. As in the differential analysis, it is important to model the variance suitably. We have chosen to use the mixture model of Delmar *et al*. (2004). This method identifies clusters of genes with equal variance and has the good properties of keeping a good control of false positive genes and having a good power of detection. We use the Bonferroni method (with a type I error equal to 5%) in order to keep a strong control of the false positives in a multiple comparison context (Benjamini and Hochberg, 1995).

RESULTS

Data

We calculate the LBI from several self–self hybridization arrays of human and *Arabidopsis thaliana* cells.

Experiments from human cells

Resting CD4+ T cells isolated from peripheral mononuclear blood cells of healthy donors were stimulated either by the SDF-1a chemokine (SDF), or infected by the NL4-3 wild-type strain of HIV-1 (WT) or left untreated (control). For each treatment, an aliquot was removed from the cell culture at 6 different time-points over a 24 h period (30 min, 2, 4, 8, 12 and 24 h) and RNA was extracted using the RNeasy mini kit (Qiagen) according to the manufacturer's recommendations. Samples of mRNA were submitted to the T7 amplification procedure described by Phillips and Eberwine (1996), in a very similar way as previously reported (Wang *et al*., 2000). An aliquot of 4 μg of amplified RNA from a given condition (SDF, wild type or control) at a chosen time (Table 1), was used for reverse transcription and aminoallylcoupling (for details see http://cmgm.stanford.edu/pbrown/protocols/amino-allyl.htm and http://www.microarrays.org/pdfs/amino-allyl-protocol.pdf). The two halves of each aminoallyl-cDNA were coupled to NHS-Cy3 and NHS-Cy5, then purified together and hybridized onto the same array to produce a self–self hybridization.

For the first six experiments of Table 1, duplicate experiments using cells from two independent donors (RNA from same time and condition) were performed on the same day. For the next two experiments, the procedures remained the same except that the amount of starting material was doubled in order to hybridize a couple of arrays (same sample duplication).

All samples were hybridized on the same type of array consisting of 11 520 clones except for the seventh dye-swap, which was hybridized on another array of 11 616 clones spotted in duplicate. These experiments are part of a larger study that will be published elsewhere.

The arrays were scanned on a GenePix 4000A scanner (Axon Instruments, Foster City, USA) and images were analyzed by the GenePix Pro 4.0 software (Axon Instruments, Foster City, USA). For each array, the raw data comprised the logarithm base 2 of median feature pixel intensity at wavelength 635 nm (red) and 532 nm (green). No background was subtracted. The array-by-array normalization was performed to remove systematic biases. First, we excluded spots that were considered badly formed features. Then we performed a global intensity-dependent normalization using the lowess procedure (Yang *et al*., 2002). Finally, for each block, the log-ratio median calculated over the values for the entire block was subtracted from each individual log-ratio value.

Experiments from *A.thaliana* cells

Four sets of 100 *A.thaliana* Col-0 plants were grown on horticultural potting soil (Tref substrate with NFU 44-571 fertilizer, BAAN SA, Vulaines, France) under cool white light at 100 μmol m-2 s-1 with a 16-h photoperiod at 22°C and 50% humidity. Pooled samples of the flowers or the buds were harvested. The RNA extraction and target labeling were described as in Lurin *et al*. (2004).

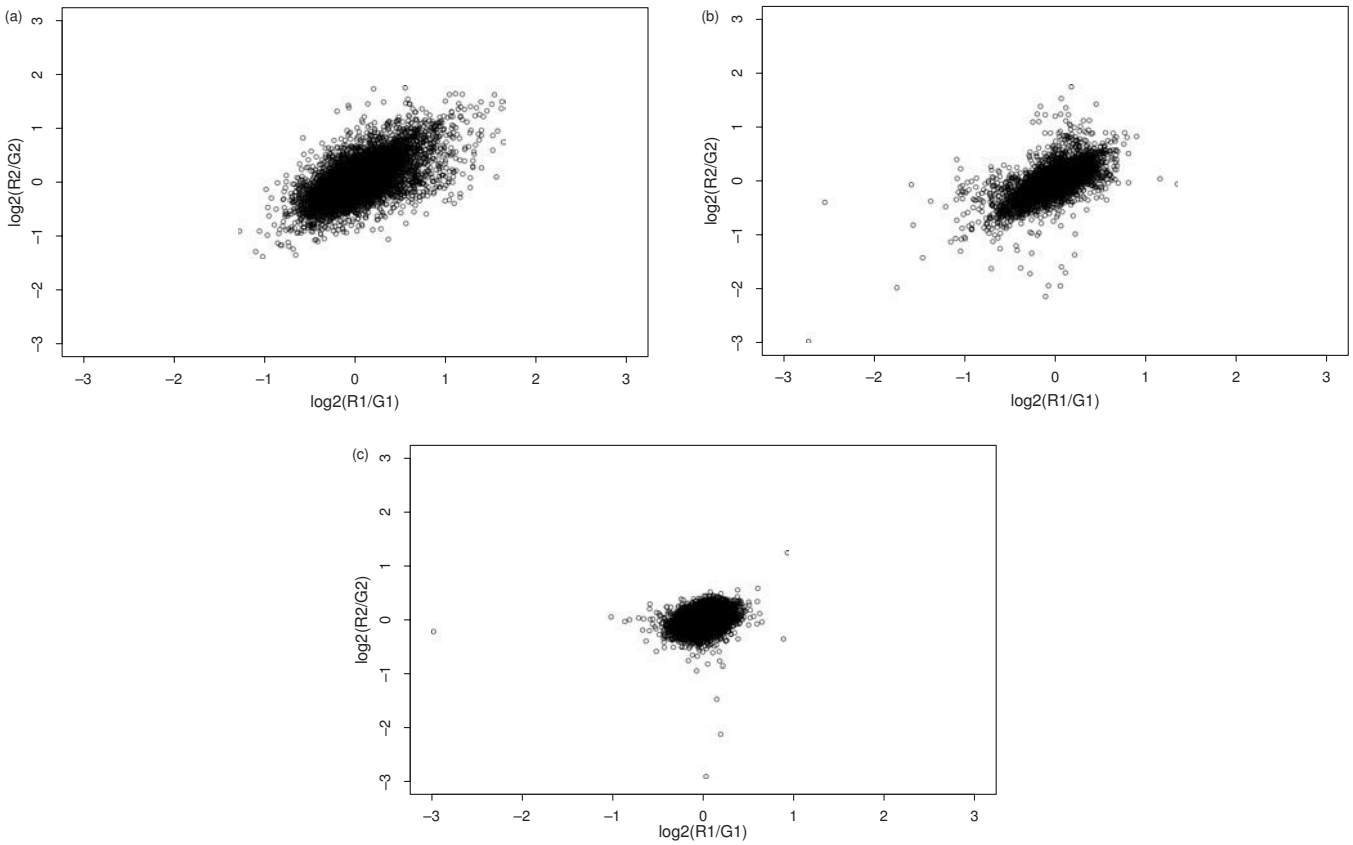


Fig. 1. Plots of the log-ratios $\log_2(R/G)$: first slide in x-axis and second slide in y-axis. (a) human/array 1, (b) human/array 2, (c) At/CATMA.

All samples were hybridized on CATMA array containing 24 576 Gene Specific Tags from *A.thaliana* (Crowe *et al*., 2003).

The arrays were scanned on a GenePix 4000A scanner (Axon Instruments, Foster City, USA) and images were analyzed by GenePix Pro 3.0 (Axon Instruments, Foster City, USA). For each array, the raw data and array-by-array normalization were respectively defined and performed as for the slides of the human cell experiments.

LBI

Table 1 summarizes the LBI computed for the 11 experiments. The LBI is the ratio between the Regression sum of squares ($\text{RegSS} = \sum_{g=1}^{G} \widehat{\beta}_g^2$) and the Residuals sum of squares ($\text{RSS} = \sum_{g=1}^{G} \widehat{\sigma}_g^2$). The RegSS, RSS and LBI values are respectively presented in the first, second and third columns of Table 1. We note that the RegSS is always $>$ RSS, so the LBI is always >1. The LBI shows that the RegSS is more than three times as high as the RSS in arrays 1 and 2 and less than twice as high as the RegSS in the CATMA array. So the dye bias is more important in the human experiments than in the experiments of *A.thaliana*. We recall that the ideal LBI (no gene-specific dye bias) is close to 1. In the experiments from *A.thaliana* cells, we have at our disposal four slides of CATMA, where the same sample of buds has been hybridized against itself. We use these four slides to evaluate the robustness of the LBI by calculating it on the six possible pairs of slides. The associated LBI varies between 1.12 and 1.26, which proves its robustness. We point out that the robustness has not been evaluated for arrays with a relatively high LBI because necessary data were not available.

To further illustrate the impact of the gene-specific dye bias, we plot the log-ratios $\log_2(R/G)$ for the two slides of the same dye-swap, for all the experiments (Fig. 1). As we have two replicates of self–self hybridization slides, nothing is expected to be seen. However one can see that there is a positive correlation between the two replicates. The only possible cause for such a correlation is the dye bias. Some genes have a higher intensity when labeled with one dye than with the other. Therefore the log-ratio $\log_2(R/G)$ is repeatedly higher (or lower) than it should be. This dye effect is higher on human experiments (correlation between 0.61 and 0.73) than on *A.thaliana* (correlation between 0.08 and 0.33). This confirms that the dye bias plays an important role in the experimental variability in the human experiments. In contrast, the dye bias seems to be better controlled in the *A.thaliana* experiments.

We also calculate the correlations between all the $\hat{\beta}_g$ for each human/array 1 experiment. These correlations are comprised between 0.45 and 0.81 (Table 2). As the array type is the same but experimental conditions vary, these correlations suggest that the dye bias may be attributed to the gene. Note that the possible gene effect is confounded with its position on the slide. Therefore it is impossible to separate the two possible causes of the labeling bias which are the nucleic composition of the probe and the spotting effect (Mary-Huard *et al*., 2004).

Table 2. Correlation between the dye bias for all the human/array 1 experiments

Experiment 1	Experiment 2	Correlation
Wild type time 1	Control time 2	0.807
Wild type time 1	SDF time 3	0.762
Wild type time 1	Wild type time 4	0.745
Wild type time 1	Control time 5	0.672
Wild type time 1	time 6	0.646
Control time 2	SDF time 3	0.718
Control time 2	Wild type time 4	0.724
Control time 2	Control time 5	0.673
Control time 2	SDF time 6	0.588
SDF time 3	Wild type time 4	0.776
SDF time 3	Control time 5	0.707
SDF time 3	SDF time 6	0.657
Wild type time 4	Control time 5	0.763
Wild type time 4	SDF time 6	0.655
Control time 5	SDF time 6	0.533

Identification of genes having a specific dye bias

After a global analysis of the gene-specific dye bias we identify the genes which are concerned. However to begin with, we assess the quality of the self–self hybridization slides by testing that each δ_g is null. Similar to the test of $\{\beta_g = 0\}$ for each gene, we use the mixture model of Delmar *et al*. (2004). The control of the false positives is done with the Bonferroni method at a level of 5%.

No gene is found to be regulated (column (a) in Table 1). Then, in order to identify genes with a significant dye bias, we test the labeling artifact using also the mixture model of Delmar *et al*. (see Methods section). Column (b) of Table 1 shows that between 0 and 189 genes have a significant gene-specific dye bias. This artifact is important in the human experiments and does not appear in the *A.thaliana* experiments. These results are in agreement with the LBI calculated in the previous section. Furthermore, all the genes having a significant dye bias are classified in the highest variance group from the differential analysis. This suggests that many genes from the highest variance group could not be detected as differentially expressed only because their 'pure' experimental variability is increased by a specific dye bias effect. This confirms that the presence of gene-specific dye bias can increase the false negative rate and so decrease the power of detection.

Table 1 contains the mean, minimal and maximal values of the $\hat{\beta}_g$ for the detected genes. One can see that the gene-specific dye bias may multiply or divide the ratio by a factor >2 which is sizeable. An analysis on the intensity level of the genes with a high specific dye bias (data not shown) shows that the intensity of these genes is in a large range between 5.5 and 15.7, with a median value between 9.5 and 10.2. Figure 2 plots the specific dye bias according to the intensity level for the first human/array 1 experiment. We can see that the magnitude of the artifact is near 0 when the intensity level is not very far from the background level. This confirms that a gene needs to be transcribed in order to reveal its specific dye bias. For higher values of the intensity level, no dependence is observed between specific dye bias and intensity level. As shown before, all expressed genes can be affected by a specific dye bias whatever their intensity level.

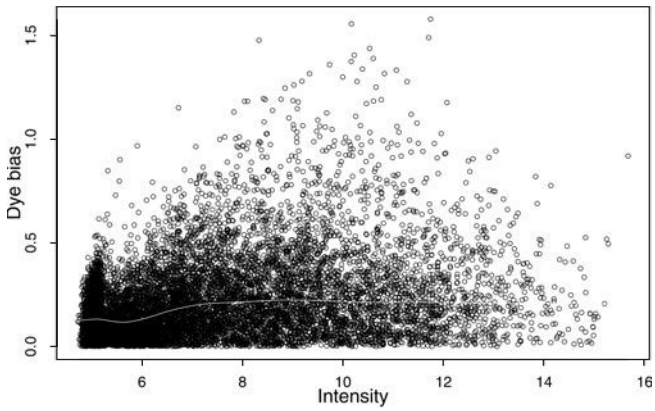


Fig. 2. Plot of the dye bias according to the intensity level in human/array 1.

DISCUSSION

Consequences of the gene-specific dye bias on direct comparison experiments

In direct comparison, two RNA samples are simultaneously hybridized on the same slide. Each sample is labeled with a dye, and it is well known that the two dyes do not have the same incorporation effectiveness. Moreover it appears that some genes are systematically badly labeled by Cy5 or Cy3 (the gene-specific dye bias). For all these reasons dye-swap design is absolutely recommended, although it is costly. Moreover in the first section we have proved that the gene*label interaction increases the experimental variability even in dye-swap experiments and thus decreases the power of the tests for detecting the differentially expressed genes.

In this paper we have proposed the LBI which is a global index to evaluate the magnitude of the gene-specific dye bias. The LBI is easily and quickly computed, and requires at least two self–self hybridization slides. After the LBI calculation we advise carrying out a gene-by-gene analysis. Even if we cannot completely describe the biochemical mechanisms of this bias, it seems that it is an artifact which involves the probes and the labeled targets, since the gene-specific dye bias can be seen only when the gene corresponding to the probe is transcribed. Consequently we advocate using a sample which hybridizes against the most possible probes. Moreover if the LBI is calculated on an array where the probes are duplicated, we think that it is better to work from the probes and not from the mean of the duplicated probes, since the gene-specific dye bias is probe-dependent. All these remarks allow us to think that the method proposed by Rosenzweig *et al*. (2004) is questionable. A condition where all genes would be transcribed simultaneously would be necessary to obtain an effective correction.

In order to investigate the gene-specific dye bias in more detail, it could be interesting for the platforms to include the LBI in their quality-control procedures, because the identification of genes which have specific dye bias is important supplementary information for the differential analysis. Moreover it could help to explain the nature of the phenomenon. According to the result of the *A.thaliana* experiment, this artifact is not an inevitability and can be well controlled. The elimination of the gene-specific dye bias could dramatically

decrease the experiment cost by removing the necessity of systematic dye-swap design.

Note that the genes can be clustered either in a group without specific dye bias ($\beta_g = 0$) or in a group with specific dye bias ($\beta_g \neq 0$). The former group has a lower experimental variability than the latter in dye-swap experiments. This explains why the mixture model on gene variances is well suited to microarray experiments (Delmar *et al*., 2004).

Consequences of the gene-specific dye bias on indirect comparison experiments

In indirect comparison an RNA sample is hybridized against a control sample. The associated design is called the reference design. As we mentioned in the introduction, it is widely assumed that reference design does not require dye-swaps. The paper of Dombkowski *et al*. (2004) demonstrated from a microarray data analysis that this assumption is not reliable. By writing the statistical model, we confirm their findings. We take the notations used throughout the paper. To take into account that the gene-specific dye bias appears only when there is transcription, we include in the model (1) the interaction between the RNA sample, the dye and the gene, say (VDG). Let us assume that the dye $j = 1$ is associated with the control sample $k = 0$, then the observed difference of expression between the i-th RNA sample and the control sample is equal to

$$\begin{aligned} Z_{ig} &= Y_{i21g} - Y_{i10g} \\ &= D_2 - D_1 + V_i - V_0 + (VG)_{ig} - (VG)_{0g} + (DG)_{2g} - (DG)_{1g} \\ &\quad + (VDG)_{i2g} - (VDG)_{01g} + \tilde{E}_{ig}. \end{aligned}$$

After the normalization step the observed difference of expression between the RNA sample and the control sample equals:

$$\begin{aligned} Z'_{ig} &= (VG)_{ig} - (VG)_{0g} + (DG)'_{2g} - (DG)'_{1g} \\ &\quad + (VDG)'_{i2g} - (VDG)'_{01g} + F_{ig}. \end{aligned}$$

Finally, the estimate for the differential expression of gene g between the two RNA samples is thus

$$Z'_{1g} - Z'_{2g} = \delta_g + (VDG)'_{12g} - (VDG)'_{22g} + \tilde{F}_g,$$

where the errors $\tilde{F}_g$ are random variates with mean 0. The gene*label interaction terms vanish but the interactions between the RNA sample, the dye and the gene remain. This is the reason why a dye-swap design is recommended even in indirect comparison.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Brem,R. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Churchill,G. (2002) Fundamentals of experimental designs for cDNA microarray. *Nat. Genet.*, **32**, 490–495.

Comander,J. *et al*. (2004) Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, **5**, 17.

Crowe,M. *et al*. (2003) CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.

Delmar,P. *et al*. (2004) Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, **21**, 502–508.

Dombkowski,A. *et al*. (2004) Gene-specific dye bias in microarray reference designs. *FEBS Lett.*, **560**, 120–124.

Kerr,M.K. *et al*. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statist. Sin.*, **12**, 203–217.

Lurin,C. *et al*. (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**, 2089–2103.

Mary-Huard,T. *et al*. (2004) Spotting effect in microarray experiments. *BMC Bioinformatics*, **5**, 63.

Phillips,J. and Eberwine,J.H. (1996) Antisense RNA amplification: a linear amplification method for analysing the mRNA population from single living cells. *Methods*, **10**, 283–288.

Pritchard,C. *et al*. (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl Acad. Sci. USA*, **98**, 13266–13271.

Rosenzweig,B. *et al*. (2004) Dye-bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.* **112**, 480–487.

Sterrenburg,E. *et al*. (2002) A common reference for cDNA microarray hybridizations. *Nucleic Acids Res.*, **30**, e116.

Tseng,G. *et al*. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.

Wang,E. *et al*. (2000) High-fidelity amplification for gene profiling. *Nat. Biotechnol.*, **18**, 457–459.

Yang,Y. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Yue,H. *et al*. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, E41-1.

Annexe D
Article sur la normalisation 3 couleurs présenté dans la section 5.1.3

Normalization for triple-target microarray experiments

Marie-Laure Martin-Magniette [*1,2], Julie Aubert[1], Avner Bar-Hen [4], Samira Elftieh[2] , Frederic Magniette [3], Jean-Pierre Renou[2] and Jean-Jacques Daudin[1]

[1] UMR AgroParisTech-INRA MIA 518, 75231 Paris Cedex05
[2] UMR INRA 1165-CNRS 8114-UEVE URGV, 91057 Evry Cedex
[3] Unité MOY300, Délégation CNRS Île de France Est, 94532 Thiais Cedex
[4] Universite Paris 5, MAP 5, PARIS cedex 06

Email: MLMM*- marie_laure.martin@agroparistech.fr; JA - julie.aubert@agroparistech.fr; ABH - avner@math-info.univ-paris5.fr; SE - elftieh@evry.inra.fr ; FM - frederic.magniette@iledefrance-est.cnrs.fr; JPR - renou@evry.inra.fr; JJD - jean-jacques.daudin@agroparistech.fr;

*Corresponding author

Abstract

Background: Most microarray studies are made using labelling with one or two dyes which allows the hybridization of one or two samples on the same slide. In such experiments, the most frequently used dyes are $Cy3$ and $Cy5$. Recent improvements in the technology (dye-labelling, scanner and, image analysis) allow hybridization up to four samples simultaneously. The two additional dyes are $Alexa488$ and $Alexa494$. The triple-target or four-target technology is very promising, since it allows more flexibility in the design of experiments, an increase in the statistical power when comparing gene expressions induced by different conditions and a scaled down number of slides. However, there have been few methods proposed for statistical analysis of such data. Moreover the lowess correction of the global dye effect is available for only two-color experiments, and even if its application can be derived, it does not allow simultaneous correction of the raw data.

Results: We propose a two-step normalization procedure for triple-target experiments. First the dye bleeding is evaluated and corrected if necessary. Then the signal in each channel is normalized using a generalized lowess procedure to correct a global dye bias. The normalization procedure is validated using triple-self experiments and by comparing the results of triple-target and two-color experiments. Although the focus is on triple-target microarrays, the proposed method can be used to normalize p differently labelled targets co-hybridized on a same array, for any value of p greater than 2.

Conclusions: The proposed normalization procedure is effective: the technical biases are reduced, the number of false positives is under control in the analysis of differentially expressed genes, and the triple-target experiments are more powerful than the corresponding two-color experiments. There is room for improving the microarray experiments by simultaneously hybridizing more than two samples.

Background

DNA microarray technology is a high throughput technique by which the expression of the whole genome is studied in a single experiment. In dual label experiments the fluorescent dyes $Cy3$ and $Cy5$ are used to label the two RNA samples co-hybridized on a same array. Recently two more dyes have been proposed (*Alexa* 488 and *Alexa* 594) allowing the simultaneous hybridization of three or four samples. [2] have evaluated triple-target microarray by comparing results of single-target, dual-target and triple-target microarrays. They have concluded that the use of triple-target microarray is valid from an experimental point of view. One year later, [7] have investigated the four-target microarray experiments. Their approach differs from that of [2], but their conclusions are in fair agreement. Their study has shown that *Alexa* 594 is best suited as a third dye and that *Alexa* 488 can be applied as a fourth dye on some microarray types. These extensions of the microarray technology are promising because they increase throughput, minimize costs and allow more powerful design of experiments. Despite these advantages, triple-target microarrays are only sparsely used [6]. The lack of guidelines for designing these experiments and for normalizing more than two-color microarray data may be an explanation. Recently [8] have proposed experimental designs for three and four-color gene expression microarrays. According to the previous work of [2], the lowess procedure [9] used to normalize data from two-color microarray is still applicable but it normalizes data sequentially because the MA-plot or the lowess correction is defined only for two dyes. Consequently, application of such a normalization method does not globally correct the dye bias due to the three dyes. Moreover the introduction of a third dye induces signal "bleeding". [2] have concluded that "it was considered as negligible between $Cy3$ and $Cy5$ signals, but seems to be important between $Alexa594$ and $Cy3$ signals", therefore signal cross-talk cannot be neglected.

We propose in this paper a normalization method for triple-target microarray experiments. First we quantify and correct the signal bleeding. Then we correct the global dye bias using a generalized lowess

procedure. Triple-target experiments with *Arabidopsis thaliana* microarrays are used to check if the proposed normalization is effective for correcting the dye bias. Moreover the comparison of the statistical power of the triple-target experiment versus the usual two-color experiment is performed. All programs and data produced for this project are available under request.

Results and Discussion

Bleeding

Using the vocabulary of [2], we call a channel, a *blank* channel when no material is hybridized for the associated dye. In theory, this blank channel should produce no signal values, and deviations from this show a bleeding phenomena from one dye-label to another. Signal bleeding from one dye-labelled sample to another is a potential source of bias. Indeed, bleeding artificially increases the signal in other channels of the same spot when the signal is high in one channel. Assume that a gene is highly expressed in condition A and weakly expressed in condition B. The difference between the two conditions is decreased by the bleeding. Therefore bleeding may induce a lowering in the statistical power for detecting differentially expressed genes. Another possibility is that the bleeding effect induces a difference between two channels for the same gene: assume that a gene is highly expressed in condition A and equally expressed in conditions B and C; if the bleeding between the channel corresponding to condition A and the channel corresponding to condition B is higher than the bleeding between A and C, then a difference between signals B and C will appear, which is a technical artifact.
In order to investigate bleeding, we have made a "single target hybridization microarray experiment" where only one dye-labelled sample is hybridized (see the dataset URGV1 in the Methods Section). We also analyzed the single target hybridization data set from Forster [2].

Experimental indications of the existence of bleeding

Figure 1 gives some plots between the hybridized and blank channels for the Forster and URGV1 datasets. These plots illustrate the bleeding. This bias depends on the channel: the bleeding bias $Cy3 \rightarrow Cy5$ is negligible but the bleeding bias $Alexa594 \rightarrow (Cy5, Cy3)$ exists. The plots from the Forster single target hybridization experiment and URGV1 datasets present the same patterns with a greater variance for the first set. As the bleeding effect seems to apply on a linear scale, we consider only raw (and not log-transformed) data in this section. Table 1 which contains the Spearman correlation coefficients between the hybridized and the blank channels for the two datasets shows that the bleeding effect exists. The

correlations between $Cy5$ and $Cy3$ are low, but the dye $Alexa594$ emits and receives significant cross-talk from the other two dyes.

Since cross-talks exist, we quantify then by using linear regression models. For example, when the sample is hybridized with $Alexa594$ and $Cy5$ and $Cy3$ are the blank channels, the following models are used: $G_i = \alpha_1 + \beta_1 Y_i + \epsilon_i$ and $R_i = \alpha_2 + \beta_2 Y_i + \epsilon'_i$, where G, Y and R stand respectively for green, yellow and red signals and i denotes the spot index. Similar models are used with swapped dyes. Estimation is performed using a robust method (R-function *rlm*, [3]) to decrease the effect of outliers. Table 2 contains the estimated parameters, which are low. This shows that the impact of bleeding on the signal is low. The greater coefficient is between $Cy3$ and $Alexa594$ (0.07). The weakness of the quantitative influence of bleeding is confirmed by the values of the standard error of the signal in the different channels: the values for the empty channels are between 6 and 200 times lower than the corresponding values for hybridized targets (Table 3). These conclusions are made for only three dyes and two experimental platforms. It is possible that other dyes or other laser technologies induce a greater bleeding bias.

Correction of bleeding

When there is a high level of bleeding it is necessary to correct it. A procedure is described in the Methods section in order to fulfill this objective. It necessitates a preliminary experiment with three single-target slides. We have used the bleeding correction for the URGV dataset in the following studies. However the results obtained are very similar with and without bleeding correction, because the importance of bleeding is not sizeable, so the data have not been corrected for bleeding in the following studies.

Note that the bleeding bias is cut down by a complete or partially dye-balanced experimental design, because the measure of the expression difference between two conditions is the mean of the individual measures of this difference taken on each slide. For example, if only one difference is distorted by the bleeding bias, its influence on the mean difference of expression is divided by the number of terms in the mean, which is equal to the number of slides containing the two conditions.

Normalization of the dye bias

Dye bias is a well characterized technical bias occuring in two-color microarray. It is mainly due to an incorporation difference between the two dyes. We refer to [4, 9] for details on this bias and also to [5] for the gene-specific dye bias. This bias is the most important technical bias and must be corrected before any transcriptome data analysis. The most used method is the lowess correction proposed by [9]. In

triple-target microarray, this bias also exists and must be corrected. Unfortunately the lowess correction is defined only for two dyes. Thus for the triple-target microarrays, [2] used the lowess correction for three dye-label combinations per array: $Cy5/Cy3$, $Cy5/Alexa594$ and $Cy3/Alexa594$. However, this procedure does not allow a global correction of the dye bias. In this paper we propose a new method generalizing the lowess correction to correct the dye bias in one step.

Let $i = 1, \cdots, n$ be the gene index (i is actually the spot index, but in the following we call it loosely the gene index), $j = 1, \cdots p$ the channel index and, y_{ij} the $\log_2$ transformed intensity measure of gene i along the channel j. Let $\overline{Y}_i = \frac{1}{p}\sum_j Y_{ij}$, be the *mean channel* raw data for gene i on the log scale, and $D_{ij} = Y_{ij} - \overline{Y}_i$, the difference between channel j and the *mean channel* for gene i. We generalize the lowess method by modelling D_{ij} as follows

$$D_{ij} = f_j(\overline{Y}_i) + E_{ij}$$

and by estimating f_j via a lowess. The value of the channel j after normalization of intensity dye-bias is defined by:

$$\widetilde{Y}_{ij} = Y_{ij} - f_j(\overline{Y}_i) = \overline{Y}_i + E_{ij}. \tag{1}$$

We point out that if this normalization procedure is applied on a two-color microarray, it leads back to the usual lowess method. Figures 2, 3 and 4 illustrate the result of the normalization procedure on an array issued from the Forster triple-self dataset. Figure 2 contains the plots showing the normalization function for each channel. In the context of two-color microarray, the MA-plot is the main graphical representation for visualizing the effect of the global dye-bias normalization. Figure 3 contains the modified MA-plots for three dyes. In such plots, the x-axis coordinate is the mean intensity of the three channels $\overline{Y}_i$ and the y-axis coordinate is the difference between intensity of channel j and the mean intensity, $D_{ij} = Y_{ij} - \overline{Y}_i$. Figure 3 contains the similar modified-MA-plots for the normalized data. The three usual MA-plot of the normalized data for each couple of dyes are represented in Figure 4.

Validation of the normalization

The normalization procedure has to be validated on two points: first it must suppress or at least cut the technical bias and second it must not reduce the difference of expression between genes. We have used different experiments to check both points. We first use an analysis of variance (Anova) approach, and then a count of the number of differentially expressed genes.

Analysis of variance of raw and normalized data

Kerr et al. [4] proposed to validate a given normalization method by analyzing the raw and the normalized data through the same Anova model. A good normalization method should cut the sum of squares due to technical factors or interactions and should not decrease the sum of squares due to the interesting biological term under consideration, the gene-condition interaction. As expected, the normalization reduces all the technical biases and the gene-condition interaction is only slightly decreased. This proves that the normalization is effective (see Table 4).

Number of genes declared differentially expressed

One way for checking the efficiency of a normalization method is to analyze self-experiments, where only one sample is labeled with all the dyes and then hybridized on the same array. In such experiments, no differentially expressed gene is expected. Differential analysis with *varmixt* ([1]) of the triple-self arrays of Forster's experiment and of the URGV2 dataset gives no genes differentially expressed after normalization. A good normalization procedure should not decrease the true difference of expression between genes. We have compared the number of differentially expressed genes for two microarray experiments, studying three conditions:

1. 3 triple-target microarrays (see URGV3 in the Methods Section)
2. 6 two-color microarrays (see URGV4 in the Methods Section), a dye-swap for each comparison between two of the three conditions.

Table 5 states the number of differentially expressed genes for each comparison and for each experiment. The two-color microarrays have been normalized using the usual lowess method and the triple-target microarrays have been normalized by equation (1). All other steps of normalization and the statistical method for differential analysis are the same for the two experiments. The experiment with three triple-target microarrays gives more differentially expressed genes than the experiment with six two-color microarrays, which proves that the proposed normalization for triple-target microarrays does not reduce the true difference between gene expression more than the usual lowess method for two dyes does.

Conclusions

The proposed normalization procedure is effective: the number of false positives is under control, and the triple-target microarray experiments are more powerful than the corresponding two-color experiments.

There is thus room for improving the routine two-color microarray experiments. The normalization procedure proposed could be used for any number of channels $p > 2$, so that it could be tested for four-target microarrays or used to evaluate the bleeding of *Alexa* 488.

Methods

Correction of bleeding

As the bleeding seems to work on a linear scale, a natural idea is to estimate $p(p-1)$ bleeding coefficients and correct the raw data using the following expression:

$$\widetilde{X}_{ij} = X_{ij} - \sum_{l \neq j} \beta_{lj} X il \tag{2}$$

where X_{ij} is the raw measure of expression of gene i on channel j, $\widetilde{X}_{ij}$ is the corresponding value corrected for bleeding, and β_{lj} is the coefficient of bleeding from channel l to channel j. This bleeding correction works under two assumptions:

1. the bleeding coefficients do not depend on the intensity of the bleeding channel,
2. the effects of the bleeding from several channels are additive on a linear scale.

The first assumption is confirmed by the preceding analysis (see Results Section) and the second one seems realistic. Two ways for estimating the coefficients β_{lj} are possible:

1. use preliminary experiment with p slides single-target hybridization,
2. use the current data set, with all the p-target hybridization slides.

The model framework for estimating the bleeding coefficients in p-target experiments is the following:

$$X_{ija} = \mu + \alpha_i + \gamma_j + \eta_{ij} + \xi_a + \tau_{ja} + \delta_{c(j,a)} + \theta_{ic(j,a)} + \sum_{l \neq j} \beta_{lja} X_{ila} + E_{ija} \tag{3}$$

where a is the array index, X_{ija} is the measure of expression for gene i, channel j and array a, $c(j,a)$ is the condition associated with channel j and slide a, α_i is the gene effect, γ_j the dye effect, η_{ij} the interaction between gene i and dye j, ξ_a the effect of array a, τ_{ja} the interaction between dye j and array a, $\delta_{c(j,a)}$ the condition $c(j,a)$ effect, $\theta_{ic(j,a)}$ the interaction gene-condition and β_{lja} is the bleeding coefficient from channel l to channel j for array a. Note that the global condition effect $\delta_{c(j,a)}$ is included in the interaction τ_{ja}. This is a standard linear model. However the size of the design matrix is huge (more than $2np$) so the computation is not routinely feasible. Even if the computation were feasible, simulations show that there

are many confounding effects in this statistical model and consequently the estimates of the β_{lja} are not reliable (data not shown).

Therefore the only possible procedure is to estimate the bleeding coefficients on preliminary one-target slides. This procedure assumes that the coefficients do not depend on the microarray and that the bleeding coefficients of the preliminary single-target experiments are the same as in real p-target experiments. The bleeding effect may depend on the platform and the technology (laser, PMT tuning, image analysis). This implies that the machine-tuning parameters are not modified during the experiment. For the bleeding correction of the URGV data sets we have used Equation (2) with the coefficients of Table 2. In practice we have only corrected the bleeding from $Cy5$ to $Alexa594$, from $Cy3$ to $Alexa594$ and from $Alexa594$ to $Cy3$.

Labelling and hybridization protocols for microarray experiments

Microarray analysis was carried out at the Unité de Recherche en Génomique Végétale (Evry, France), using the CATMA array (Crowe et al., 2003; Hilson et al., 2004), containing 24 576 gene-specific tags from *Arabidopsis thaliana*. Total RNA was extracted from each sample using TRIzol extraction (Invitrogen, Carlsbad, CA) followed by two ethanol precipitations, then checked for RNA integrity with the Bioanalyzer from Agilent (Waldbroon, Germany). cRNAs were produced from 1 μg of total RNA from each sample with the "Message Amp aRNA" kit (Ambion, Austin, TX). Then 5μg of cRNAs were reverse transcribed in the presence of 200 u of SuperScript II (Invitrogen, Carlsbad, CA), in presence of Amino-allyl-dUTP (Sigma-Aldrich, St. Louis, MO). The samples are then labelled by coupling with Cy3 or Cy5 monoreactive dyes (G.E. Healthcare, UK) or Alexa Fluor 594 (Invitrogen, Carlsbad, CA). Labelled samples were purified and concentrated with Qiaquick columns (Qiagen, Hilden, Germany). Slides were pre-hybridized for 1h and hybridized overnight at 42°C in 25% formamide. Slides were washed in 2 × SSC+ 0.1% SDS 4', 1 × SSC 4', 0.2× SSC 4', 0.05 × SSC1' and dried by centrifugation. The slides were scanned on a Genepix Professionnal 4200A scanner (Molecular Devices Corporation, St. Grégoire, France) and images were analysed by Genepix Pro 6.0 (Molecular Devices, St. Grégoire, France).

URGV Dataset description

URGV1 single target hybridization microarray experiment

Total RNA sample from Arabidopsis thaliana flowers was reverse-transcribed and labelled in a one-dye fashion either with cy3, cy5 or Alexa Fluor 594 and hybridized separately on two slides each (i.e. six hybridizations).

URGV2 triple-self hybridization microarray experiment

One pool of total RNA from Arabidopsis thaliana roots, leaves and flowers was separated in three aliquots and reverse-transcribed and labelled with the three fluorochromes, then melted and hybridized on the same slides in three technical replicates (i.e. three hybridizations).

URGV3 Triple target experiment

Total RNA from Arabidopsis thaliana roots, leaves and flowers were labelled independently with the three fluorochromes in a one-dye fashion either with cy3, cy5 or Alexa Fluor 594. Then the three samples were hybridized on the same slide, each being labelled with a different fluorochrome, in three technical replicates with fluorochrome switch (i.e. three hybridizations).

URGV4 dual target experiment

Total RNA from Arabidopsis thaliana roots; leaves and flowers were labelled independently with the three fluorochromes in a one-dye fashion either with cy3, cy5 or Alexa 594. Then two samples were hybridized on the same slide, each being labelled with a different fluorochrome. Each comparison was performed with a technical replicate with fluorochrome switch: regular dye-swap (i.e. six hybridizations).

Author contributions

MLMM, JA, ABH and JJD designed the method. MLMM, JA and JJD wrote the manuscript. JA implemented part of the software and performed the statistical analysis. SE made the URGV experiments under the direction of JPR. FM implemented part of the software. All authors contributed to the discussion and have approved the final manuscript.

Acknowledgements

This research has been sustained by the Genoplante ANR program ANR05GPLA030034 *AgriArray*.

References

1. P. Delmar, S. Robin, and J.J Daudin. Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, 21(4):502–508, 2005.
2. T Forster, Y Costa, D Roy, H J Cooke, and Maratou K. Triple-target microarray experiments : a novel experimental strategy. *BMC Genomics*, 2004.
3. P J Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.
4. M.K. Kerr, C.A Afshari, L. Bennett, P. Bushel, J. Martinez, N.J. Walker, and G.A. Churchill. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12:203–217, 2002.
5. M.-L. Martin-Magniette, J. Aubert, E. Cabannes, and J.-J. Daudin. Evaluation of the gene-specific dye bias in cdna microarray experiments. *Bioinformatics*, 21(9):1995–2000, 2005.

6. N. Reynolds, B. Collier, K. Maratou, V. Bingham, R.M. Speed, M. Taggart, C.A. Semple, N.K. Gray, and J. Cooke Howard. Dazl binds in vivo to specific transcripts and can regulate the pre-meiotic translation of mvh in germ cells. *Human Molecular Genetics*, 14:3899–3909, 2005.
7. Y CM Staal, M HM van Herwijnen, F J van Schooten, and J HM van Delft. Application of four dyes in gene expression analyses by microarrays. *BMC Genomics*, 2005.
8. Y Woo, W Krueger, A Kaur, and G Churchill. Experimental design for three-color and four-color gene expression microarrays. *Bioinformatics*, 21:459–467, 2005.
9. YH Yang, S Dudoit, P Luu, DM Lin, V Peng, J Ngai, and TP Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30, 2002.

Figures

Figure 1 - Bleeding

First row: Forster data, last row: URGV1 data. In the first column the hybridized dye is $Cy3$ and the *empty* dye is $Cy5$, in the second column the hybridized dye is $Alexa594$ and the *empty* dyes are $Cy3$ (black) and $Cy5$(green). x-axis: signal along the hybridized channel, y-axis: signal along the *empty* channel(s).

Figure 2 - Normalization function

x-axis: raw data for one channel, y-axis: normalized data from the same channel. First column: $Cy5$, second column: $Cy3$, third column: $Alexa594$.

Figure 3 - Modified-MA-plots

x-axis: mean intensity, y-axis: difference between channel and mean intensities. First row: raw data, last row: normalized data. First column: $Cy5$, second column: $Cy3$, third column: $Alexa594$.

Figure 4 - Usual MA-plots

x-axis: mean intensity between two channels, y-axis: difference between two channels. First column: $Cy5 - Cy3$, second column: $Cy3 - Alexa594$, third column: $Cy5 - Alexa594$.

Tables

Table 1 - Bleeding: correlations between hybridized and empty channels.

Mean (standard error of the mean) Spearman correlations between hybridized and empty channels.

Data	Cy5 → Cy3	Cy5 → Alexa	Cy3 → Cy5	Cy3 → Alexa	Alexa → Cy5	Alexa → Cy3
Forster	0.29 (0.06)	0.75 (0.002)	0.39 (0.06)	0.84 (0.11)	0.82 (0.02)	0.83 (0.02)
URGV1	0.13 (0.06)	0.58 (0.04)	0.02 (0.03)	0.47 (0.05)	0.26 (0.03)	0.61 (0.04)

Table 2 - Bleeding: regression coefficient between hybridized and empty channels.

Mean (se) of the regression coefficient (x1000) between hybridized and empty channels (robust regression method).

Data	Cy5 → Cy3	Cy5 → Alexa	Cy3 → Cy5	Cy3 → Alexa	Alexa → Cy5	Alexa → Cy3
Forster	1 (1)	6 (3)	0.5 (0.5)	26 (14)	2.5 (0.3)	27 (5)
URGV1	1.0 (0.1)	52 (5)	0.0 (0)	26 (2)	5 (0.4)	70(15)

Table 3 - Bleeding: Standard deviation of the signal in each channel

The hybridized target signal values are in bold.

Forster experiment

Slide	*Alexa*	*Cy3*	*Cy5*
6s	**8043**	277	193
11s	**6845**	251	191
16s	**6704**	368	245
10s	1132	1210	**6802**
17s	585	819	**6936**
18s	264	**7240**	219
23s	1033	**4188**	939

URGV experiment

Slide	*Alexa*	*Cy3*	*Cy5*
3	**1249**	73	10
6	**1124**	96	10
2	65	16	**1323**
5	78	41	**1346**
1	48	**1313**	7
4	45	**1739**	7

Table 4 - Anova Sum of Squares before and after normalization (URGV3 data set)

Source	Before normalization	After normalization
Array	1191	1184
Dye	13269	11
Array*Dye	425	43
Gene	310836	309177
Array*Gene	6362	6378
Dye*Gene	10595	2739
Condition*Gene	2387	2105
Residual	24890	23929

Table 5 - Number of genes declared differentially expressed for triple-target and two-color experiments

Number of differentially expressed genes (FDR=5%).

Comparison	triple-target experiments	two-color experiments	Common
C1 versus C2	3353	2188	1925
C1 versus C3	3986	3384	2737
C2 versus C3	4519	3465	2928

Annexe E
Article de comparaison de méthodes de normalisation pour les données de RNA-Seq présenté dans la section 5.1.4

Briefings in Bioinformatics Advance Access published September 17, 2012

doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Abstract

During the last 3 years, a number of approaches for the normalization of RNA sequencing data have emerged in the literature, differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data continue to accumulate, there has been no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis. In this work, we focus on a comprehensive comparison of seven recently proposed normalization methods for the differential analysis of RNA-seq data, with an emphasis on the use of varied real and simulated datasets involving different species and experimental designs to represent data characteristics commonly observed in practice. Based on this comparison study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

Keywords: *high-throughput sequencing; RNA-seq; normalization; differential analysis*

INTRODUCTION

During the last decade, advances in Molecular Biology and substantial improvements in microarray technology have enabled biologists to make use of high-throughput genomic studies. In particular, the simultaneous measurement of the expression levels of tens of thousands of genes has become a mainstay of biological and biomedical research. For example, microarrays have been used to discover genes differentially expressed between two or more groups of interest in a variety of applications. These include the identification of disease biomarkers that may be important in the diagnosis of the different types and subtypes of diseases, with several implications in terms of prognosis and therapy [1, 2].

In recent years, the continuing technical improvements and decreasing cost of next-generation sequencing technology have made RNA sequencing (RNA-seq) a popular choice for gene expression studies. Such sequence-based methods have revolutionized studies of the transcriptome by enabling a wide range of novel applications, including detection of alternative splicing isoforms [3, 4], genome-guided [5, 6] or *de novo* assembly of transcripts [7–9], transcript fusion detection [10] or strand-specific expression [11]. In addition, RNA-seq has become an attractive alternative to microarrays for the identification of differentially expressed genes between several conditions or tissues, as it allows for high coverage of the genome and enables detection of weakly expressed genes [12].

Corresponding author. Marie-Agnès Dillies, Institut Pasteur, PF2 Plate-forme Transcriptome et Epigénome, 28 rue du Dr Roux, Paris CEDEX 15, F-75724 France. Tel.: +33 (0) 145688651; Fax: +33 (0) 145688406; E-mail: marie-agnes.dillies@pasteur.fr
*These authors have contributed equally to this work.

The French StatOmique Consortium gathers more than 40 statisticians and bioinformaticians involved in high-throughput transcriptome data analysis in France. The objective of this group, created in 2008, is to share among researchers and practitioners knowledge of the statistical analysis of high-throughput data.

In many ways, the progression of methodological development for RNA-seq data mirrors that of microarray data, although the bioinformatic and analytical pipelines differ considerably [13]. In particular, fragmented transcripts (short reads) are sequenced, rather than hybridized onto a chip, and must be assembled or aligned to a reference genome. Different sequencing technologies and protocols are currently available and share the same general pre-processing and analytical steps as follows [13]: (i) short reads are pre-processed (e.g. in order to remove adapters and low-quality sequences) and either mapped onto a genome reference sequence or assembled, (ii) the expression level is estimated for each biological entity (e.g. a gene) [5], (iii) the data are normalized and (iv) a statistical analysis is used to identify differentially expressed biological features [14]. Questions regarding all of these steps are still open and can have a strong impact on the analysis. In this work, we focus specifically on the third step, namely the issue of normalization for RNA-seq data in the context of differential analysis.

With both microarray and sequencing data, it has been shown that normalization is an essential step in the analysis of gene expression [15–17]. In microarray data analysis, normalization enables accurate comparisons of expression levels between and within samples by adjusting for systematic biases such as dye effect and hybridization artifacts [15, 18]. Although the technical biases inherent to microarray technology are not present in RNA-seq experiments, other sources of systematic variation have been reported, including between-sample differences such as library size (i.e. sequencing depth) [19] as well as within-sample gene-specific effects related to gene length [20] and GC-content [21]. In particular, larger library sizes result in higher counts for the entire sample. Although differences in library composition between samples may not be considered to be a source of systematic variation, they may contribute to a high level of biological variability.

During the last 3 years, a number of normalization approaches to treat RNA-seq data have emerged in the literature differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data accumulate, there is still no clear indication of how the choice of normalization method impacts the downstream analysis. In addition, although effective and relevant methods have been derived and implemented to normalize RNA-seq data, they are not always properly used in practice. A small number of publications have compared normalization methods [16], providing useful yet preliminary results that must be confirmed with additional data to yield clear and robust guidelines to the community. To this end, we propose a systematic comparison of seven representative normalization methods for the differential analysis of RNA-seq data: Total Count (TC), Upper Quartile (UQ) [16], Median (Med), the DESeq normalization implemented in the `DESeq` Bioconductor package [14], Trimmed Mean of M values (TMM) implemented in the `edgeR` Bioconductor package [17], Quantile (Q) [22, 23] and the Reads Per Kilobase per Million mapped reads (RPKM) normalization [19].

In the past, comparisons among normalization methods for gene expression analysis have either made use of simulation studies or real calibration data [24–29]. Our comparison process is based on four real datasets sequenced using an Illumina sequencing machine, involving different species [*Homo sapiens* [30], *Mus musculus* (D. Castel, unpublished data), *Aspergillus fumigatus* (G. Janbon, unpublished data) and *Entamoeba histolytica* (C. C. Hon et al, submitted for publication)] and experimental designs, and dealing with both messenger RNAs and microRNAs. These four datasets were chosen to represent a broad range of characteristics and diversity typical of RNA-seq data analyses. Our comparison relies on both the qualitative characteristics of normalized data and the impact of the normalization method on the results from a differential expression (DE) analysis. In addition, a simulation study allows a further investigation of the impact of the normalization method on the false-positive rate and power of a DE analysis. Based on this study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

METHODS

In this section, we describe the normalization methods and real datasets used in our study, as well as the specific criteria used in our comparison.

Definitions

The datasets included in this study were obtained from two different Illumina sequencing machines, differing in their read length and overall throughput but sharing the same sequencing technology that

takes place on a glass slide called a 'flow cell'. A flow cell is made up of eight independent sequencing areas, or 'lanes'. Libraries are deposited on these lanes in order to be sequenced. A library contains cDNAs representative of the RNA molecules that are extracted from a given culture or tissue and are pre-processed in order to be adapted to the sequencing procedure. Similarly to microarrays, the library composition reflects the RNA repertoire expressed in the corresponding culture or tissue. The 'library size' refers to the number of mapped short reads obtained from the sequencing process of the library. In this study, a single library was sequenced in each lane.

Normalization methods

Because the most obvious source of variation between lanes is the differences in library size (i.e. sequencing depth), the simplest form of inter-sample normalization is achieved by scaling raw read counts in each lane by a single lane-specific factor reflecting its library size. We consider five different methods for calculating these scaling factors, described as follows:

Total count (TC): Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset.

Upper Quartile (UQ): Very similar in principle to TC, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors [16].

Median (Med): Also similar to TC, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors.

DESeq: This normalization method [14] is included in the `DESeq` Bioconductor package (version 1.6.0) [14] and is based on the hypothesis that most genes are not DE. A DESeq scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its geometric mean across all lanes. The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1. Assuming most genes are not DE, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfill the hypothesis. By calling the `estimateSizeFactors()` and `sizeFactors()` functions in the `DESeq` Bioconductor package, this factor is computed for each lane, and raw read counts are divided by the factor associated with their sequencing lane.

Trimmed Mean of *M*-values (TMM): This normalization method [17] is implemented in the `edgeR` Bioconductor package (version 2.4.0). It is also based on the hypothesis that most genes are not DE. The TMM factor is computed for each lane, with one lane being considered as a reference sample and the others as test samples. For each test sample, TMM is computed as the weighted mean of log ratios between this test and the reference, after exclusion of the most expressed genes and the genes with the largest log ratios. According to the hypothesis of low DE, this TMM should be close to 1. If it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (and not the raw counts) in order to fulfill the hypothesis. The `calcNormFactors()` function in the `edgeR` Bioconductor package provides these scaling factors. To obtain normalized read counts, these normalization factors are re-scaled by the mean of the normalized library sizes. Normalized read counts are obtained by dividing raw read counts by these re-scaled normalization factors.

In addition to these scaling methods, we consider two alternative strategies:

Quantile (Q): First proposed in the context of microarray data, this normalization method consists in matching distributions of gene counts across lanes [22, 23]. It is implemented in the Bioconductor package `limma` [31] by calling the `normalizeQuantiles()` function.

Reads Per Kilobase per Million mapped reads (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample and combines between- and within-sample normalization, as it re-scales gene counts to correct for differences in both library sizes and gene length [19]. However, it has been shown that attempting to correct for differences in gene length in a differential analysis actually has the effect of introducing a bias in the per-gene variances, in particular for lowly expressed genes [20]. Despite these findings, the RPKM method continues to be a popular choice in many practical applications.

All of these methods can be divided into two subgroups referring to the library size concept

(TMM and DESeq) or distribution adjustment of read counts (TC, UQ, Med, Q, RPKM). Both TMM and DESeq rely on the hypothesis that most of the genes are not DE. They both propose a scaling factor based on a mean, or median, ratio. However, for TMM this ratio is computed between each test lane and the reference one, while for DESeq all samples are taken into account. Finally, DESeq scaling factors apply to read counts, while those calculated using TMM apply to library sizes. The second group is composed of methods that assume similarities between read count distributions, either on a single quantile (TC, Med, UQ, RPKM) or on all quantiles (Q). RPKM includes both a TC and gene length normalization.

Finally, in addition to the main methods described above, some proposed strategies for RNA-seq data normalization focus on the use of housekeeping genes [16] or on the putative bias associated to GC-content [30, 32]. We did not include such a normalization strategy in our study because a close inspection of our datasets did not confirm the presence of such a bias (Supplementary Figure S13). As such, we assume that the GC bias associated with each gene is constant across conditions and does not need to be corrected in the context of a differential analysis. However, these normalization methods are further discussed in Supplementary Data.

The seven normalization methods are also compared to the raw unnormalized data, denoted by Raw Counts (RC). All the analyses are performed with `R` 2.14; the scripts used to implement each method are available in Supplementary Data. It is worth noting that all of the scaling normalization approaches described above can easily be modified to produce an offset parameter to be incorporated within a statistical model for DE.

Real data

The seven normalization methods previously described are compared based on four real RNA-seq datasets involving different species and experimental designs as well as very different characteristics in terms of reproducibility between replicates, the presence of high-count sequences, the library sizes, differences in library composition between biological conditions and the importance of gene length in estimates of gene expression (Table 1). The four datasets as well as additional details about each experiment, data pre-processing and bioinformatics steps are included in Supplementary Data.

Dataset descriptions

Homo sapiens melanoma cell lines (Hs): These human data correspond to a comparison between a melanoma cell line expressing the Microphtalmia Transcription Factor (MiTF) and a melanoma cell line in which small interfering RNAs (siRNAs) are used against MiTF in order to lower its expression [33].

Entamoeba histolytica strains (Eh): *Entamoeba histolytica* is a unicellular protozoa that can be ingested through soiled water. This human parasite is the causative agent of amebiasis, one of the three most common causes of death worldwide. The data included in this study compare gene expression between two strains of *E. histolytica* (Eh), one being virulent (HM1:IMSS) and the other being attenuated (Rahman) (C. C. Hon et al, submitted for publication).

Aspergillus fumigatus (Af): *Aspergillus fumigatus* is a fungus whose spores are present not only in the air we breathe but also in soils and decaying organic matter. It does not normally cause illness but can induce fatal pulmonary infections to individuals with a weakened immune status. These RNA-seq data compare the transcriptome of *A. fumigatus* strain 1163 in two different growth media.

Mus musculus muscle stem cells (Mm): These data are related to a transcriptome study where the expression of miRNAs was measured in three different cellular stages of the skeletal muscle lineage in adult mouse.

Comparison procedures

Qualitative characteristics of normalized data. For each dataset, the seven normalization methods are compared based on qualitative characteristics of normalized data, including the count distributions and variability between biological replicates. Boxplots of raw and normalized read counts are calculated as log2(read count + 1) in order to avoid problems associated with zero values. The within-condition variability measure is based on the coefficient of variation per gene. Boxplots represent the distribution of this coefficient across samples.

We also investigated the average variation of a set of 30 housekeeping genes in the human data, assuming that these genes are similarly expressed across samples (lanes). The housekeeping genes were selected from a previously described list [34] and presented the least variation across the 84 human cell types of the GeneAtlas data [35] available on `GEO`

Table 1: Summary of datasets used for comparison of normalization methods, including the organism, type of sequencing data, number of genes, number of replicates per condition, minimum and maximum library sizes, Pearson correlation between replicates and between samples of different conditions (minimum, maximum), percentage of reads associated with the most expressed RNA (minimum, maximum), library type (SR = single-read or PE = paired-end read, read length, D = directional or ND = non-directional) and Illumina sequencing machine

Organism	Type	Number of genes	Replicates per condition	Minimum library size	Maximum library size	Correlation between replicates	Correlation between conditions	% Most expressed gene	Library type	Sequencing machine
H. sapiens	RNA	26 437	{3, 3}	2.0×10^7	2.8×10^7	(0.98, 0.99)	(0.93, 0.96)	≈ 1%	SR 54, ND	GaIIx
A. fumigatus	RNA	9248	{2, 2}	8.6×10^6	2.9×10^7	(0.92, 0.94)	(0.88, 0.94)	≈ 1%	SR 50, D	HiSeq2000
E. histolytica	RNA	5277	{3, 3}	2.1×10^7	3.3×10^7	(0.85, 0.92)	(0.81, 0.98)	6.4–16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	{3, 2, 2}	2.0×10^6	5.9×10^6	(0.95, 0.99)	(0.09, 0.75)	17.4–51.1%	SR 36, D	GaIIx

(http://www.ncbi.nlm.nih.gov/geo) with the accession number GSE1133.

Differential expression analysis. The seven normalization methods are compared based on results from a DE analysis performed with the Bioconductor package `DESeq` and the Two-Stage Poisson Model (TSPM) [36], both described below. In addition to comparing the number of DE genes and the number of common DE genes found among the methods, we generate, for each real dataset, a dendrogram representing the similarity between the DE gene lists obtained with each normalization method, based on the binary distance and the Ward linkage algorithm (`dist()` and `hclust()` functions in `R`). The four dendrograms (Supplementary Figure S4) are subsequently merged into a consensus dendrogram resulting from the mean of the distance matrices obtained from each real dataset.

Simulations

Simulation model

The simulation model is similar to one previously used [29] and adapted to counts. Let N be the number of genes and M the number of samples divided into two conditions, and let x_{ij} be the expression value of a given gene i in sample j. We assume x_{ij} follows a Poisson distribution of parameter λ_{jk} according to the condition k to which sample j belongs. Under this model, the null hypothesis H_0 of no difference between the two conditions is equivalent to $\lambda_{i2} = \lambda_{i1}$; the alternative hypothesis H_1 of DE between the two conditions is equivalent to $\lambda_{i2} \neq \lambda_{i1}$. Finally, let π_0 (resp. π_1) be the proportion of genes generated under H_0 (resp. H_1) among the N genes.

Data were simulated with $N = 15\,000$, $M = 20$ (10 samples per condition) and π_1 increasing from 0% to 30%. In order to generate realistic data, the parameter λ_{i1} used to sample the gene i from a Poisson distribution for the first condition corresponds to the observed mean expression for each gene estimated from the *M. musculus* data; the parameter λ_{i2} used to sample the gene i from a Poisson distribution for the second condition is equal to λ_{i1} under H_0 and to $(1 + \tau)\lambda_{i1}$ under H_1, with $\tau = \pm 0.2$. To assess the impact of non-equivalent library sizes, we added the possibility of multiplying all gene expression values x_{ij} for a given sample j by a constant K_j taken to be equal to $|\varepsilon_j|$, where ε_j is drawn from a $N(1, 1)$ distribution. In addition, the *M. musculus* data contain a set of highly expressed genes contributing to the majority of total counts, which enables an assessment of the impact of such high-count genes in the simulated data.

False-positive rate and power

For each simulated dataset, the false-positive rate (power) can be estimated based on the genes simulated under H_0 (H_1). We consider three settings: (i) equivalent library sizes across lanes and no high-count genes, (ii) non-equivalent library sizes across lanes and no high-count genes and (iii) equivalent library sizes across lanes and presence of high-count genes. For each scenario and proportion of H_1 tested, the false-positive rate and the power were averaged over 10 simulated datasets to ensure a reasonable precision.

Differential expression analysis

In both the real and simulated data, the impact of the normalization methods is assessed using the results from a DE analysis. For this test, we choose to use two methods based on different models: the `DESeq`

Bioconductor package [14] and the TSPM [36], which may be implemented using an `R` script found at the corresponding author's website. The DESeq method, which was specifically developed to find differentially expressed genes between two conditions for RNA-seq data with small sample size and overdispersion, uses a model based on a negative binomial distribution and local regression to estimate the relationship between the mean and variance of each gene. `DESeq` was chosen because it is widely used in practice. In addition, it allows scaling factors to be easily included in the statistical test, and in contrast to `edgeR`, the statistical test does not assume comparable distribution of read counts. The `DESeq` Bioconductor package (version 1.6.0) with default setting was employed. The package accommodates each normalization method via the specification of size factors in the following function:

```
AnnotatedDataSet(pData(cds)$sizeFactor <-...)
```

In order to confirm these results using an alternative method, we have also applied the TSPM [36], which makes use of a model based on the Poisson, rather than negative binomial, distribution. The TSPM evaluates the presence of overdispersion on a gene-by-gene basis in a first stage, and subsequently tests for DE using a standard likelihood approach for genes displaying evidence of overdispersion, or a likelihood ratio test statistic for those without evidence of overdispersion.

For both methods, raw *P*-values were adjusted for multiple comparisons by the Benjamini–Hochberg procedure [37], which controls the false discovery rate. Genes with an adjusted *P*-value < 0.05 were considered to be differentially expressed.

RESULTS

In comparing the seven normalization methods, we aim to identify methods that appear to have both robust and stable performance across real or simulated datasets exhibiting a variety of characteristics. We note that RNA-seq technology provides the opportunity to explore the expression of transcripts rather than genes for organisms exhibiting complex transcription patterns. Although some of the normalization strategies included in this study apply to both read counts and estimated expression levels, others are adapted only to read counts. As such, all data included in this study contain gene-level read counts rather than estimated transcript expression levels.

Real data

We consider two criteria for the comparisons made on four real datasets, described in detail in Table 1: (i) the qualitative characteristics of normalized data and (ii) results from DE analyses. For the former, we focus on boxplots of the distribution of counts as well as a study of intra-group variability. We remark that drawing definitive conclusions from these qualitative comparisons concerning the performance of each normalization method is typically not possible; however, such exploratory analyses are often undertaken in the early stages of an analysis and help shed light on characteristics of the data and the impact of the normalization process on the data distribution prior to further analysis. For the latter, we study the lists of differentially expressed genes between conditions identified following the use of each normalization method in each dataset.

Qualitative characteristics of normalized data

Like microarray data, in typical DE analyses the majority of genes under consideration are often assumed to be non-differentially expressed between conditions. For this reason, it is useful to examine boxplots of counts across samples in each dataset, both before and after normalization; an effective normalization scheme should result in a stabilization of read count distributions across replicates. For data with small differences in library size and little inter-sample variability (e.g. the *H. sapiens* data), it is perhaps unsurprising that all methods, including the unnormalized raw counts, yield comparable results (Supplementary Figure S1). However, when large differences in library size exist (e.g. the *A. fumigatus* and *M. musculus* data), these sample-to-sample differences are evident in the boxplots for the unnormalized raw counts.

In the case of the *M. musculus* miRNA-seq data, we note that although most of the other methods appear to perform similarly in stabilizing these differences, TC and RPKM do not improve over the raw counts (Figure 1a). A similar pattern may be seen in the results for the *A. fumigatus* data (Supplementary Figure S1). In addition to large differences in library size, the *M. musculus* data also exhibit the presence of high-count genes (i.e. a few genes whose read counts contribute to a large proportion of the total count for a given sample) associated with different expressed RNA repertoires. The TC normalization method thus corrects for differences in sequencing depth, but it is unable to handle differences in RNA

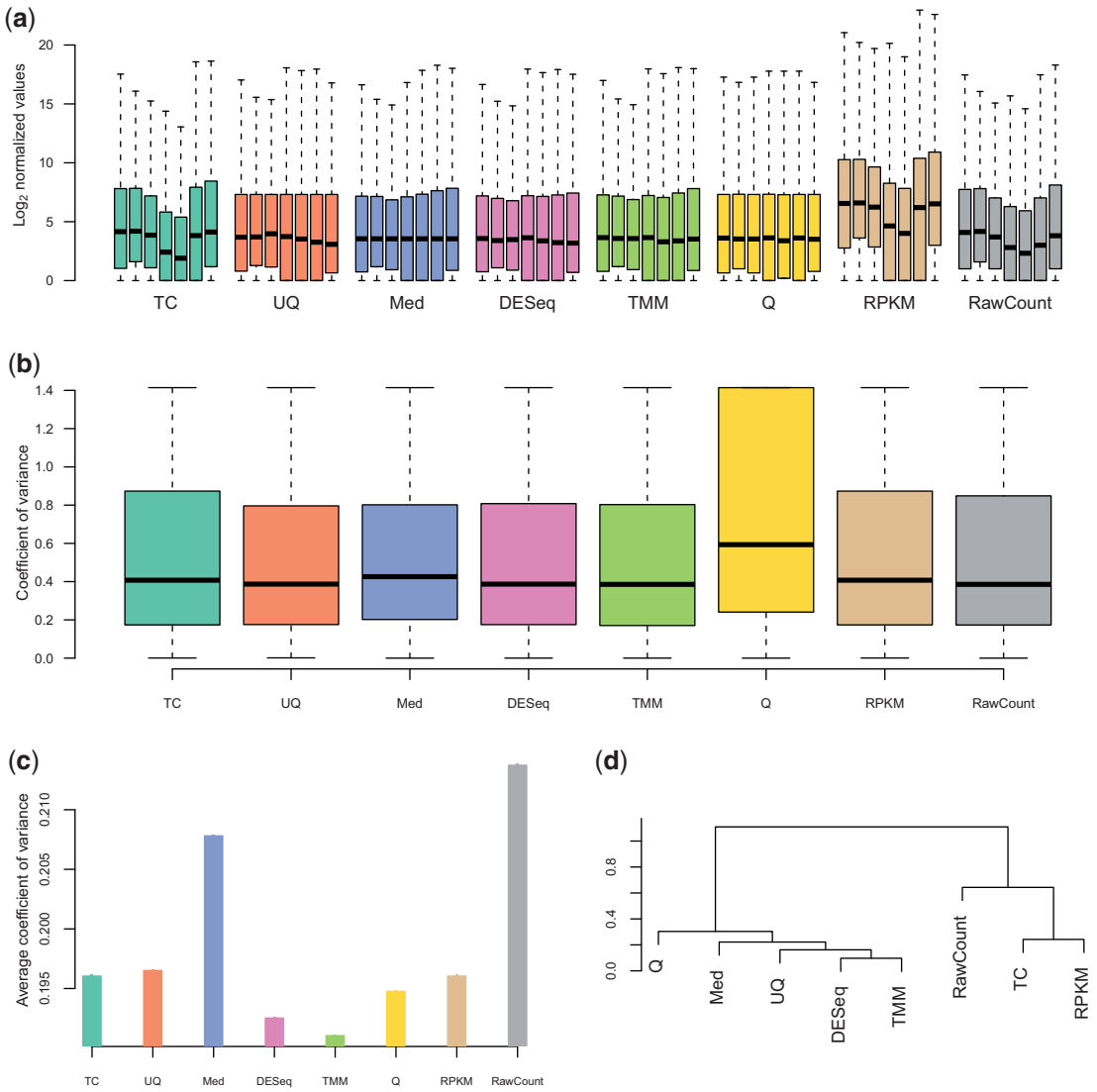


Figure 1: Comparison of normalization methods for real data. (**A**) Boxplots of log2(counts + 1) for all conditions and replicates in the *M. musculus* data, by normalization method. (**B**) Boxplots of intra-group variance for one of the conditions (labeled 'B' in the corresponding data found in Supplementary Data) in the *M. musculus* data, by normalization method. (**C**) Analysis of housekeeping genes for the *H. sapiens* data. (**D**) Consensus dendrogram of differential analysis results, using the **DESeq** Bioconductor package, for all normalization methods across the four datasets under consideration.

composition among samples [17]. In addition, we note that even following a normalization using the Q method, the first quantiles of the samples in the *M. musculus* are not aligned; this is due to a subset of samples that contain a much higher proportion of 0 counts (Supplementary Figure S12).

These boxplots of normalized values also indicate subtle discrepancies between normalization methods that are similar in nature. As an example, we remark upon the differences between the Med and UQ methods in the *M. musculus* data; the former aligns the median values for counts across all samples, while the latter aligns the upper quantile of counts across all samples. However, differences in library composition across samples, such as the aforementioned presence of high-count genes or a large numbers of 0 counts, affect the calculation of these scaling factors unequally (Supplementary Figures S9–S11 and S14).

It is also of interest to consider which normalizations are able to minimize intra-condition variance. In most of the datasets considered here, little difference is observed among the normalization methods (Supplementary Figure S2). One exception occurs in the *M. musculus* data, where Q normalization actually

appears to increase, rather than decrease, the intra-group variance for one of the conditions (Figure 1b). This can be explained by looking at read count distributions across the seven mouse samples. In particular, the read count distributions in one of the conditions (labeled 'B' in the corresponding data found in Supplementary Data) are quite different from those in the other two conditions, with more extreme counts (very low or very high) but fewer moderate counts (data not shown). As the Q normalization process corrects gene counts by matching distributions across all samples on the basis of the mean distribution, read counts of this condition are corrected more than read counts of the others. This over-correction in turn increases intra-condition variability, especially for genes with moderate counts.

Finally, we consider the effect of the various normalization methods on the variation in expression among a set of housekeeping genes in the human data, which may be assumed to be similarly expressed across samples. Figure 1c represents the average coefficient of variation of 30 known housekeeping genes in the human data (see Supplementary Data for further detail). Considering that these genes are assumed to have relatively constant expression, we note that the DESeq and TMM normalization methods lead to the smallest coefficient of variation. Although choosing an appropriate set of such housekeeping genes can be difficult, these results complement the previous qualitative observations concerning the behavior of the normalization methods under consideration.

Differential expression analysis

Because the aim of this comparative study is to determine the downstream effect of the choice of normalization method, we also consider results from a DE analysis based on the `DESeq` Bioconductor package and TPSM method. With real data, it is difficult to determine whether a particular normalization method is superior to the others (e.g. through the false-positive rate). However, the advantage of such a comparison is that it allows us to determine which methods perform similarly.

Table 2 indicates that there is a great deal of overlap among all of the normalization methods in data with little inter-sample variability (e.g. the *E. histolytica* data) using the `DESeq` package; the same general trend may be seen with results from the TSPM (Supplementary Table S10). However, across datasets Q and RPKM tend to uniquely identify weakly expressed genes as differentially expressed (Supplementary Figure S3). These same patterns were observed across all datasets for the `DESeq` method (Supplementary Figure S4, Supplementary Tables S5–S9) and are displayed in the consensus dendrogram tree in Figure 1d. This consensus dendrogram illustrates a trend, namely that in the results from a DE analysis, the TC normalization tends to group with RPKM and the unnormalized raw counts, while the remaining methods tend to group together. We note that although the number of genes identified as differentially expressed differs between the DESeq and TSPM methods (Supplementary Tables S5–S10), the same general relationships may be observed among the different normalization methods, and the consensus dendrogram tree constructed using results from the TSPM is nearly identical to that constructed from the DESeq results (Supplementary Figure S15). This suggests that the relationships identified among the normalization methods are not simply linked to the model used for the differential analysis.

Simulations

Although comparisons using real data are informative, simulations complement these results by allowing different factors, including differences in library size and RNA composition, to be controlled. With this in mind, the false-positive rate and power resulting from the DE analysis may be calculated in a variety of scenarios: equivalent or non-equivalent library sizes between lanes and presence or not of high-count genes contributing to a large proportion of the total count for a given sample. By varying these factors, differences among the normalization methods become more apparent.

In situations where library sizes are simulated to be equivalent and no high-count genes are present, all normalization methods considered perform nearly identically to the unnormalized raw counts in terms of the false-positive rate and power, using the `DESeq` Bioconductor package; this is unsurprising, as normalization is unneeded in such a case (Supplementary Figure S5a). In situations where library sizes are different (Supplementary Figure S5b), we note that the nominal false-positive rate is not maintained and the power is significantly decreased for the unnormalized data. All of the normalization methods are able to correct for these differences in library sizes, as all control the false-positive rate and

Table 2: Number of differentially expressed genes found in common for each of the normalization methods using the **DESeq** Bioconductor package, as well as the unnormalized raw counts (RC), in the *E. histolytica* data

	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1213**	**1195**	**1160**	**1172**	**1054**	416	184
Med			**1218**	**1147**	**1160**	**1043**	416	183
DESeq				**1249**	**1169**	**1058**	413	184
TMM					**1190**	**1051**	516	184
Q						**1092**	414	184
RPKM							417	149
RC								184

Counts along the diagonal indicate the number of DE genes per method (i.e. 548 DE genes for the TC method, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e. 547 DE genes in common between TC and UQ). Numbers in bold correspond to pairs of methods with very similar lists of DE genes.

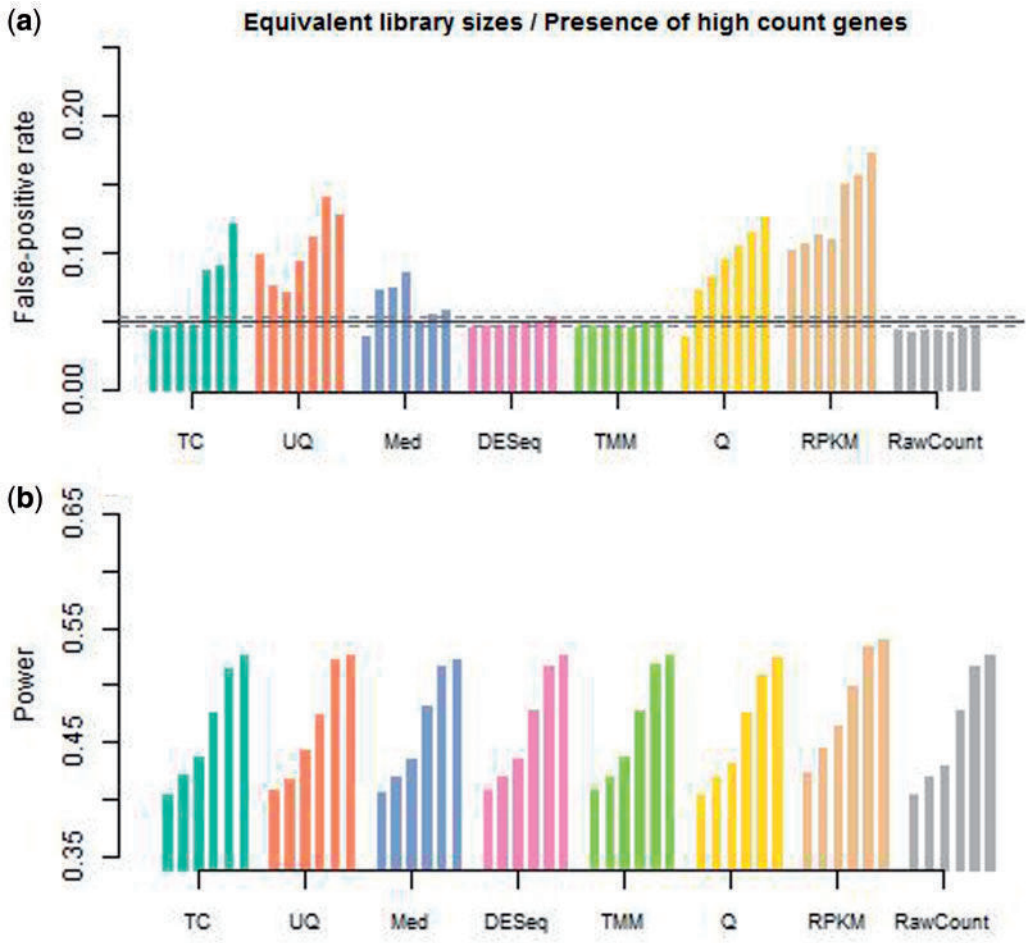


Figure 2: Comparison of normalization methods for simulated data with equal library sizes and the presence of high-count genes. (**A**) Average false-positive rate over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 0% to 30% for each normalization method). (**B**) Power over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 5% to 30% for each normalization method).

maintain a reasonable power. Figure 2 presents results from the most discriminant simulation setting, where the library sizes are simulated to be equivalent for all samples with the presence of a few high-count genes. This setting indicates that contrary to the situation with varying library sizes, the presence of high-count genes does not impact the performance of raw counts; this seemingly contradictory result is due to the fact that the data are simulated under the model used for the differential analysis. However, the presence of these high-count genes clearly results in an inflated false-positive rate for five of the normalization methods (TC, UQ, Med, Q and RPKM). Only DESeq and TMM are able to control the false-positive rate while also maintaining the power to detect differentially expressed genes.

DISCUSSION

Despite initial optimistic claims that RNA-seq data do not require sophisticated normalization [38], in practice normalization remains an important issue since raw counts are often not directly comparable within and between samples. While this subject has received some attention in the literature, the increasing number of RNA-seq normalization methods makes it challenging for scientists to decide which method to use for their data analysis. Given the fact that the choice of normalization has a great influence on the subsequent statistical analyses, the quality and credibility of these methods need to be assessed fairly [39]. To this end, our comparison study deals with seven representative normalization strategies compared on four real datasets involving different species and experimental designs, and on simulated datasets representing various scenarios.

Based on three real mRNA and one miRNA-seq datasets, we confirm previous observations that RPKM and TC, both of which are still widely in use [40, 41], are ineffective and should be definitively abandoned in the context of differential analysis. The RPKM approach was initially proposed to account for differences in gene length [19]; however, the relationship between gene length and DE actually varies among the datasets considered here (Supplementary Figures S6–S8). Even in cases where a strong positive association between gene counts and length is observed, scaling counts by gene length with RPKM is not sufficient for removing this bias [16, 20]. Several alternative approaches to account for gene length at the steps of normalization, differential analysis or gene-set analysis have been proposed [19, 32, 42], but no standard strategy has yet been identified. The TC approach, on the other hand, ignores the fact that different biological samples may express different RNA repertoires. In addition, it may too often be biased by the behavior of a relatively small number of high-count genes that are not guaranteed to have similar levels of expression across different biological conditions [16]. Similarly, Q is based on the strong assumption that all samples must have identical read count distributions. As shown in our comparison, this may lead to increased within-condition variability and should be avoided. The other normalization methods (UQ, Med, DESeq and TMM) perform similarly on the varied datasets considered here, both in terms of the qualitative characteristics of the normalized data and the results of DE analyses.

Simulations allow a further discrimination of the seven methods, in particular in the presence of high-count genes, where it appears that only DESeq and TMM are able to maintain a reasonable false-positive rate without any loss of power. One should note that DESeq and TMM are also indicated through an investigation of the variation of housekeeping genes in the *H. sapiens* data, although this analysis should be interpreted with caution. Housekeeping genes are assumed to have similar expression levels across samples of different tissues, but there is no guarantee that this hypothesis holds in every condition tested. However, taken together with the previous conclusions, these results confirm the satisfactory behavior of the DESeq and TMM methods. We also remark that in terms of the scaling factors used, DESeq and TMM are the most similar normalization methods. Finally, these two methods do not explicitly include an adjustment of count distributions across samples, allowing samples to exhibit differences in library composition. It is not surprising, then, that these two methods performed much better than the others for data with differences in library composition. A summary of these conclusions is shown in Table 3.

It is important to keep in mind that most normalization strategies (including DESeq and TMM) rely on the rather strong assumptions that most genes are not differentially expressed, and that for those differentially expressed there is an approximately balanced proportion of over- and under-expression [22, 43]. Though these assumptions appear reasonable in many studies, including those considered here, there are experiments in which they are not met.

Table 3: Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Unfortunately, these assumptions are rarely checked in practice; in fact, it would be extremely difficult to do so. In recent work, to address the observation that the proportion of DE genes can affect normalization quality, Kadota *et al*. [44] proposed an alternative multi-step normalization strategy in which genes that are determined to be potentially DE are removed prior to estimation of scaling factors using the TMM normalization method. This work suggests that in some cases, the appropriate choice of parameters can lead to slight improvements in performance in the TMM method.

On a practical note, DESeq and TMM are straightforward to apply through a command of the `DESeq` and `edgeR` Bioconductor packages, respectively. We note that unlike the other methods, TMM and DESeq use a normalization factor within the statistical model for differential analysis, rather than on the data themselves; one consequence of this approach is that the corresponding packages do not automatically provide normalized read counts to the end user, although this information is often appreciated and requested by biologists. However, normalized read counts for the DESeq and TMM methods can be obtained through a simple command in the `DESeq` package or a series of `R` commands, respectively, as shown in Supplementary Data. As the two packages implement normalization methods with comparable performance, a comparison of their respective statistical models dedicated to differential analysis may provide further arguments to favor one of the two methods.

The present study represents a major step toward a more comprehensive use of normalization methods for RNA-seq data and will be of great help to biologists that are confronted with RNA-seq data analyses. As sequencing technology continues to mature, the use of multiplexed experiments will likely become increasingly common, paving the way to a dramatic growth in the amount of data produced; additional work will be needed to determine how to include such multiplexed samples within a normalization scheme. In addition, this work is restricted to normalization methods for processing read counts, and as such its conclusions are limited to this context. In particular, it assumes that complex transcriptomes are studied at the gene, rather than transcript, level. Normalization and differential analysis at the transcript level require the use of sophisticated statistical models such as Cufflinks [5] or RSEM [45] in order to estimate, rather than count, expression levels of these transcripts. These estimates do not have the same statistical properties as read counts and may not be described by the same models or processed by the same normalization algorithms. An exception can be made for the `DEXseq` Bioconductor package [46], which proposes a detection of differential exon usage based on read counts *per exon* and applying the DESeq normalization. Another comparative study will be carried out in the future to address this more complex yet fruitful area.

SUPPLEMENTARY DATA

Supplementary Data are available online at http://bib.oxfordjournals.org/.

Key points

- Normalization of RNA-seq data in the context of differential analysis is essential in order to account for the presence of systematic variation between samples as well as differences in library composition.
- The Total Count and RPKM normalization methods, both of which are still widely in use, are ineffective and should be definitively abandoned in the context of differential analysis.
- Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions, both of which are typical of real RNA-seq data.

Acknowledgements

We thank Chung Chau Hon from Institut Pasteur (whose work was supported by the French National Research Agency (ANR-10-GENM-011)) for his very helpful support in providing up-to-date annotations of *E. histolytica*, as well as Thomas Strub, Irwin Davidson and the IGBMC sequencing platform for supplying the *H. sapiens* RNA-seq data. We also thank Guilhem Janbon from Institut Pasteur, who designed the *A. fumigatus* experiment and kindly accepted that his data be included in this study, and Delphine Charif, who participated in discussions concerning this work. On behalf of the French StatOmique Consortium and in alphabetical order: J.A., M.-A. D., M.G., C.H.-A., F.J., M.J., C.K., S.L.C., A.R. and N.S. wrote the manuscript. J.A., D.C., M.-A.D., J.E., M.G., G.G., C.H.-A., F.J., B.J., M.J., L.J., C.K., D.L., S.L.C., C.L.G., G.M., A.R., B.S. and N.S. designed and performed the analyses.

FUNDING

This work was supported by the Groupe de Rercherche Bioinformatique Moléculaire (GdR BiM, http://www.gdr-bim.u-psud.fr). D.C. was supported by a DIM STEM-Pole fellowship and Association Française contre les Myopathies.

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
2. Sørlie T, Tibshirani R, Parker J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003;**100**:8418–23.
3. Wang ET, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008; **456**(7221):470–6.
4. Pan Q, Shai O, Lee LJ, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;**40**(12):1413–5.
5. Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–5.
6. Guttman M, Garber M, Levin JZ, *et al*. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010;**28**:503–10.
7. Robertson G, Schein J, Chiu R, *et al*. De novo assembly and analysis of RNA-seq data. *Nat Methods* 2010;**11**:909–12.
8. Grabherr MG, Haas BJ, Yassour M, *et al*. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 2011;**29**(7):644–52.
9. Schulz MH, Zerbino DR, Vingron M, *et al*. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;**28**(8):1086–92.
10. Maher CA, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458(7234)**:97–101.
11. Levin JZ, Yassour M, Adiconis X, *et al*. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 2010;**7(9)**:709–15.
12. Marguerat S, Bähler J. RNA-seq: from technology to biology. *Cell Mol Life Sci* 2010;**67(4)**:569–79.
13. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol* 2010; **11**(220).
14. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11(R106)**:R106.
15. Park T, Yi SG, Kang SH, *et al*. Evaluation of normalization methods for microarray data. *BMC Bioinformatics* 2003; **4**(33).
16. Bullard JH, Purdom E, Hansen KD, *et al*. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* 2010; **11**(94).
17. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**(R25).
18. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;**32**:496–501.
19. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 2008;**5**:621–8.
20. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;**4**(14).
21. Pickrell JK, Marioni JC, Pai AA, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**(7289): 768–72.
22. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**:185–93.
23. Yang YH, Thorne NP. Normalization for two-color cDNA microarray data. *Science and Statistics: A Festschrift for Terry Speed*, Vol. 40. IMS Lecture Notes – Monograph Series, 2003, 403–18.
24. Shedden K, Chen W, Kuick R, *et al*. Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics* 2005;**6**(26).
25. Qin LX, Beyer RP, Hudson FN, *et al*. Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics* 2006;**7**(23).
26. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006;**7**(359).
27. Jaffrézic F, Marot G, Degrelle S, *et al*. A structural mixed model for variances in differential gene expression studies. *Genet Res* 2007;**89**:19–25.
28. McCall MN, Irizarry RA. Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Res* 2008; **36**(17):e108.
29. Jeanmougin M, de Reynies A, Marisa L, *et al*. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE* 2010;**5**(9):e12336.

30. Strub T, Giuliano S, Ye T, *et al*. Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene* 2011;**30**:2319–32.
31. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, (eds). *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, 2005;397–420.
32. Risso D, Schwartz K, Sherlock G, *et al*. GC-content normalization for RNA-seq. *BMC Bioinformatics* 2011;**12**:480.
33. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 2012;**13**(2):204–216.
34. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* 2003;**19(7)**:362–5.
35. Su AI, Wiltshire T, Batalov S, *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004;**101(16)**:6062–7.
36. Auer PL, Doerge RW. A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol* 2011;**10**:1–28.
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Stat Methodol* 1995;**57**(1):289–300.
38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
39. Hofmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol* 2002;**3**.11:research0033–research0033.
40. Liu S, Lin L, Jiang P, *et al*. A comparison of RNA-seq and high-density exon array for detecting differential gene expression between related species. *Nucleic Acids Res* 2011; **39**(2):578–88.
41. Isabella VM, Clark VL. Deep sequencing-based analysis of the anaerobic stimulon in Neisseria gonorrhoeae. *BMC Genomics* 2011;**12**(51).
42. Young MD, Wakefield MJ, Smyth GK, *et al*. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;**11(2)**:R14.
43. Calza S, Pawitan Y. Normalization of gene-expression microarray data. *Methods Mol Biol* 2010;**673**:37–52.
44. Kadota K, Nishiyama T, Shimizu K. A normalization strategy for comparing tag count data. *Algorithms Mol. Biol.* 2012;**7**:5.
45. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323.
46. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Research* 2012. doi:10.1101/gr.133744.111.

Supplementary Materials for "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis"

Authors: Marie-Agnès Dillies[1]*, Andrea Rau[2]*, Julie Aubert[3,4]*, Christelle Hennequet-Antier[5]*, Marine Jeanmougin[6,7]*, Nicolas Servant[8,9,10]*, Céline Keime[11]*, Guillemette Marot[12,13], David Castel[14] , Jordi Estelle[2,15,16], Gregory Guernec[17] , Bernd Jagla[1] , Luc Jouneau[18] , Denis Laloë[2] , Caroline Le Gall[19] , Brigitte Schaëffer[20] , Stéphane Le Crom[21,22,23,24,25]*, Mickaël Guedj[6]* and Florence Jaffrézic[2]*, on behalf of the French StatOmique Consortium

[1]Institut Pasteur, PF2 - Plate-forme Transcriptome et Epigénome, Paris, F-75015 France

[2]INRA, UMR 1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, F-78352 France

[3]AgroParisTech, UMR 518 Mathématiques et Informatique Appliquées, Paris, F-75005 France

[4]INRA, UMR 518 Mathématiques et Informatique Appliquées, Paris, F-75005 France

[5]INRA, UR83 Recherches Avicoles, Nouzilly, F-37380 France

[6]Pharnext, Department of Biostatistics, Issy-les-Moulineaux, F-92130, France

[7]Laboratoire Statistique et Génome, Université d'Evry Val d'Essonne, UMR CNRS 8071 - USC INRA

[8]Institut Curie, Paris, F-75248 France

[9]INSERM, U900, Paris, F-75248 France

[10]Ecole des Mines de Paris, Fontainebleau, F-77300 France

[11]Institut de Génétique et de Biologie Moléculaire Cellulaire, CNRS UMR 7104, INSERM U 596, Université de Strasbourg, Illkirch, France

[12]Université Lille Nord de France, UDSL, EA2694 Biostatistics

[13]Inria Lille Nord Europe, MODAL team

[14]Institut Pasteur, Stem Cells and Development, CNRS URA 2578, 25 rue du Dr Roux, Paris, F-75015, France

[15]AgroParisTech, UMR 1313 GABI, Jouy-en-Josas, F-78352, France

[16]CEA, DSV/iRCM/SREIT/LREG, Jouy-en-Josas, F-78352, France

[17]INRA, UR1012 SCRIBE, IFR140, GenOuest, 35000 Rennes, France

[18]INRA UR0892, Unité de Virologie et Immunologie Moléculaires, Jouy-en-Josas, F-78352, France

[19]Institut de Mathématiques, Université de Toulouse and CNRS (UMR5219), Toulouse, F-31062 France

[20]UR341, Mathématiques et Informatique Appliquées (MIA), INRA, Jouy-en-Josas, F-78352 France

[21]Ecole normale supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France

[22]INSERM, U1024, Paris, F-75005 France

[23]CNRS, UMR 8197, Paris, F-75005 France

[24]UPMC Univ Paris 06, UMR 7622, Laboratoire de Biologie du Développement, 9 quai St. Bernard, F-75005, Paris, France

[25]CNRS, UMR 7622, Laboratoire de Biologie du Développement, 9 quai St. Bernard, F-75005, Paris, France

*These authors have contributed equally to the work

Corresponding author: marie-agnes.dillies@pasteur.fr

Marie-Agnès Dillies is a statistician at Institut Pasteur, PF2 - Plate-forme Transcriptome et Epigénome in Paris, France. She has ten years of experience in transcriptomic data analysis.

Andrea Rau is a Junior Research Scientist at INRA, UMR 1313 Génétique Animale et Biologie Intégrative, in Jouy-en-Josas, France. Her research interests focus on the development of statistical methodology for the analysis of high-throughput 'omics data.

Julie Aubert is a statistician with more than eight years experience in transcriptomic data analysis. She works at AgroParisTech, UMR 518 Mathématiques et Informatique Appliquées and INRA, UMR 518 Mathématiques et Informatique Appliquées, in Paris, France.

Christelle Hennequet-Antier is a Statistical Consultant at INRA, UR83 Recherches Avicoles, in Nouzilly, France, and has over seven years of experience in analyzing transcriptomic data.

Marine Jeanmougin is a Ph.D. student in biostatistics at Pharnext, Department of Biostatistics, in Issy-les-Moulineaux, France, and Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 - USC INRA, in Evry, France. Her projects focus on transcriptomic data analysis, including normalization and gene selection issues as well as network inference strategies.

Nicolas Servant is the coordinator of the NGS data analysis group of the Bioinformatic Platform at Institut Curie, in Paris, France; INSERM, U900, in Paris, France; and Ecole des Mines de Paris in Fontainebleau, France. He is involved in the development of computational approaches dedicated to next generation sequencing analysis (small RNA-seq, high-throughput Chromosome Conformations Capture).

Céline Keime is a research engineer at the IGBMC Microarray and Sequencing platform (CNRS-INSERM-University of Strasbourg, France), where she manages the bioinformatics team, working on the analysis of NGS data.

Guillemette Marot is an Assistant Professor in the department of Biostatistics at Université Lille Nord de France, UDSL, EA2694 Biostatistics, and in the MODAL team at Inria Lille Nord Europe. Her research interests focus on statistical modeling for 'omics data.

David Castel is a postdoctoral fellow at Institut Pasteur, Stem Cells and Development, CNRS URA 2578, in Paris, France. His research interests focus on the genetic regulation of quiescence and self-renewal of muscle stem-cells.

Jordi Estelle is a Junior Research Scientist at INRA, UMR 1313 Génétique Animale et Biologie Intégrative, in Jouy-en-Josas, France; AgroParisTech, UMR 1313 GABI, in Jouy-en-Josas, France; and CEA, DSV/iRCM/SREIT/LREG, in Jouy-en-Josas, France.

Gregory Guernec is a Research Engineer at INSERM UMR1027 Epidemiology and Public Health analyses, in Toulouse, France. He spent the last five years at INRA, UR1037 Laboratoire de Physiologie et de Génomique des Poissons (LPGP), in Rennes, France.

Bernd Jagla is a bioinformatician at Institut Pasteur, PF2 - Plate-forme Transcriptome et Epigénome, in Paris, France.

Luc Jouneau is a Research Engineer at INRA UR0892, Unité de Virologie et Immunologie Moléculaires, in Jouy-en-Josas, France.

Denis Laloë is a Senior Research Engineer at INRA, UMR 1313 Génétique Animale et Biologie Intégrative, in Jouy-en-Josas, France. His research focuses on applying geometric data analysis to different fields of animal genetics and genomics.

Caroline Le Gall joined the Institut de Mathématiques, Université de Toulouse and CNRS (UMR5219), in Toulouse, France, in 2010 for seventeen months as a biostatistician and worked on 'omics and high-

throughput data.

Brigitte Schaëffer is a statistician engineer at UR341, Mathématiques et Informatique Appliquées (MIA), INRA, in Jouy-en-Josas, France.

Stéphane Le Crom is a Professor in Functional Genomics at the University Pierre et Marie Curie (Paris, France) and is responsible for the Genomic Platform at École normale supérieure, Institut de Biologie de l'ENS, IBENS, Paris ; INSERM, U1024, Paris; and CNRS, UMR 8197, in Paris, France.

Mickaël Guedj is head of the Bioinformatics/Biostatistics department at Pharnext in Issy-les-Moulineaux, France, and Adjunct Professor at the École Nationale de la Statistique et de l'Analyse de l'Information (Ensai). His research focuses on statistical genetics, computational biology, and their applications to cancer and rheumatological, neurodegenerative, and metabolic diseases.

Florence Jaffrézic is a Senior Research Scientist at INRA, UMR 1313 Génétique Animale et Biologie Intégrative, in Jouy-en-Josas, France.

Contents

---

1 Real data sets

In this section we provide additional detail about the data described in the main text. The four datasets used in the paper are also provided as supplementary data as text files with raw counts. The human melanoma cell lines data correspond to file Hs.txt, the *Entamoeba histolytica* data to Eh.txt, the *Mus musculus* miRNA-seq data to Mm.txt and the *Aspergillus fumigatus* data to Af.txt. The genes have been rendered anonymous for all these datasets. The housekeeping genes in the human melanoma cell lines data are noted HK00001 to HK00034, and correspond to the list of genes given in the file housekeepingEnsemblId34.txt (not necessarily in the same order).

1.1 *H. sapiens* (Hs) Melanoma cell lines

The human data (Hs) are related to a transcriptome study of the effect of Micropthalmia Transcription Factor (MITF) on a human melanoma cell line (501Mel), where gene expression in this cell line was observed following small interfering RNA-mediated MITF knockdown (siMITF) as compared to control siluciferase (siLUC) cells. The transfection and culture of 501Mel cells are as previously described [1]. Subsequently, RNA-seq libraries were prepared using mRNA-seq Sample Prep Kit (RS-100-0801, Illumina) following Illumina's protocol with some modifications. For these libraries, mRNA were purified from 1µg total RNA using poly-T oligo-attached magnetic beads and fragmented using divalent cations at 95◦ C during 5 minutes. The cleaved mRNA fragments were reverse transcribed in cDNA using random primers. This was followed by second strand cDNA synthesis using Polymerase I and RNase H. These double stranded cDNA fragments were blunted, phosphorylated and ligated to double-stranded adapters. Size selection was performed by electrophoresis on a 2% agarose gel and DNA fragments in the range of ∼250-350bp were excised and purified using QIAquick Gel Extraction Kit (Qiagen). Then a PCR amplification was performed (30 sec at 98◦ C; [10 sec at 98◦ C, 30 sec at 65◦ C, 30 sec at 72◦ C] × 13 cycles; 5 min at 72◦ C). After PCR amplification, PCR products were purified using AMPure beads (Agencourt Biosciences Corporation), according to the manufacturer's instructions. DNA libraries were checked for quality and quantified using 2100 Bioanalyzer (Agilent). The libraries were loaded in the flowcell at 8pM concentration and clusters were generated following Illumina's instructions. Libraries were sequenced on an Illumina Genome Analyzer IIx as single-end 54 bases reads, following the manufacturer's protocols. Image analysis and base calling were performed using the Illumina pipeline v1.8.0. Reads were then trimmed to remove adapters (with length ≥ 5 and error rate ≤10%) and low quality sequences (phred score ≤ 25) using cutadapt (http://code.google.com/p/cutadapt). Sequences shorter than 25bp after clipping were discarded from further analysis. Reads were then mapped onto the hg19 assembly of the human genome using GSNAP v2011-09-14 [2]; four mismatches and novel splicing were allowed. This

aligner was chosen as in a comparative analysis of RNA-seq alignment algorithms [3], the results provided by GSNAP were very good for nearly all criteria evaluated. Quantification of gene expression was done on uniquely aligned reads using HTseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) with intersection-nonempty mode and gene annotations from Ensembl release 64. For the RPKM calculation, gene length was calculated as the median of the length of all transcripts corresponding to each gene. For additional information about the Hs data, see Supplementary Table 1.

1.2 *E. histolytica* (Eh) strains

Entamoeba histolytica is an unicellular parasite which causes amebiasis in humans. The data used in this study correspond to a transcriptome experiment comparing gene expressions between two strains of *Entamoeba histolytica* (Eh), one being virulent (HM1:IMSS) and the other being attenuated (Rahman) . (C.C. Hon et al., submitted for publication). Briefly, both strains were cultivated axenically in a TYI-S-33 medium at 36◦ C. Total RNA were extracted using TRIzol reagents and poly(A)+ mRNA were purified from total RNA using Dynabeads according manufacturer's instructions (Invitrogen). Strand non-specific pair-end libraries (mRNA-Seq 8-Sample Prep Kit) were prepared according manufacturer's instructions (Illumina) and were sequenced using an Illumina HiSeq2000 instrument (Illumina). Genome scaffolds of *E. histolytica* HM1:IMSS version 1.3 were used for mapping [4]. Unspliced and spliced reads were mapped using Bowtie [5] and HMMSplicer [6], respectively. Gene models were retrieved from AmoebaDB version 1.3 and were revised using the RNA-seq data with manual curation (C.C. Hon et al., submitted for publication). For this study, we only used a subset of 5,277 genes out of the 7,312 revised gene models. This subset includes all genes without introns, genes with no alternative splicing as well as genes where alternative splicing events are restricted to exon skipping. The remaining genes have been excluded from this study because of the uncertainty in their gene models. For additional information about the Eh data, see Supplementary Table 2.

1.3 *A. fumigatus* (Af) in two different growth conditions

Aspergillus fumigatus is a fungus whose spores are not only present in the air we breathe, but also in soils and in organic matter in decay. It does not normally cause illness but it can induce fatal pulmonary infections to individuals with a weakened immune status. The strain used for these experiments was *Aspergillus fumigatus* A1163 [7]. For preparation of the RNA samples, 100 ml of liquid YEG (Yeast Extract 1%, glucose 3%) were inoculated with 8 x 105 conidia/ml and incubated at 37◦ C for 16h at 150 rpm. At this point the culture was equally divided into two aliquots to which either Caspofungin (Merck) at a final concentration of 25 ng or distilled water was added, followed by incubation at 37◦ C for 60 min at 150 rpm. Fungal tissue was then collected by filtration and immediately lyophilized. RNA was extracted with TRIzol Reagent (Invitrogen) following the manufacturer's instructions. Two independent biological replicates were prepared for each growth condition. Directional single read libraries were prepared according to Illumina instructions with one modification. Poly(A)+ mRNA were purified from total RNA using Dynabeads mRNA Purification kit according to manufacturer's instructions (Invitrogen). Strand-specific single-end libraries were prepared based on the RNA ligation method (Lister et al. 2008; Levin et al. 2010) using the reagents of Small RNA Sample Prep Kit v1.5 (Illumina). All libraries were sequenced using an Illumina HiSeq2000 instrument (Illumina). Adapter sequences and sequences of low quality were removed using a KNIME workflow [8]. The cleaned sequences were then aligned to the reference genome using Bowtie [5] (-a –best -q -m50 -e50 -solexa1.3-quals). For additional information about the Af data, see Supplementary Table 3.

1.4 *Mus musculus* (Mm) miRNA-seq in muscle stem cells

These data are related to a transcriptome study where the expression of miRNAs was measured in three different cellular stages of the skeletal muscle lineage in adult mouse. Mouse muscle satellite cells were isolated from adult Tg: Pax7-nGFP mice as previously described [9]. Briefly, limb muscles were dissected and digested in 0.1% Collagenase and 0.25% Trypsin [2], then sorted based on GFP fluorescence using a FACSAria (BD Biosciences). Cells were either directly sorted into TRIzol®LS buffer (Invitrogen, CA) or plated on Matrigel coated dishes in 1: 1 DMEM to MDCB containing 20% fetal calf serum, Insulin-Transferin-Selenium and 2% Ultroser [10] for 60H (Activated myoblasts) or 7 Days (Differentiated cells)

before RNA extraction using TRIzol®. To allow differentiation the culture was maintained for 7 days without changing the medium. Total RNA corresponding to the 3 cellular states (quiescent satellite cells, activated myoblasts, and differentiated myotubes) were purified using the miRneasy kit (Qiagen, Germany) following the manufacturer's instructions. Small RNAs were cloned using the DGE-Small RNA Sample Prep Kit and the Small RNA Sample Prep v1.5 Conversion Kit from Illumina, following manufacturer instructions. Libraries were sequenced using the Illumina Genome Analyzer II. Small RNA sequence files from in-house libraries were obtained and processed as fastq format files. Sequence reads were trimmed from the adapter sequence and matched to the *M. musculus* genome release NCBI37/Mm9 using novoalign (http://www.novocraft.com). Only 19-29nt reads matching the genome with 0 or 1 mismatch were retained for subsequent analysis. For annotations, we used the release 14 fasta reference files available in Mirbase (http://www.mirbase.org). Small RNA mapping were generated using in-house Perl software to parse novoalign outputs (code available on request). For additional information about the Mm data, see Supplementary Table 4.

2 Alternative normalization methods

Housekeeping-gene normalization: The housekeeping-gene approach borrows the idea from standard laboratory procedures (e.g. Northern blot or quantitative RT-qPCR), where an internal control is used for data normalization. It assumes that some genes are similarly expressed across samples/lanes, so that they can be used as a reference for the relative expression levels of other genes. However, there is a serious concern about the assumption of invariant expression of the so-called housekeeping genes as they are often affected by various factors that are not controlled in the experiment. Also, those genes are usually highly expressed, thus not representing genes of low intensities. Furthermore, they are usually a very small subset, so fluctuations in their intensities are highly affected by random or systematic errors. Any normalization based on such a limited number of internal references would be unreliable. Finally, housekeeping-gene normalization depends on the a priori knowledge of genes with stable expression levels, which may not always be feasible in practice. Therefore, normalization based on housekeeping genes selected a priori is not recommended and hence is not included in our comparison study.

GC-content normalization: Some normalization methods have recently been proposed to correct for the bias associated with the GC content of genes [11, 12]. Indeed, Pickrell et al. have suggested that this bias is sample-specific and should be corrected prior to a differential analysis [13]. We did not include such a normalization strategy in our study because a close inspection of our datasets did not confirm the presence of such a bias (Supplementary Fig. 13). As such, we assume that the GC bias associated with each gene is constant across conditions, and does not need to be corrected in the context of a differential analysis. Finally, it has been suggested [14] that any bias associated with the GC content should be taken into account at the quantification level, prior to the normalization step.

3 `R` code

The `R` scripts and functions used for the normalization and differential analysis of the *Entamoeba histolytica* dataset (using the `DESeq` Bioconductor package) are provided in `R` files, as well as the `R` script used for the simulations.

4 List of supplementary figures

Supp Figure 1. Boxplots of $\log_2$(normalized values + 1) for all conditions and replicates for the four datasets, by normalization method.

Supp Figure 2. Boxplots of intra-group variance within each condition for the four datasets, by normalization method.

Supp Figure 3. MA-plots for the four datasets, by normalization method. The log-ratio (M) is plotted

against the mean intensity (A) for differentially expressed genes. Black dots correspond to genes that are detected as DE by all normalization methods; the other (colored) dots represent genes that are specifically detected as DE by a single normalization method.

Supp Figure 4. Clustering trees of differential analysis results (using the `DESeq` Bioconductor package), for all normalization methods, for each of the four datasets.

Supp Figure 5. Comparison of normalization methods for simulated data (using the `DESeq` Bioconductor package) under scenarios 2 (equal library sizes and no majority genes) and 3 (non-equivalent library sizes and no majority genes). (a) Average type I error rate over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 0% to 30% within each color grouping). (b) Average power for the same set of simulated data.

Supp Figure 6. Relationship between gene length and differential expression (using the `DESeq` Bioconductor package), by normalization method, for the *A. fumigatus* data. Blue lines represent loess curves.

Supp Figure 7. Relationship between gene length and differential expression (using the `DESeq` Bioconductor package), by normalization method, for the *E. histolytica* data. Blue lines represent loess curves.

Supp Figure 8. Relationship between gene length and differential expression (using the `DESeq` Bioconductor package), by normalization method, for the *H. sapiens* data. Blue lines represent loess curves.

Supp Figure 9. Scaling factors for the the UQ (green) and Med (blue) normalization methods for each of the four datasets. Replicates from different conditions are differentiated by the shaded areas in the background.

Supp Figure 10. Median (blue) and upper quartile values (green) for each sample in each of the four datasets. Replicates from different conditions are differentiated by the shaded areas in the background. Blue and green horizontal lines represent mean values for the median and upper quartile, respectively, across all samples within each dataset.

Supp Figure 11. (Left) Tail-area histogram of $\log_2$(counts + 1) for counts greater than the upper quartile in sample B2 of the *M. musculus* data, which had a small difference between the Med and UQ scaling factors. (Middle) Tail-area histogram of $\log_2$(counts + 1) for counts greater than the upper quartile in sample C2 of the *M. musculus* data, which had a large difference between the Med and UQ scaling factors. (Right) Density plots for all gene counts from samples B2 (light blue) and C2 (dark blue) in the *M. musculus* data.

Supp Figure 12. (Top left) Log_2(counts + 1) versus normalized ranks for the *M. musculus* data, where per-gene means are represented by the black line. Each sample is represented by the same color throughout the figure. (Top right) Difference between $\log_2$(counts + 1) and $\log_2$(normalized counts + 1) by normalized ranks for the *M. musculus* data. (Bottom left) Number of genes with 0 counts for each sample in the *M. musculus* data. (Bottom right) Log_2(normalized values + 1) for the *M. musculus* data following quantile normalization.

Supp Figure 13. Log_2-transformed counts versus GC-content for each sample in the *E. histolytica* (top left), *H. sapiens* (top right), and *M. musculus* (bottom left) data. Replicates in each condition are represented by the same color, as indicated in the legends.

Supp Figure 14. Comparison of scaling factors calculated for each sample using the TC, UQ, Med, TMM, and DESeq normalization methods, for each of the four datasets: *E. histolytica* (top left), *M. musculus* (top right), *H. sapiens* (bottom left), and *A. fumigatus* (bottom right). Grey background panels distinguish the experimental conditions for each sample (i.e., for the *E. histolytica* data, samples 1-3 belong to one experimental condition, and samples 4-6 to another).

Supp Figure 15. Consensus dendrogram of differential analysis results, using the TSPM method, for all normalization methods across the four datasets under consideration.

References

1. Strub T, Giuliano S, Ye T et al. Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene* 2011; 30:2319–2332.

2. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010; 26(7):873–81.

3. Grant GR, Farkas MH, Pizarro AD et al. Comparative analysis of RNA-seq alignment algorithms and the RNA-seq unifed mapper (RUM). *Bioinformatics* 2011; 27(18):2518–2528.

4. Aurrecoechea C, Barreto A, Brestelli J et al. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res* 2011; 39:D612–619.

5. Langmead B, Trapnell C, Pop M et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; 10(R25).

6. Dimon MT, Sorber K, DeRisi JL HMMSplicer: A tool for efficient and sensitive discovery of known and novel splice junctions in RNA-seq data. *PLoS ONE* 2010; 5(11):e13875.

7. Fedorova ND, Khaldi N, Joardar VS et al. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet* 2008; 4(4):e1000046.

8. Jagla B, Wiswedel B, JY C. Extending knime for next-generation sequencing data analysis. *Bioinformatics* 2011; 27(20).

9. Sambasivan R, Gayraud-Morel B, Dumas G et al. Distinct regulatory cascades govern extraocular and pharyngeal arch muscle progenitor cell fates. *Dev Cell* 2009; 16(6):810–821.

10. Gayraud-Morel B, Chrétien F, Flamant P et al. A role for the myogenic determination gene Myf5 in adult regenerative myogenesis. *Dev Biol* 2007; 312(1):13–28.

11. Risso D, Schwartz K, Sherlock G et al. GC-content normalization for RNA-seq. *BMC Bioinformatics* 2011;12:480.

12. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-Seq data using conditional quantile normalization. *Biostatistics* 2012; 13(2):204-216.

13. Pickrell JK, Marioni JC, Pai A et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010; 464(7289):768–772.

14. Roberts A, Trapnell C, Donaghey J et al. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol* 2010; 12(3):R22.

5 Supplementary tables

Condition	Replicate	Total number of reads	Number of mapped reads	% mapped reads
A	1	27,193,572	23,176,430	98.40
A	2	35,606,942	26,610,393	98.62
B	1	30,891,601	28,471,528	96.04
B	2	30,455,011	22,659,643	97.87
A	3	28,692,182	27,640,912	98.39
B	3	31,580,414	29,752,354	98.20

Table 1: Data characteristics for each sample in the *H. sapiens* data, including the total number of reads, the number of mapped reads, and the percent of mapped reads.

Condition	Replicate	Total number of reads	Number of mapped reads	% mapped reads
A	1	67,450,000	64,010,050	94.90
A	2	65,200,000	61,222,800	93.90
A	3	98,900,000	90,493,500	91.50
B	1	100,100,000	94,594,500	94.50
B	2	84,950,000	68,384,750	80.50
B	3	77,200,000	72,876,800	94.40

Table 2: Data characteristics for each sample in the *E. histolytica* data, including the total number of paired reads, the number of mapped paired reads, and the percent of mapped paired reads.

Condition	Replicate	Total number of reads	Number of mapped reads	% mapped reads
A	1	22,185,444	17,068,739	76.94
B	1	80,086,374	41,061,241	51.27
A	2	64,450,566	12,232,921	18.98
B	2	8,938,278	6,561,361	73.41

Table 3: Data characteristics for each sample in the *A. fumigatus* data, including the total number of reads, the number of mapped reads, and the percent of mapped reads.

Condition	Replicate	Total number of reads	Number of mapped reads	% mapped reads
A	1	5,293,970	4,266,749	80.60
B	1	5,248,466	4,745,649	90.42
A	2	4,927,435	4,255,355	86.36
B	2	5,114,680	4,581,687	89.58
A	3	4,635,284	3,924,984	84.67
C	1	3,472,791	2,742,337	78.97
C	2	5,711,974	4,358,146	76.30

Table 4: Data characteristics for each sample in the *M. musculus* data, including the total number of reads, the number of mapped reads, and the percent of mapped reads.

DESeq								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	52	50	43	41	43	50	43	0
UQ		51	42	40	42	49	44	0
Med			43	41	42	42	37	0
DESeq				44	44	43	36	0
TMM					46	45	38	0
Q						72	51	0
RPKM							52	0
RC								0
TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	1041	907	676	669	707	333	79	4
UQ		996	707	663	700	344	75	4
Med			933	818	825	345	87	5
DESeq				992	934	346	83	5
TMM					992	356	82	5
Q						576	53	4
RPKM							534	0
RC								10

Table 5: Number of differentially expressed genes (using the `DESeq` Bioconductor package and TSPM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *A. fumigatus* data. Counts along the diagonal indicate the number of DE genes per method (i.e., 52 DE genes for the TC method using DESeq, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 50 DE genes in common between TC and UQ using DESeq).

DESeq								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	2,362	2,229	1,986	2,296	2,332	2,251	944	1,656
UQ		2,367	1,902	2,306	2,261	2,284	932	1,660
Med			2,496	1,935	2,030	1,991	864	1,670
DESeq				2,379	2,307	2,283	943	1,660
TMM					2,405	2,275	942	1,668
Q						2,488	943	1,660
RPKM							950	799
RC								1,68
TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	628	503	430	556	557	441	282	197
UQ		728	447	567	536	551	370	168
Med			658	430	471	388	306	227
DESeq				632	573	488	301	173
TMM					627	453	273	203
Q						848	369	124
RPKM							2199	18
RC								296

Table 6: Number of differentially expressed genes (using the `DESeq` Bioconductor package and TSPM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *H. sapiens* data. Counts along the diagonal indicate the number of DE genes per method (i.e., 2,362 DE genes for the TC method using DESeq, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 2,229 DE genes in common between TC and UQ using DESeq).

DESeq								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	231	207	225	224	230	220	231	185
UQ		250	214	231	224	237	227	193
Med			242	234	235	226	232	186
DESeq				251	242	241	241	193
TMM					251	236	246	191
Q						258	238	193
RPKM							263	195
RC								195
TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	229	169	200	200	212	195	172	109
UQ		252	201	218	205	205	126	93
Med			255	236	240	219	153	92
DESeq				256	241	226	153	98
TMM					257	223	161	103
Q						259	155	95
RPKM							180	86
RC								127

Table 7: Number of differentially expressed genes (using the `DESeq` Bioconductor package and TSPM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *M. musculus* data (A vs. C). Counts along the diagonal indicate the number of DE genes per method (i.e., 231 DE genes for the TC method using DESeq, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 207 DE genes in common between TC and UQ using DESeq).

DESeq								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	270	65	96	74	83	103	266	109
UQ		90	86	87	86	89	70	68
Med			117	95	101	110	101	88
DESeq				96	94	95	79	72
TMM					104	101	88	82
Q						132	109	93
RPKM							286	115
RC								116
TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	129	36	41	43	44	49	32	50
UQ		110	74	91	89	58	21	27
Med			108	86	84	76	22	29
DESeq				104	94	65	22	30
TMM					100	67	23	30
Q						104	15	33
RPKM							53	21
RC								65

Table 8: Number of differentially expressed genes (using the `DESeq` Bioconductor package and TSPM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *M. musculus* data (B vs. C). Counts along the diagonal indicate the number of DE genes per method (i.e., 270 DE genes for the TC method using DESeq, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 65 DE genes in common between TC and UQ using DESeq).

DESeq								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	250	62	44	43	60	80	259	191
UQ		90	71	69	87	89	62	66
Med			81	68	70	77	44	48
DESeq				69	69	68	43	47
TMM					87	86	60	64
Q						125	80	83
RPKM							275	191
TC								195
TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	276	43	37	53	47	79	151	116
UQ		69	51	59	65	57	34	30
Med			71	62	56	57	32	26
DESeq				89	66	69	44	40
TMM					73	62	37	34
Q						122	56	61
RPKM							161	75
TC								119

Table 9: Number of differentially expressed genes (using the `DESeq` Bioconductor package and TSPM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *M. musculus* data (A vs. B). Counts along the diagonal indicate the number of DE genes per method (i.e., 259 DE genes for the TC method using DESeq, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 62 DE genes in common between TC and UQ using DESeq).

TSPM								
	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	63	44	44	38	42	44	37	13
UQ		240	233	210	201	179	108	15
Med			253	224	197	184	113	15
DESeq				258	191	171	108	16
TMM					206	167	97	16
Q						218	105	14
RPKM							576	6
TC								24

Table 10: Number of differentially expressed genes (using the TPSM) found in common for each of the normalization methods, as well as the unnormalized raw counts (RC), in the *E. histolytica* data. Counts along the diagonal indicate the number of DE genes per method (i.e., 63 DE genes for the TC method, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e., 44 DE genes in common between TC and UQ).

Annexe F

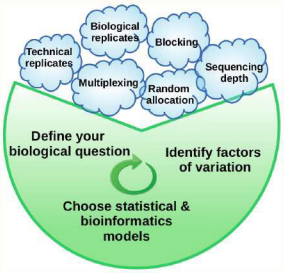
Poster de bonnes pratiques de planification d'expériences de RNA-Seq mentionné dans la section 5.2.1

How to Design a good RNA-Seq experiment in an interdisciplinary context?

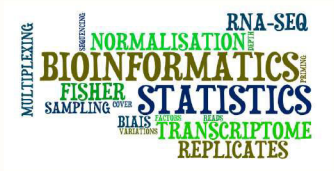
Pôle Planification Expérimentale, PEPI IBIS [1],
[1]INRA, France

RNA-seq technology is a powerful tool for characterizing and quantifying transcriptome. Upstream careful experimental planning is necessary to pull the maximum of relevant information and to make the best use of these experiments.

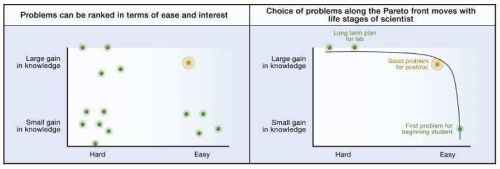
An RNA-seq experimental design using Fisher's principles



Rule 1: Share a minimal common language



Rule 2: Well define the biological question



From Alon, 2009

- Choose scientific problems on feasibility and interest
- Order your objectives (primary and secondary)
- Ask yourself if RNA-seq is better than microarray regarding the biological question

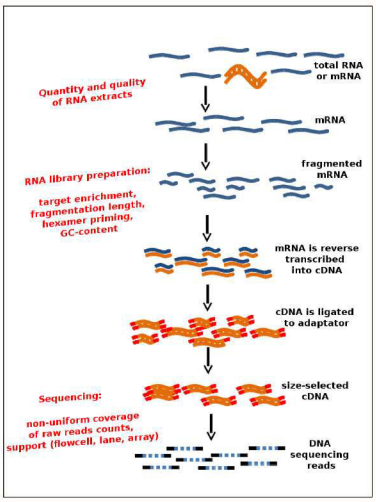
Make a choice

- Identify differentially expressed (DE) genes?
- Detect and estimate isoforms?
- Construct a de Novo transcriptome?

Rule 3: Anticipate difficulties with a well designed experiment

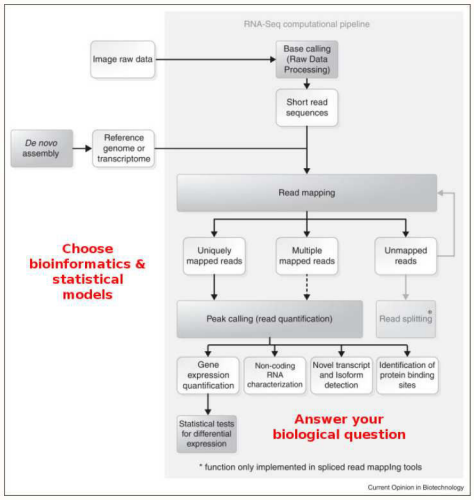
1. Prepare a checklist with all the needed elements to be collected,
2. Collect data and determine all factors of variation,
3. Choose bioinformatics and statistical models,
4. Draw conclusions on results.

Be aware of different types of bias



Keep in mind the influence of effects on results: lane $\leq$ run $\leq$ RNA library preparation $\leq$ biological (Marioni, 2008), (Bullard, 2010)

RNA-seq experiment analysis: from A to Z



Adapted from Mutz, 2013

Rule 4: Make good choices

How many reads?

- 100M to detect 90% of the transcripts of 81% of human genes (Toung, 2011)
- 20M reads of 75bp can detect transcripts of medium and low abundance in chicken (Wand, 2011)
- 10M to cover by at least 10 reads 90% of all (human and zebrafish) genes (Hart, 2013)...

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate to a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)

More biological replicates or increasing sequencing depth?

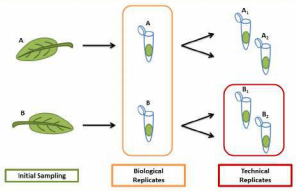
It depends! (Haas, 2012), (Liu, 2014)

- DE transcript detection: (+) biological replicates
- Construction and annotation of transcriptome: (+) depth and (+) sampling conditions
- Transcriptomic variants search: (+) biological replicates and (+) depth

A solution: multiplexing.
Decision tools available: Scotty (Busby, 2013), RNAseqPower (Hart, 2013)

Some definitions

Biological and technical replicates:



Sequencing depth: Average number of a given position in a genome or a transcriptome covered by reads in a sequencing run
Multiplexing: Tag or bar coded with specific sequences added during library construction and that allow multiple samples to be included in the same sequencing reaction (lane)
Blocking: Isolating variation attributable to a nuisance variable (e.g. lane)

Conclusions

- Clarify the biological question
- All skills are needed to discussions right from project construction
- Prefer biological replicates instead of technical replicates
- Use multiplexing
- Optimum compromise between replication number and sequencing depth depends on the question
- Wherever possible apply the three Fisher's principles of randomization, replication and local control (blocking)

And do not forget: budget also includes cost of biological data acquisition, sequencing data backup, bioinformatics and statistical analysis.

Who are we?

julie.aubert@agroparistech.fr, anne.delafoye@clermont.inra.fr, **cyprien.guerin@jouy.inra.fr**, **christelle.hennequet@tours.inra.fr**, frederique.hilliou@sophia.inra.fr, fabrice.legeai@rennes.inra.fr, delphine.labourdette@insa-toulouse.fr, nmarsaud@insa-toulouse.fr, brigitte.schaeffer@jouy.inra.fr

ECCB14, 7 - 10 September 2014, Strasbourg, France

Annexe G
Article sur l'étude des plans en dye-switch présenté dans la section 5.2.2

BMC Bioinformatics

Methodology article

Open Access

Statistical methodology for the analysis of dye-switch microarray experiments

Tristan Mary-Huard*[1], Julie Aubert[1], Nadera Mansouri-Attia[2], Olivier Sandra[2] and Jean-Jacques Daudin[1]

Address: [1]UMR AgroParisTech/INRA 518, 16, rue Claude Bernard 75231 Paris CEDEX 05, France and [2]UMR INRA/ENVA/CNRS 1198, Jouy en Josas, France

Email: Tristan Mary-Huard* - maryhuar@agroparistech.fr; Julie Aubert - julie.aubert@agroparistech.fr; Nadera Mansouri-Attia - nadera.mansouri@jouy.inra.fr; Olivier Sandra - olivier.sandra@jouy.inra.fr; Jean-Jacques Daudin - daudin@agroparistech.fr

* Corresponding author

Published: 13 February 2008

BMC Bioinformatics 2008, **9**:98 doi:10.1186/1471-2105-9-98

Received: 25 June 2007
Accepted: 13 February 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/98

Abstract

Background: In individually dye-balanced microarray designs, each biological sample is hybridized on two different slides, once with *Cy3* and once with *Cy5*. While this strategy ensures an automatic correction of the gene-specific labelling bias, it also induces dependencies between log-ratio measurements that must be taken into account in the statistical analysis.

Results: We present two original statistical procedures for the statistical analysis of individually balanced designs. These procedures are compared with the usual ML and REML mixed model procedures proposed in most statistical toolboxes, on both simulated and real data.

Conclusion: The UP procedure we propose as an alternative to usual mixed model procedures is more efficient and significantly faster to compute. This result provides some useful guidelines for the analysis of complex designs.

Background

DNA microarray technology is a high throughput technique by which the expression of the whole genome is studied in a single experiment. Experiments must be well organized and design issues are crucial, see [1,2]. In dual label experiments Cy3 and Cy5 are used as fluorescent dyes allowing to compare two RNA samples on the same slide. It is now well known that there exists a differential effect of the two dyes [3,4], that can be gene-specific. An efficient way to remove this technical artifact is to use balanced reverse dye designs [5]. Balanced reverse dye designs can be divided into three classes along a line of strengthening balancing constraints:

1. Balanced reverse dyes for which each biological sample is hybridized only one time and therefore present with only one dye, on only one array (Table 1.1). These designs are *globally balanced* but not individually balanced.

2. *Individually-balanced* design for which each biological sample is divided into two parts, one hybridized with *Cy*3 on one array and the other with *Cy*5 on another array. Each biological sample is hybridized exactly two times (Table 1.2).

3. Dye-swaps for which each couple of biological samples from two conditions are hybridized on two arrays with

reverse dyes. Dye-swaps are constrained to be *couple-balanced* (Table 1.3).

Dye-swap design is mostly used when the technical error is higher than the biological variability, either to reduce the technical variance, or when gene-specific dye-bias is of concern [6,7]. When the biological variability is greater than the technical error, *globally balanced* designs are statistically more efficient [5]. However the number of biological samples is sometimes limited, therefore this design is not always possible in practice.

The term *Dye-switch* is used for the first and sometimes also for the second classes. Dye-switch designs of the second class are sometimes described and proposed in papers dealing with microarrays experiments. For example loop designs are often members of this class [8,9], although the distinction between the first and the second class is not always clearly made.

A major point to notice is that the statistical analysis may be very different for the three classes of design. The analysis of the first and third classes is straightforward and well described in articles (see for example [4,10,11]): the experimental units are mutually independent (we consider as usual that the two conjugate arrays of the dye-swaps are summed up to only one experimental unit), and simple statistical procedures such as Student *T*-tests (or regularized *T*-tests) can be performed. On the contrary, if we consider the second class of designs, the experimental units are not independent, a feature that must (or must not) be accounted for. The literature about the statistical study of such designs is limited: some papers proposed some theoretical contributions for their analysis [12,13], but simple guidelines for experimenters and practical considerations (computational burden, choice of a strategy for parameter estimation) are not available.

Table 1: Three different balanced reverse dye designs for the comparison of 2 treatments

1	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B5	A3	B9	A5	B6	A7	B10	A9	B9
	Cy3	B3	A2	B8	A4	B2	A6	B1	A8	B4	A10
2	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
	Cy3	B1	A2	B2	A3	B3	A4	B4	A5	B5	A1
3	array	1	2	3	4	5	6	7	8	9	10
	Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
	Cy3	B1	A1	B2	A2	B3	A3	B4	A4	B5	A5

Three different balanced reverse dye designs for the comparison of 2 treatments (A and B), with an equal number of slides. A_i stands for the i^{th} biological sample in condition A. (1) Globally balanced design, with 10 biological samples per condition. (2) Individually-balanced design with 5 biological samples per condition. (3) Dye-swap design with 5 biological samples per condition.

We consider here the simplest individually-balanced dye-switch design: two conditions *A* and *B* are compared in a two-color cDNA microarray experiment, with *n* biological samples for each condition. The design is the following: each RNA sample (A_1 to A_n for condition *A*, and B_1 to B_n for condition *B*) is divided into two parts, one labelled with $C\gamma5$ and the second labelled with $C\gamma3$. Then $2n$ microarrays are hybridized with respectively $A_1C\gamma5$ and $B_1C\gamma3$, $B_1C\gamma5$ and $A_2C\gamma3$, $A_2C\gamma5$ and $B_2C\gamma3$, and so on till $B_nC\gamma5$ and $A_1C\gamma3$, (see Table 1.2). There are $2n$ samples, $4n$ labelled samples, $2n$ microarrays, and each sample is hybridized two times (one with $C\gamma5$ and one with $C\gamma3$) on two different arrays. We propose a simple, efficient and robust method for the statistical analysis of this experiment.

Model on the measure of the expression of genes

After the normalization step, X_i is the expression measure on the log-scale, for a given gene, corresponding to condition *A* on array *i*. Let $j(i)$ denote the sample number corresponding to condition *A* and array *i*.

Similarly, Y_i is the expression measure for the condition *B* sample on the same array, and $j'(i)$ the sample number corresponding to condition *B* and array *i*. In the following the gene index is not present in order to simplify the mathematical expressions, but it is important to note that all the terms in the following models are gene-specific. Here we use an analysis of variance (ANOVA) model for the expression measure as introduced by [10].

The model for X_i and Y_i is the following:

$$X_i = \boldsymbol{m}_A + \boldsymbol{d}_{l(i)} + B_{j(i)} + M_i + T_i$$
$$Y_i = \boldsymbol{m}_B + \boldsymbol{d}_{l'(i)} + B_{j'(i)} + M_i + T'_i,$$

where

• $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are the population mean expression measures for condition *A* and *B*.

• $\boldsymbol{\delta}_{l(i)}$ is a two-level fixed effect corresponding to the dye effect. $\boldsymbol{\delta}_{l(i)} = \boldsymbol{\delta}_1$ (resp. $\boldsymbol{\delta}_2$) for all the samples labelled with Cy5 (resp. Cy3). This term accounts for the *gene-specific dye bias*.

• $B_{j(i)}$ represents an independent gaussian random term with mean 0 and standard deviation $\boldsymbol{\sigma}_B$, corresponding to the random effect of sample $j(i)$. This variable is specific to the biological sample and is called *biological error*, related

to the variability of the biological material inside each population A and B.

• M_i represents an independent gaussian random term with mean 0 and standard deviation σ_M. M_i is the effect of the spot associated to the gene under concern in microarray *i* and has the same value for the two samples which are hybridized on array *i*. This error term takes into account the spatial heterogeneity in each array that affects both *Cy*3 and *Cy*5 measurements.

• T_i represents an independent gaussian random term with mean 0 and standard deviation σ_T, corresponding to the technical variability, including the steps of labelling, hybridization and measure of intensity of fluorescence. This variable has a specific value for each combination gene×dye×sample, even if the samples are hybridized on the same array and at the same spot, so that T_i and T_i' are independent random variables. T_i and M_i are the two components of the so-called *technical error*.

Model on the difference of expression on one array

Let $D_i = X_i - Y_i$, $i = 1,...,2n$. Using equation (1) we obtain:

$$D_i = \mu_A - \mu_B + B_{j(i)} - B_{j'(i)} + \delta_{l(i)} - \delta_{l'(i)} + T_i - T_i'$$

which may be written

$$D_i = \mu + BD_i + (-1)^{i+1}\,\delta + TD_i$$

where

• $\mu = \mu_A - \mu_B$ is the true differential expression between conditions *A* and *B* for the gene under concern,

• $BD_i = B_{j(i)} - B_{j'(i)}$ is a random variable with mean 0 and standard deviation $\sqrt{2}\,\sigma_B$,

• $TD_i = T_i - T_i'$ is an independent random variable with mean 0 and standard deviation $\sqrt{2}\,\sigma_T$,

• $\delta = \delta_1 - \delta_2$ is the difference between the Cy3 and Cy5 dye effects. This term accounts for the *gene-specific dye bias*.

Each variable D_i follows a Gaussian distribution with mean $E(D_i) = \mu + (-1)^{i+1}\delta$ and variance $V(D_i) = 2\sigma_B^2 + 2\sigma_T^2$. All the covariances $\mathrm{cov}(D_i, D_j)$ are equal to zero except the following ones:

$$\mathrm{cov}(D_i, D_{i+1}) = \sigma_B^2$$

with the convention that $2n + 1 = 1$.

In this study, we present and compare different strategies for the statistical analysis of individually-balanced designs. The article is organized as follows. In the Results section, five statistical procedures to analyze individually balanced designs (Table 1.2) are compared on both simulated and real data. The Conclusion section draws the main conclusions and gives some useful guidelines for the analysis of individually-balanced designs. The details of the computation are given in the Methods section.

Results

Statistical procedure comparison

In this section, we investigate the efficiency of five test procedures for the differential analysis of datasets corresponding to the design of Table 1.2. The procedures are the following (see the Methods section for more details):

• Naive Method NM: for each gene, the naive test statistic

$$T_N = \sqrt{2n}\,\frac{\bar{D}}{\sqrt{S^2}}$$

is computed.

• Unbiased Paired Method (UP): for each gene, the unbiased paired statistic

$$T_{UP} = \sqrt{2n}\,\frac{\bar{D}}{\sqrt{(S^2+2C)}}$$

is computed. Notice that from the Methods section, the theoretical value of *C* must be positive. In practice, the estimated value may be negative. In such a case, *C* is truncated at 0.

• Unbiased Unpaired Method (UU): for each gene, the unbiased unpaired statistic

$$T_{UU} = \sqrt{n}\,\frac{\bar{X}-\bar{Y}}{\sqrt{S_X^2+S_Y^2-2C_{XY}}}$$

is computed. As for the previous method, the value of C_{XY} must be positive. If not, C_{XY} is truncated at 0. Furthermore, the unbiased variance estimator is $S_X^2 + S_Y^2 - 2C_{XY}$. Since C_{XY} is non-negative, the variance estimator may have a negative value. In such a case, the variance can be fixed at a given threshold (0.001 in the following).

• Mixed Model with ML estimation (ML): for each gene, model (1) is adjusted with the Maximum Likelihood algorithm.

• Mixed Model with REML estimation (REML): for each gene, model (1) is adjusted with the Restricted Maximum Likelihood algorithm.

It is important to consider both the ML and REML algorithms for the mixed model since each algorithm has its own advantages. While ML is known to provide biased estimates of the variance components, computations are faster and the algorithm does converge. REML gives unbiased estimates of the parameters, but may not converge if the number of observations is small. Both ML and REML computations were performed using the R package Maanova [10].

Simulations

To study the behavior of the 5 procedures, we performed a simulation study using model (1). We considered 3 different values for σ_B^2 (0.5, 1, 2) and σ_M^2 (1, 2, 5), 4 values for the number of samples in one condition (5, 10, 20, 30) and 5 possible values for the differential expression $\mu = \mu_A - \mu_B$ (0, 1, 2, 3, 4). The parameter σ_T was fixed at 1. For each combination of the parameters, 10,000 genes were simulated.

Control of the Type I error rate

We first consider the case $\mu = 0$. Table 2 shows the actual Type I error rate level of the 5 test procedures, when the requested nominal level is 5%. Different behaviors can be observed: NM and ML result in a type I error rate higher than the nominal level, and procedure UU is conservative. UP results in an actual level that is close to the expected one, whatever the conditions. In most cases, REML enables an efficient control of the type I error. Yet, when the biological variability is high and the number of samples is low, REML yields a high type I error because of inconsistent estimations of the variance (see the next section). When $\sigma_B^2 = 2$ and $n = 5$, the discrepancy between the theoretical and the actual level is even worse for REML than for the other methods.

From these first observations we conclude that we can discard procedures NM and ML, since in differential analysis an effective control of the Type I error rate is necessary.

Performance analysis

We now compare the performance of the 3 remaining procedures to detect differentially expressed genes. Table 3 shows the proportion of detected differentially expressed genes, for different values of the parameter set. It clearly appears that the power of procedure UU is low compared with procedures UP and REML. This may be the consequence of the Student approximation (each test statistic is compared with the quantile of a Student distribution with $2n$ - 2 degrees of freedom), that could be more erroneous in the case of the UU statistic.

An interesting point is that UP results are more stable than the REML results. If we consider sample sizes n larger than 20, we observe that the absolute values of the approximate REML T-test range from 0 to 32, except for some genes where the absolute value is larger than 400. These outliers come from an erroneous estimation of the variance of the mean difference, that is evaluated to be (almost) 0. This does not happen with (UP) since the estimated variance is $\max(S^2, S^2 + 2C)$, i.e. the variance is overestimated to avoid outliers. Notice that despite this overestimation in many cases the power of UP is larger than the power of REML.

Computational burden and convergence

We now consider the important question of computational time for the 2 competitive procedures UP and REML. Since microarray experiments can involve hun-

Table 2: Actual level of the 5 test procedures in one simulation of 10 000 genes

	$\sigma_B^2 = 0.5$				$\sigma_B^2 = 2$			
Method	5	10	20	30	5	10	20	30
Naive	6.9 (0.2)	7.3 (0.2)	7.3 (0.2)	7.5 (0.2)	13.2 (0.3)	13.9 (0.3)	14.0 (0.3)	14.2 (0.3)
Unbiased Paired	5.2 (0.2)	5.2 (0.2)	5.2 (0.2)	5.3 (0.2)	8.2 (0.3)	6.9 (0.2)	6 (0.2)	5.8 (0.2)
Unbiased Unpaired	2.1 (0.1)	1.3 (0.1)	1.0 (0.1)	1 (0.1)	4.6 (0.2)	3.4 (0.2)	2.7 (0.1)	2.9 (0.2)
ML	8.5 (0.3)	8.6 (0.3)	8.3 (0.3)	8.3 (0.3)	12.5 (0.4)	11.1 (0.3)	9.9 (0.3)	9.8 (0.3)
REML	4.7 (0.2)	4.2 (0.2)	4.5 (0.2)	4.9 (0.2)	14.7 (0.4)	8.5 (0.3)	5.9 (0.2)	5.5 (0.2)

Actual mean level (standard error) of the 5 test procedures, for low ($\sigma_B^2 = 0.5$, left) and high ($\sigma_B^2 = 2$, right) values of biological variance, and different number of samples n in each condition (in column). The requested nominal threshold is 5%.

Table 3: Power of the UU, UP and REML test procedures

		$\mu = 1$			$\mu = 3$		
Nb Samples	s_B^2	UU	UP	REML	UU	UP	REML
5	0.5	5.6	13.6	10.6	55.5	92.1	86.75
5	2	2.8	5.0	12.95	17.9	29.4	34.75
10	0.5	13.2	39.3	33.97	77.8	100.0	99.64
10	2	3.5	7.8	9.06	45.0	63.5	63.06
20	0.5	35.0	80.1	78.13	98.8	100.0	100.0
20	2	7.3	14.5	13.93	82.6	94.8	94.53
30	0.5	51.9	95.5	95.05	100.0	100.0	100.0
30	2	12.1	22.5	21.74	96.2	99.6	99.53

Power (probability of rejecting H_0 × 100) of the different test procedures to detect a low ($\mu = 1$, left) or high ($\mu = 3$, right) differential expression.

dreds of thousands of genes, it becomes critical to use efficient algorithms for the statistical analysis of the data. Table 4 gives the user CPU time associated to each procedure for the complete analysis of 10,000 genes. While the computational time is constant whatever the condition for the (UP) procedure, (REML) is 8 to 330 times longer than (UP), depending on the number of samples.

Furthermore, REML can result in inconsistent estimates of the variance, as shown in the previous sections, or may not converge. Table 4 provides the number of genes for which the REML algorithm did not converge.

Embriogenomic data

The impact of pregnancy on the cattle endometrium transcriptom is investigated in [14]. In Mammals, the implantation of the embryo is a key event in the establishment of a pregnancy. A microarray experiment has been made to analyze the gene expression of the bovine pregnant endometrium and determine key pathways that control the endometrium physiology during the implantation process. The expression of 13300 genes in the endometrium of cows (n = 5) has been investigated. Only 5 animals were available for each condition so that the dye-switch design of Table 1.2 was used. Gene profiling has been established to analyze the impact of pregnancy by comparing the endometrium of cyclic (day 20 of cycle) versus pregnant animals (day 20 of pregnancy). In the following, the results of the five statistical procedures defined above are compared using this dataset.

Table 4: CPU times of procedures UP and REML

n	UP CPU	REML CPU	No REML CV
5	2.3	787	56.9
10	2.6	212	5
20	2.8	467	0
30	3.2	1046	0.16

User CPU time of procedures (UP) and (REML), for $\sigma^2 = 0.5$ and different numbers of samples. The last column provides the average number of genes for which REML did not converge.

The Venn diagram of Figure 1 shows the number of genes declared differentially expressed (DE) by 4 methods using the Bonferroni method with a 5% level. The UU method gives the least number of DE genes (4) and is not presented in the diagram. REML (which did not converge for 3 genes) gives the greater number of DE genes (93), among which 23 are also found by the other methods, and 70 are specifically found by REML (70 REML specific genes). 70 genes are found DE by ML (22 ML specific genes), and 58 by the naive method (9 Naive specific). Finally 33 genes are declared DE by UP, and all of them are also found by one, two or all of its competitors. Therefore UP provides the less discordant results. The higher number of DE genes obtained with the naive and the ML methods was expected, since it is known from the theory and the simulation study that these methods yield more false positives than the nominal risk. Figure 2 (right) shows that the ML and UP estimates of the standard error are coherent but that the ML estimate are lower than the ones obtained by the UP method. This point is in keeping with the statistical theory which assesses that the UP estimate of the variance is unbiased while the ML estimate has a negative bias.

The high number of DE genes specifically found with REML is odd. Figures 2 and 3 show that this comes from very low estimates of variance for some genes, so that these genes are declared DE not because the mean difference of expression between the two conditions is high,

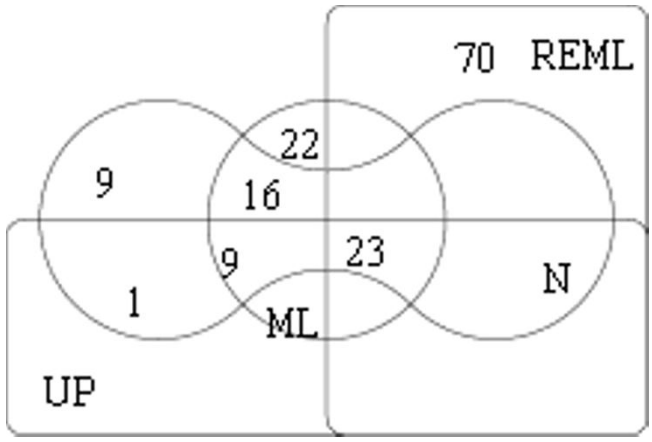


Figure 1
Venn diagram for the embriogenomics experiment. Comparison of the DE genes obtained by four methods. Vertical right rectangle: REML, horizontal low rectangle: UP, bone: N and circle: ML.

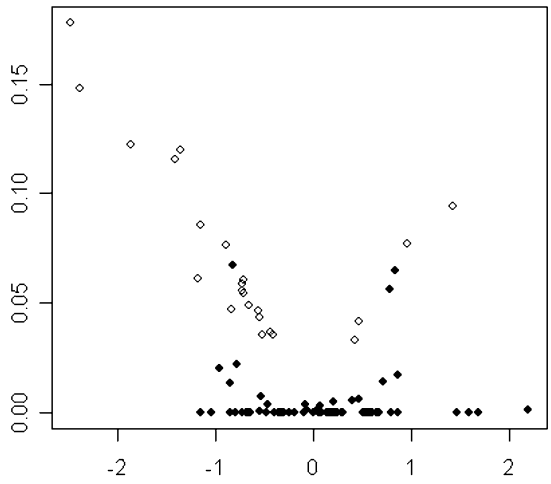


Figure 2
Comparison of the standard errors obtained with ML, REML and UP for the REML-DE genes of the embriogenomics experiment. **Left:** REML estimates (y-axis) versus UP estimates (x-axis) of the standard error. **Center:** REML estimates versus ML estimates. **Right:** UP estimates versus ML estimates.

but because this mean difference is divided by a very low standard error. So most of the 70 genes only found by the REML method are due to too low estimates of the gene variance obtained by the REML algorithm. This observation is in keeping with the results of the simulation study. Therefore the UP method or the naive method should be preferred in this particular experiment. The use of REML without a sharp biological analysis of the results gene by gene would be misleading.

Teleost fish dataset

An important application of the methodology proposed in the previous section is the analysis of loop design experiments. Loop and interwoven loop designs were initially proposed in [2] to compare p treatments, where p is 3 or higher. Figure 4 displays a particular interwoven loop design where 3 different 2-by-2 loop comparisons of treatments are combined in a single experiment. The 3 loop comparisons are

• $N1 \rightarrow S1 \rightarrow N3 \rightarrow S3 \rightarrow N5 \rightarrow S5 \rightarrow N2 \rightarrow S2 \rightarrow N4 \rightarrow S4 \rightarrow N1$

• $S1 \rightarrow G1 \rightarrow S3 \rightarrow G3 \rightarrow S5 \rightarrow G5 \rightarrow S2 \rightarrow G2 \rightarrow S4 \rightarrow G4 \rightarrow S1$

• $N1 \rightarrow G2 \rightarrow N3 \rightarrow G4 \rightarrow N5 \rightarrow G1 \rightarrow N2 \rightarrow G3 \rightarrow N4 \rightarrow G5 \rightarrow N1$

Each of these comparisons corresponds to the design of Table 1.2 discussed in the previous section. Such experimental designs have been studied both theoretically [15] and practically [8,9]. Here, we briefly present the Teleost fish data of [8].

The Teleost fish experiment aims to compare 3 populations of fish (Northern Fundulus heteroclitus, Southern Fundulus heteroclitus and Fundulus grandis). Five individuals were examined in each population to determine the variation in gene expression between populations. Each individual is used to probe four cDNA microarrays, according to the design of Figure 4. The raw data consist of 120 measurements (15 individuals × 4 slides × 2 duplicates per slide) for 907 genes.

In [8], the signal is modelled as follows (after per slide duplicate averaging):

$$Y_{ijkg} = m + A_i + D_j + (AD)_{ij} + G_g + (AG)_{ig} + (DG)_{jg} + (V\,G)_{kg} + e_{ijkg},$$

where A, D, G and V stand for Array, Dye, Gene and Variety, respectively. Then the 4 measurements corresponding to a given individual are averaged, and an F statistic is computed per gene to check whether the variety effect is significant or not.

This strategy roughly amounts to the UU test procedure of section when the number of treatments is higher than 2. The main difference is that in model (4), the model does not include the array random effect which takes into account the dependency between two measures on the same array. According to the results of section, this implies that the variance estimator is biased, leading to a loss of power.

As an alternative, we perform the statistical analysis using the UP procedure. Each pairwise comparison between 2 varieties is made, and a gene is declared differentially expressed if at least 2 of the 3 tests are significant. Each test is performed at the level 0.02, meaning that for a given gene, the nominal level is roughly 0.001 (3×0.02^2 for 2 of the 3 tests to be significant under H_0 at level 0.02). This is a good compromise between the 0.01 threshold adopted in the original articles with no correction for multiple testing, and the 0.5×10^{-4} (0.05/907) threshold given by a 5% level per test combined with a Bonferronni multiple testing correction. While the drawback of our strategy is to replace one test by three, the advantage is that the variance estimate is unbiased.

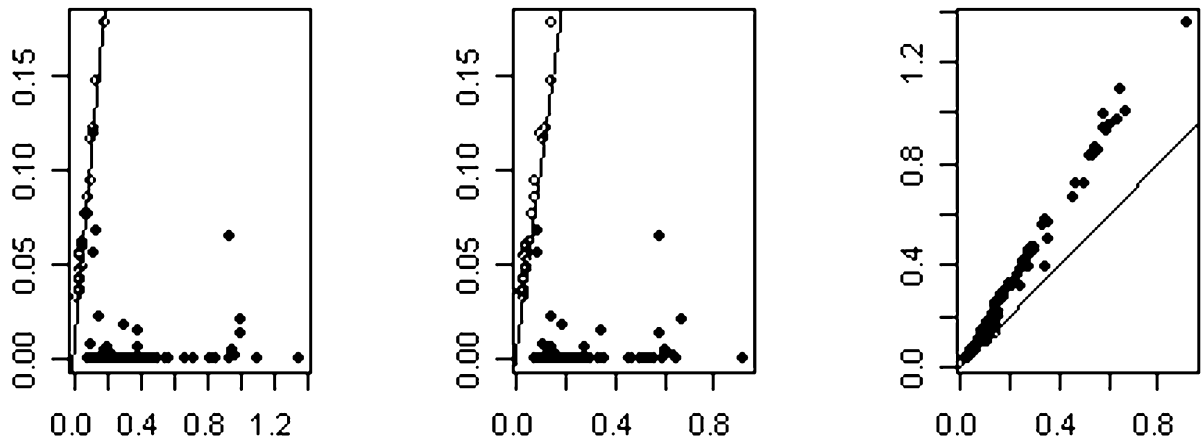


Figure 3
Mean difference versus standard error for the REML-differentially expressed genes of the embriogenomics experiment. Standard error of the difference obtained by REML (y-axis) versus mean difference between the two conditions (x-axis). Black points are not found DE by other methods than REML.

Table 5 gives the Oleksiak original list of differentially expressed genes found with the original method and the UP list of genes found with the UP procedure.

Among the 15 genes originally identified, 5 are also declared differentially expressed with the (UP) method. At a first glance, the (UP) procedure seems less powerful than the original method since only 9 genes are found here compared with the 15 genes of the original article. But due to the threshold adopted by the authors in [8], the expected number of false positives is 9 for the Oleksiak list, whereas for the (UP) list we expect only 1 false positive. Therefore most of the 10 extra genes found in [8] may be false positives. To examine the discriminant effect of the 9 genes of the (UP) list, we performed as in the original publication a clustering of the individuals, according to the significative genes. The corresponding tree is given in Figure 5. A cutoff of the tree at 0.15 gives the following 3 classes :

$$\{S1, S2, S4, S5, N1\},\ \{G1, G2, G3, G4, G5, S3\},\ \{N2, N5, N3, N4\}.$$

These 3 classes roughly correspond to the three populations of interest, up to 2 misclassified observations. In the original article, the partition in 3 classes gave

$$\{N1, N2, N3, N4, N5\},\ \{S1, S4\},\ \{G1, G2, G3, G4, G5, S2, S3, S5\}.$$

With only 9 genes (rather than 15), the classification obtained with (UP) is improved compared with the classification of the original method.

Discussion

Random terms taking into account the array and the sample effects must be included in the statistical model at the gene level for dye-switch experiments. We showed on simulations that the naive paired T-test, which does not take into account the biological sample effect, leads to more false positives than expected, especially when the biological sample effect is high. It may be safely used only when the biological variance is lower than the technical variance. The REML estimate for mixed model provides an approximatively correct false positive rate, at the price of high computational complexity, lack of convergence for low or medium sample sizes and sometimes spurious results. To the contrary, the UP method we propose is easy to implement and not computationally intensive. The method is protected against spurious results, leading to a more robust and powerful analysis than REML when the biological variability is high and the number of samples low, an usual situation in microarray experiments.

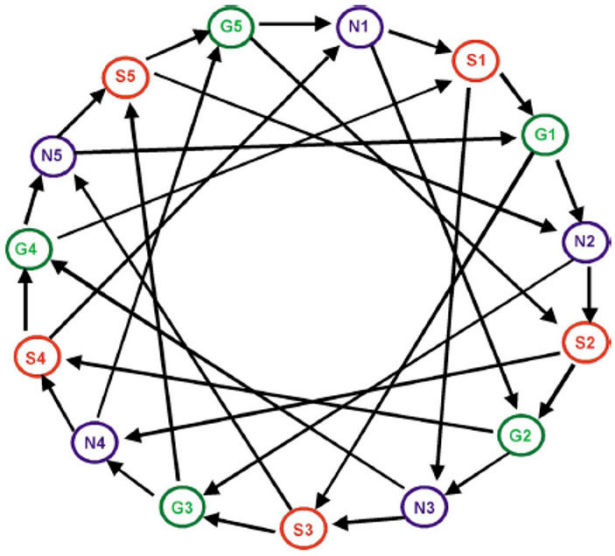


Figure 4
The Teleofish experiment design.

For small sample size experiments, it is advised to use regularized T-test, see [16-19]. Regularization strategies are based on statistical methods that take the individual variance of each gene as input and give a regularized variance for each gene as output. The UP procedure proposed in this paper gives an estimate for the variance of the differential expression for each gene, so it allows a further application of all these regularization methods.

Conclusion

In this paper the proposed estimate of the variance of the differential expression is assessed for the comparison between two conditions in a dye-switch design. The same methodology could be extended to more complex designs involving more than two conditions and duplicate hybridizations of the same biological sample on different arrays.

Methods

Paired test procedure

According to expression (2), an unbiased estimator of μ is $\bar{D} = \frac{1}{2n}\sum D_i$. The variance $V_{\bar{D}} = V(\bar{D})$ of this estimator is

$$\begin{aligned} V_{\bar{D}} &= V\left(\frac{1}{2n}\sum D_i\right) \\ &= \frac{1}{4n^2}\left[\sum V(D_i) + 2\sum \mathrm{cov}(D_i, D_{i+1})\right] \\ &= \frac{1}{4n^2}\left[2n(2\sigma_B^2 + 2\sigma_T^2) + 4n\sigma_B^2\right] \\ &= \frac{1}{2n}(4\sigma_B^2 + 2\sigma_T^2) \\ &= \frac{1}{2n}(V(D) + 2\,\mathrm{cov}(D_i, D_{i+1})) \end{aligned}$$

To perform a statistical test on parameter μ, we need to estimate $V_{\bar{D}}$.

Naive variance estimate

The naive estimate of $V_{\bar{D}}$ is

$$\frac{1}{2n-1}\sum_i (D_i - \bar{D})^2$$

which is used to perform paired *T*-tests. But in a dye-switch experiment the variables D_i - $\bar{D}$ are not centered, since the means of D_i and $\bar{D}$ are $\mu + (-1)^{i+1}\delta$ and μ, respectively. Hence we consider the alternative estimator

$$S^2 = \frac{1}{2n-2}\sum_i (D_i - \bar{D}_{(i)})^2,$$

where

$$\bar{D}_i = \frac{1}{n}\sum_{i \text{ is odd}} D_i \quad \text{if } i \text{ is odd},$$

$$\bar{D}_{(i)} = \frac{1}{n}\sum_{i \text{ is even}} D_i \quad \text{otherwise}.$$

Table 5: Lists of genes for the Teleofish experiment

Oleksiak list [8]	UP list
RAN GTP binding protein hypo P FLJ20727 ribosomal protein L27 dihydrolipoamide dehydrogenase GTP binding protein Steroidogenic acute regulatory protein hypo P FLJ11275 capping protein muscle Z line orla C4 surface glycoprotein HT7 precursor methionine adeno. regulatory Von Willebrand factor succinate dehydrogenase complex KIAA1481 protein protein disulfide isomerase annexin V	Thioredoxin nascent polypeptide associated dnaK type molec. chap. prec. ribosomal protein S16

Lists of genes whose expression was significantly different between populations. The first 5 genes are found differentially expressed by both methods.

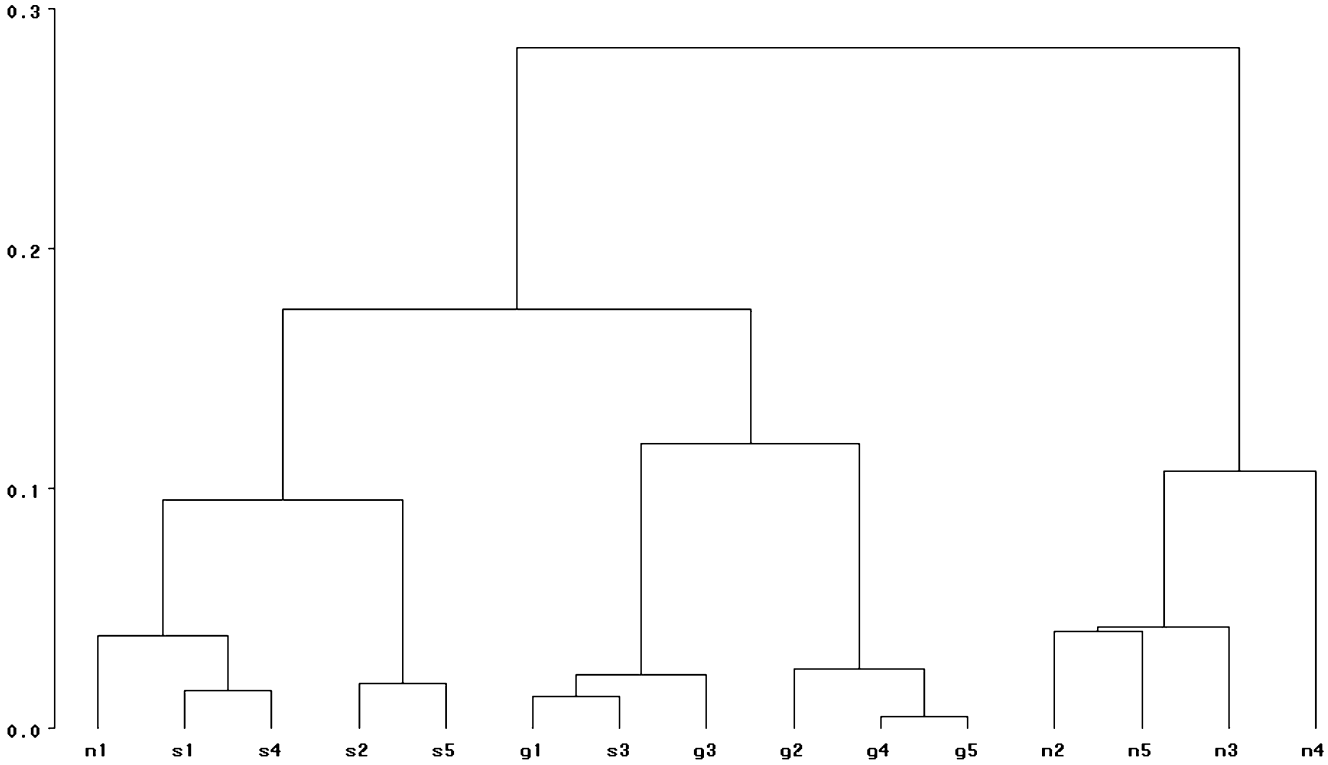


Figure 5
Clustering tree for the Teleofish dataset. Clustering tree for the Teleofish dataset, obtained from the second list of differentially expressed genes of Table 5.

The expectation of this alternative estimator is

$$\begin{aligned} E\left[\frac{1}{2n-2}\sum_i (D_i-\bar{D}_{(i)})^2\right] &= \frac{1}{2n-2}\sum_i\left[E(D_i^2)-E(\bar{D}_{(i)}^2)\right] \\ &= \frac{1}{2n-2}\sum_i\left[(2\sigma_B^2+2\sigma_T^2)-\frac{1}{n}(2\sigma_B^2+2\sigma_T^2)\right] \\ &= \frac{1}{2n-2}\times 2n\times\frac{n-1}{n}(2\sigma_B^2+2\sigma_T^2) \\ &= 2\sigma_B^2+2\sigma_T^2. \end{aligned}$$

S^2 is a downward biased estimator of $V(\bar{D})$. The higher σ_B^2 compared with σ_T^2, the higher the bias:

$$V(\bar{D}) = E(S^2)\left[1+\frac{\sigma_B^2/\sigma_T^2}{1+\sigma_B^2/\sigma_T^2}\right].$$

From this naive estimate of the variance we can derive a first T-test statistic to be used for the differential analysis:

$$T_N = \sqrt{2n}\,\frac{\bar{D}}{\sqrt{S^2}}$$

Unbiased variance estimate

Let $C = \frac{1}{2n-4}\sum_i (D_i-\bar{D}_{(i)})(D_{i+1}-\bar{D}_{(i+1)})$. We have

$$\begin{aligned} E(C) &= \frac{1}{2n-4}\sum_i E[D_iD_{i+1}-D_i\bar{D}_{(i+1)}-D_{i+1}\bar{D}_{(i)}+\bar{D}_{(i)}\bar{D}_{(i+1)}] \\ &= \frac{1}{2n-4}\sum_i\left[\sigma_B^2-\frac{2}{n}\sigma_B^2-\frac{2}{n}\sigma_B^2+\frac{1}{n^2}\times n\times 2\sigma_B^2\right] \\ &= \sigma_B^2. \end{aligned}$$

From this and equation (5) we can deduce the following unbiased estimate of $V_{\bar{D}}$:

$$S_{\bar{D}}^2 = \frac{1}{2n}(S^2+2C)$$

Finally, the "unbiased paired t-statistic" for testing the null hypothesis $H_0 = \{\mu_1 = \mu_2\}$ is

$$T_{UP} = \sqrt{2n}\,\frac{\bar{D}}{\sqrt{(S^2+2C)}}$$

which is approximately distributed as a Student distribution with $2n$ - 2 df under H_0.

Unpaired test procedure

Let $\bar{X}_j$ (respectively $\bar{Y}_j$) be the mean of the 2 results obtained with the same biological sample (in 2 different arrays and with the 2 dyes) for condition A (respectively condition B). From model (1) one obtains

$$\bar{X}_j = \mu_A + (\delta_1 + \delta_2)/2 + B_j + M_{i(j)} + M_{i'(j)})/2 + (T_{i(j)} + T_{i'(j)})/2$$
$$\bar{Y}_j = \mu_B + (\delta_1 + \delta_2)/2 + B'_j + M_{i(j)} + M_{i'(j)})/2 + (T'_{i(j)} + T'_{i'(j)})/2$$

where j is the biological sample index (recall that sample j is different for the two conditions), $i(j)$ and $i'(j)$ are the arrays on which sample j has been hybridized. $\bar{X}_j$ and $\bar{Y}'_j$ may be correlated as a result of a possible common array effect. $\bar{X}_j$ and $\bar{X}'_j$ are uncorrelated because the two different biological samples of the same condition cannot be present together on the same array. From result (5) we have:

$$V(\bar{X}-\bar{Y}) = V_{\bar{D}} = \frac{1}{2n}(4\sigma_B^2 + 2\sigma_T^2).$$

The usual unpaired estimate of $V(\bar{X}-\bar{Y})$ is equal to $(S_X^2 + S_Y^2)/n$, where $S_X^2 = \frac{1}{n-1}\sum_j(\bar{X}_j - \bar{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_j(\bar{Y}_j - \bar{Y})^2$, whose common mean (under the homoscedastic model (1)) is equal to

$$\sigma_B^2 + \frac{1}{2}(\sigma_T^2 + \sigma_M^2).$$

Therefore

$$E\left[(S_X^2 + S_Y^2)\right] = 2\sigma_B^2 + \sigma_T^2 + \sigma_M^2.$$

This method overestimates the true variance

$$V_{\bar{D}} = \frac{1}{2n}\left[4\sigma_B^2 + 2\sigma_T^2\right].$$

The overestimation is more dramatic as σ_M^2 increases. This estimate may be corrected: σ_M^2 may be estimated using the empirical covariance between $\bar{X}_j$ and $\bar{Y}_{j'}$. Let

$$C_{XY} = \frac{1}{n-2}\left(\sum_{j=1}^{n}(\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y}) + \sum_{j=1}^{n}(\bar{X}_j - \bar{X})(\bar{Y}_{j-1} - \bar{Y})\right)$$

with the convention that $\bar{Y}_0 = \bar{Y}_n$. The mean of the first sum is

$$\begin{aligned}\frac{1}{n-2}\sum_{j=1}^{n}E\left[(\bar{X}_j - \bar{X})(\bar{Y}'_j - \bar{Y})\right] &= \frac{1}{n-2}\sum_{j=1}^{n}E\left[((M_{i(j)} + M_{i'(j)})/2 - \bar{M})((M_{i(j)} + M_{i'(j)})/2 - \bar{M})\right]\\ &= \frac{1}{n-2}\sum_{j=1}^{n}\frac{\sigma_M^2}{4} - \frac{2\sigma_M^2}{4n}\\ &= \frac{\sigma_M^2}{4}\end{aligned}$$

It is easy to see that the second sum in C_{XY} has the same mean. Therefore an unbiased estimate of $V_{\bar{D}}$ is $\frac{1}{n}(S_X^2 + S_Y^2 - 2C_{XY})$, and the approximate unpaired t-statistic is

$$T_{UU} = \sqrt{n}\,\frac{\bar{X}-\bar{Y}}{\sqrt{S_X^2+S_Y^2-2C_{XY}}}.$$

Authors' contributions

TMH and JJD conceived the method and prepared the manuscript. TMH and JA implemented part of the software and performed the statistical analysis. NM made the Embriogenomics experiment under the direction of OS. All authors contributed to the discussion and approved the final manuscript.

Acknowledgements

The authors thank Douglas L. Crawford who provided the Teleofish dataset.

References

1. Yang Y, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3(8):**579-88.
2. Churchill G: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32:**490-495.
3. Yang Y, Dudoit S, Luu P, Speed T: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nuclear Acids Res* 2002, **30(4):**e15.
4. Kerr M, Afshari C, Bennett L, Bushel P, Martinez J, Walker N, Churchill G: **Statistical Analysis of a gene expression microarray experiment with replication.** *Statistica Sinica* 2002, **12:**203-217.
5. Dobbin K, Shih J, Simon R: **Statistical design of reverse dye microarrays.** *Bioinformatics* 2003, **19(7):**803-810.
6. Martin-Magniette M, Aubert J, Cabannes E, Daudin J: **Evaluation of the gene-specific dye bias in cDNA microarray experiments.** *Bioinformatics* 2005, **21(9):**1995-2000.

7. Dobbin K, Kawasaki E, Petersen D, Simon R: **Characterizing dye bias in microarray experiments.** *Bioinformatics* 2005, **21(10):**2430-2437.
8. Oleksiak M, Churchill G, Crawford D: **Variation in gene expression within and among natural populations.** *Nat Genet* 2002, **32:**261-266.
9. Whitehead A, Crawford D: **Neutral and adaptive variation in gene expression.** *Proc Natl Acad Sci USA* 2006, **103(14):**5425-5430.
10. Kerr M, Martin M, Churchill G: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7:**819-837.
11. Wit E, McClure J: *Statistics for Microarrays: Design, Analysis and Inference* Chichester: Wiley; 2004.
12. Landgrebe J, Bretz F, Brunner E: **Efficient two-sample designs for microarray experiments with biological replications.** *Silico Biology* 2004, **4:**.
13. Landgrebe J, Bretz F, Brunner E: **Efficient design and analysis of two colour factorial microarray experiments.** *Comput Stat & Data Anal* 2006, **50(2):**499-517.
14. Mansouri N, Sandra O, Aubert J, Everts R, Galio L, Heyman Y, Audouart C, Degrelle S, Hue I, Yang X, Lewin H, JP R: **Identification of differentially regulated genes in the endometrium of cyclic and pregnant cows using a high-throughput transcriptome analysis.** *American Journal of Reproductive Immunology* 2007, **58(5):**204-220.
15. Wit E, Nobile A, Khanin R: **Near-optimal designs for dual channel microarray studies.** *J R Stat Soc C* 2005, **54(5):**817-830.
16. Baldi P, Long A: **A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes.** *Bioinformatics* 2001, **17:**509-519.
17. Delmar P, Robin S, Daudin J: **VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data.** *Bioinformatics* 2005, **21(4):**502-508.
18. Smyth G: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3:**Article 3.
19. Tusher V, Tibshirani R, Chu C: **Significance analysis of microarrays applied to transcriptional response to ionizing radiations.** *PNAS* 2001, **98:**5116-5121.

Annexe H
Article publié présenté dans la section 6.1

Classification of Human Chromosome 21 Gene-Expression Variations in Down Syndrome: Impact on Disease Phenotypes

E. Aït Yahya-Graison, J. Aubert, L. Dauphinot, I. Rivals, M. Prieur, G. Golfier, J. Rossier, L. Personnaz, N. Créau, H. Bléhaut, S. Robin, J. M. Delabar, and M.-C. Potier

Down syndrome caused by chromosome 21 trisomy is the most common genetic cause of mental retardation in humans. Disruption of the phenotype is thought to be the result of gene-dosage imbalance. Variations in chromosome 21 gene expression in Down syndrome were analyzed in lymphoblastoid cells derived from patients and control individuals. Of the 359 genes and predictions displayed on a specifically designed high-content chromosome 21 microarray, one-third were expressed in lymphoblastoid cells. We performed a mixed-model analysis of variance to find genes that are differentially expressed in Down syndrome independent of sex and interindividual variations. In addition, we identified genes with variations between Down syndrome and control samples that were significantly different from the gene-dosage effect (1.5). Microarray data were validated by quantitative polymerase chain reaction. We found that 29% of the expressed chromosome 21 transcripts are overexpressed in Down syndrome and correspond to either genes or open reading frames. Among these, 22% are increased proportional to the gene-dosage effect, and 7% are amplified. The other 71% of expressed sequences are either compensated (56%, with a large proportion of predicted genes and antisense transcripts) or highly variable among individuals (15%). Thus, most of the chromosome 21 transcripts are compensated for the gene-dosage effect. Overexpressed genes are likely to be involved in the Down syndrome phenotype, in contrast to the compensated genes. Highly variable genes could account for phenotypic variations observed in patients. Finally, we show that alternative transcripts belonging to the same gene are similarly regulated in Down syndrome but sense and antisense transcripts are not.

Down syndrome (DS [MIM #190685]) results from the triplication of chromosome 21 and is the most common genetic cause of mental retardation in humans, occurring in ~1 in 800 newborns. The phenotype of DS is characterized by >80 clinical features, including cognitive impairments, muscle hypotonia, short stature, facial dysmorphisms, congenital heart disease, and several other anomalies.[1] These clinical features can vary considerably in number and in severity,[2] and certain abnormalities, such as acute megakaryoblastic leukemia and Hirschsprung disease, occur at higher frequencies in patients with DS than in the general population.

Trisomy 21 has been known to be the cause of DS since 1959, when Lejeune and colleagues demonstrated the presence in three copies of chromosome 21 in persons with DS.[3] The phenotype of DS is thus thought to be the result of gene-dosage imbalance. However, the molecular mechanisms by which such dosage imbalance causes abnormalities remain poorly understood. Two different hypotheses have been proposed to explain the phenotype of DS: "developmental instability" (loss of chromosomal balance) and "gene-dosage effect." According to the developmental instability hypothesis, the presence of a supernumerary chromosome globally disturbs the correct balance of gene expression in DS cells during development.[4,5] However, this hypothesis is weakened by the fact that other autosomal trisomy syndromes do not lead to the same clinical pattern.[6] Moreover, correlations between genotype and phenotype in patients with partial trisomies indicate that a restricted region in 21q22.2 is associated with the main features of DS, including hypotonia, short stature, facial dysmorphies, and mental retardation.[7–9] This DS chromosomal region (DCR) supports the alternative gene dosage–effect hypothesis, which postulates that the restricted number of genes from chromosome 21 that are overexpressed in patients with segmental trisomies contributes to the phenotypic abnormalities.

To determine which hypothesis applies to the etiology of DS, several gene-expression studies of human DS cells or tissues have been conducted.[10–17] Most of these studies have shown a global up-regulation of the three-copy genes mapping to the trisomic chromosome, but the limited number of studied DS cases restricted the statistical analysis and did not allow the identification of precise gene deregulation. Moreover, these studies were performed using a small number of three-copy genes. Several other ex-

From the Neurobiologie et Diversité Cellulaire, Unité Mixte de Recherche 7637 du Centre National de la Recherche Scientifique et de l'Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (E.A.Y.-G.; L.D.; G.G.; J.R.; M.-C.P.), Université Paris Diderot-Paris (E.A.Y.-G.; N.C.; J.M.D.), Unité Mixte de Recherche de l'Ecole Nationale du Génie Rural, des Eaux et des Forêts, de l'Institut National Agronomique Paris-Grignon et de l'Institut National de la Recherche Agronomique (J.A.; S.R.), Equipe de Statistique Appliquée (I.R.; L.P.), Service de Cytogénétique, Hôpital Necker Enfants Malades (M.P.), and Institut Jérôme Lejeune (H.B.), Paris

Received March 13, 2007; accepted for publication May 21, 2007; electronically published July 19, 2007.

Address for correspondence and reprints: Dr. M.-C. Potier, Neurobiologie et Diversité Cellulaire, CNRS UMR 7637, ESPCI, 10 rue Vauquelin, 75005 Paris, France. E-mail: marie-claude.potier@espci.fr

Am. J. Hum. Genet. 2007;81:475–491. 0002-9297/2007/8103-0006$15.00
DOI: 10.1086/520000

Table 1. Experimental Design

	Men with DS							Women with DS		
Controls	1	2	3	4	5	6	7	8	9	10
Men:										
11	−1	1		−1		1			−1	1
12				1	−1			1		−1
13		−1	1			−1		1	1	−1
14					1			−1	−1	1
Women:										
15		1						−1		
16					−1				1	
17					1		−1	1		−1
18			−1			1		−1	1	
19	1								−1	
20		−1							1	
21						−1	1		−1	1

NOTE.—Microarray experiments were performed using LCLs from individuals with DS and control individuals in accordance with a mixed model (see the "Material and Methods" section). Each "1" indicates one experiment. "+1" means that DS and control samples were labeled with Cy5 and Cy3, respectively. "−1" means that DS and control samples were labeled with Cy3 and Cy5, respectively.

periments have been done on animal models of DS with a greater number of chromosome 21 gene orthologs by use of microarray and quantitative PCR experiments.[18–21] In these studies, the three-copy genes were overexpressed, with a mean ratio of 1.5, which is proportional to the gene-dosage imbalance. However, some of these triplicated genes appeared to escape the "1.5-fold rule." Yet, these animal models are not trisomic for all chromosome 21 orthologs. Thus, a comprehensive classification of all human genes on chromosome 21, according to their level of expression in DS, does not yet exist.

The goal of the present study was to fill this knowledge gap and to find the genes that are likely to be involved in DS phenotypes through their transcriptional dysregulation.[22] For this purpose, we designed an oligonucleotide microarray containing all chromosome 21 genes, ORFs, antisense transcripts, and predicted genes listed in the most common databases (NCBI Gene Database, Eleanor Roosevelt Institute, and Max Planck Institute), except for the 53 genes of the keratin-associated protein cluster. Gene expression was measured on lymphoblastoid cell lines (LCLs) from 10 patients with DS and 11 control individuals. LCLs are easy to obtain and are widely used to study genotype-phenotype correlations.[23] To our knowledge, this is the most comprehensive study so far that has been done using triplicated genes in DS human cells. In addition, we analyzed data with a mixed-model analysis of variance, to find genes that are differentially expressed in DS independent of sex and interindividual variations. Our data show a global gene dosage–dependent expression of chromosome 21 genes in LCLs, with no effect of sex. In addition, by use of our data-analysis protocol, chromosome 21 genes can now be classified into four classes: class I genes are overexpressed with a mean ratio very close to 1.5, proportional to the gene-dosage effect of trisomy 21; class II genes are overexpressed with ratios significantly >1.5, reflecting an amplification mechanism; class III genes have ratios significantly <1.5, corresponding to compensated genes; and class IV genes have expression levels that are highly variable between individuals. This classification should have an impact on the search for genes that are involved in the DS phenotype.

Material and Methods

Cell Lines and Culture Conditions

LCLs were derived from the B lymphocytes of 10 patients with DS collected from the cytogenetic service of the hospital Necker Enfants Malades and the Institut Jérôme Lejeune. Parents of patients from the Institut Jérôme Lejeune gave their informed consent, and the French biomedical ethics committee gave its approval for this study (Comité de Protection des Personnes dans la Recherche Biomédicale number 03025). Written informed consent was obtained from the participants or from their families by the cytogenetic service of Hôpital Necker Enfants Malades. Cell lines from 11 control individuals were also obtained with their written informed consent, for comparison of chromosome 21 gene-expression profiles. Culture media consisted of Opti-MEM with GlutaMax (Invitrogen) supplemented with 5% fetal bovine serum from a unique batch and 1% penicillin and streptomycin mix (10,000 U/ml). Cell lines were grown at 37°C in humidified incubators, in an atmosphere of 5% CO_2. Each culture was grown to at least 60×10^6 cells. All cell lines were karyotyped, to confirm their trisomic or euploid status and also to verify that immortalization by the Epstein-Barr virus (EBV) did not produce any visible chromosomal rearrangement other than trisomy 21. Cells

Table 2. List of Oligonucleotide Primers Used in the QPCR Experiments

Primer	Sequence (5′→3′)
CHAF1B_UP	CCATCATATGGGATGTCAGCAA
CHAF1B_LOW	CTTCATGCTGTCGTCGTGAAAC
CSTB_UP	GCCACCGCCGAGACCCAGCA
CSTB_LOW	TGGCTTTGTTGGTCTGGTAG
DSCR1_UP	GCACAAGGACATTTGGGACT
DSCR1_LOW	TTGCTGCTGTTTTCACAACC
DYRK1A_UP	ATCCGACGCACCAGCATC
DYRK1A_LOW	AATTGTAGACCCTTGGCCTGGT
GART_UP	CTGGGATTGTTGGGAACCTGAG
GART_LOW	ACCAAAGCAGGGAAGTCTGCAC
H2BFS_UP	CAGAAGAAGGACGGCAGGAA
H2BFS_LOW	GAAGCCTCACCTGCGATGCG
MX1_UP	GCCAGTATGAGGAGAAGGTGCG
MX1_LOW	GTTTCAGCACCAGCGGGCATCT
SNF1LK_UP	GCCGCTTCCGCATCCCCTTCTT
SNF1LK_LOW	CTCATCGTAGTCGCCCAGGTTG
SOD1_long_UP	TCGCCCAATAAACATTCCCTTG
SOD1_long_LOW	AAGTCTGGCAAAATACAGGTCATTG
STCH_UP	GGACGTGGCCTTTCTGATAA
STCH_LOW	CTTGACGGATCCGAGGAATA
TMEM1_UP	CGTGCAGGAACTGAAGCTCTTA
TMEM1_LOW	TCTGAGCTGTGTTGGCTGTTTC
L13852_UP	CTCCAATCTCAGCCGTCAGT
L13852_LOW	AGCCACACCATCCACACGGG
AB000468_UP	CAAGAAAGCGTCGTGGTGGA
AB000468_LOW	ATCGTCACTGCTCACCACAC

Table 3. HSA21 Oligoarray Content

Putative Expressed Sequences	HSA21 Content[a]	HSA21 Oligoarray Content	
		No. of Sequences[b]	HSA21 Coverage[c] (%)
Genes	182	145	82.42
ORFs	93	58	62.37
Predictions	Not represented	118[d]	NA
Antisense transcripts	Only 1 represented	18	NA

[a] NCBI Gene Database build 36.2 was used to estimate HSA21 gene content. Only current sequences were considered, with the exception of pseudogenes and hypothetical proteins.

[b] The number of HSA21 sequences represented by at least one probe on the HSA21 oligoarray.

[c] The percentage of HSA21 sequences currently annotated in NCBI Gene Database that are represented on the microarray. NA = not available.

[d] Of the 118, 20 are represented with their reverse sequence.

were harvested by centrifugation, were washed in 5 ml PBS, followed by another centrifugation, and were stored at −80°C.

Human Chromosome 21 (HSA21) Oligoarray

A dedicated oligonucleotide microarray—named "HSA21 oligoarray," containing 664 50-mer amino-modified oligonucleotides representing 145 genes, 58 ORFs, 118 predictions (20 of them represented in both orientations), and 18 antisense transcripts assigned to chromosome 21—was used in the present study. Predictions represented on the array included cDNAs and exons from the *CBR-ERG* region on 21q deduced from cDNA isolation and exon-trapping experiments[8] and gene or exon predictions produced from *in silico* analysis of the complete sequence of human chromosome 21.[24] Nonredundant transcript sequences and antisense transcripts were also included in this oligoarray.[25–27] Thirty-nine genes assigned to chromosomes other than chromosome 21, represented by 58 oligonucleotides—showing a wide range of expression levels according to UniGene and no variations between DS and control samples as demonstrated by the first version of the HSA21 oligoarray (data not shown)—were added for data normalization. All probes present on the array were designed using the SOL software (G.G., S. Lemoine, A. Bendjoudi, J.R., S. Lecrom, and M.-C.P., unpublished data). Sequences were then synthesized by EuroGentec and were spotted onto CodeLink activated glass slides (Amersham Biosciences) by use of a MicroGrid II spotter (Biorobotics). Each array contained two matrices with eight blocks each, in which probes were present in duplicates so that each oligonucleotide was present in four replicates on each slide.

Table 4. Classification of HSA21 Genes by Statistical Analysis

DS/Control Ratio	Class by DS/Control Ratio	
	Not Significantly Different from 1	Significantly Different from 1
Significantly >1.5	…	II
Not significantly different from 1.5	IV	I
Significantly <1.5	III	…

Experimental Design

The experiment comprised 10 patients with DS (7 men and 3 women) and 11 controls (4 men and 7 women). Samples from the same individual were used in different hybridizations.

For each gene, we used the linear model

$$y_{ijklm} = \mu + D_i + S_j + A_l + F_m + DS_{ij} + DF_{im} + SF_{jm} + I(DS)_{ijk} + \varepsilon_{ijklm}\ , \qquad (1)$$

where y_{ijklm} is the normalized expression of the gene in $\log_2$ for factor i (DS or control sample), sex is j, patient number is k ($k = 1, \ldots, 21$), dye label is m (m is red, for Cy5, or green, for Cy3) on the HSA21 oligoarray l. The symbols *D, S, A,* and *F* represent the fixed effects due to the disease, sex, array, and fluorochrome, respectively. For example, *D* represents the modifications of the gene expression level due to the disease. *A* and *F* are both nuisance parameters that account for potential technological biases. *DS, DF,* and *SF* correspond to interacting effects: disease and sex, disease and fluorochrome, and sex and fluorochrome, respectively. The symbol *I*(*DS*) refers to the patient (nested within disease and sex) random effect. This last effect accounted for the correlation between samples used in different hybridizations but collected from the same patient.

We assumed the independence between all $I(DS)_{ijk}$ and E_{ijklm}. We also assumed that $I(DS)_{ijk}$ was independent with a distribution $N(0,s^2)$ and that E_{ijklm} was independent with distribution $N(0,\sigma^2)$.

Model (1) can be rewritten under the matrix form

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{ZU} + \mathbf{E}\ , \qquad (2)$$

where θ is the vector of fixed effects (*D, S, A, F, DS, DF,* and *SF*), $\mathbf{U}$ is the vector of $I(DS)_{ijklm}$, and $\mathbf{E}$ is the vector of E_{ijklm}. $\mathbf{Y} \sim N(X\theta,\Sigma)$, where $\Sigma = 2\sigma_g^2 Id + s_g^2 ZZ^T$. $\mathbf{Y}$ has n rows (n is the total number of samples) and one column, $\mathbf{X}$ is the matrix describing the status of the patient (disease and sex) from which the sample was collected. $\mathbf{Z}$ has n rows and I columns (I is the total number of patients) and describes the correspondence between samples and patients.

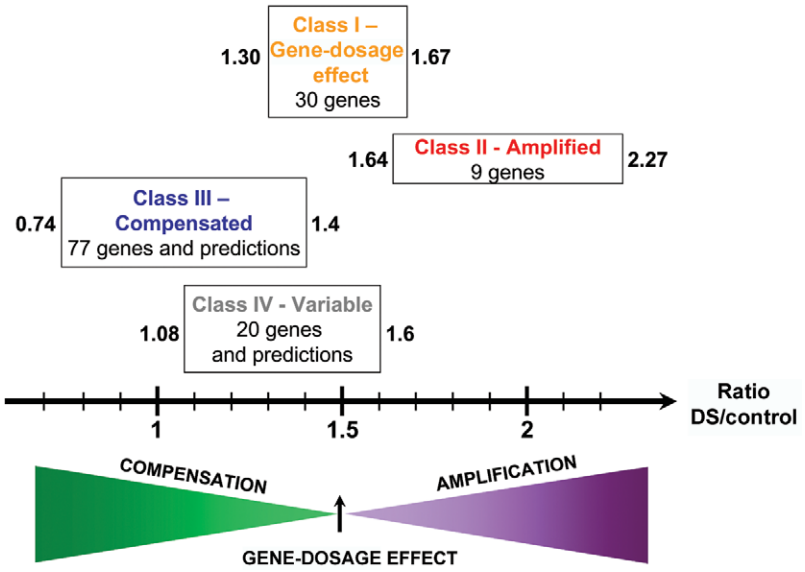


Figure 1. Classification of HSA21 genes according to the expression ratio between DS and control LCLs. The sum of classified genes is 136 genes minus 2 (*C21ORF108* and *PRMT2*) that appear twice, depending on the oligonucleotide probe considered (see the "Results" section for details).

On each array *l*, we actually observed the differential expression (red signal minus green signal)

$$\begin{aligned} y_{ii'jj'kk'lmm'} = y_{ijklm} - y_{i'j'k'lm'} &= (D_i - D_{i'}) + (S_j - S_{j'}) \\ &+ (DS_{ij} - DS_{i'j'}) + (F_m - F_{m'}) \\ &+ (DF_{im} - DF_{i'm'}) + (SF_{im} - SF_{i'm'}) \\ &+ [I(DS)_{ijk} - I(DS)_{i'j'k'}] + (\varepsilon_{ijklm} - \varepsilon_{i'j'k'l'm'}) \ . \end{aligned} \quad (3)$$

This model can be rewritten under another matrix, Δ, describing the comparisons performed on each array. This matrix had L rows (L is the total number of arrays) and n columns. The lth row of Δ was zero except for the value 1 in the column corresponding to the sample labeled in red and −1 in the column corresponding to the sample labeled in green. The model for the differential expression was obtained by multiplying all terms of equation (2) by Δ:

$$\Delta\mathbf{Y} = \Delta\mathbf{X}\theta + \Delta\mathbf{Z}\mathbf{U} + \Delta\mathbf{E} \ .$$

The vector of differential expression $\Delta\mathbf{Y}$ has distribution $N(\Delta\mathbf{X}\theta, \Delta\Sigma\Delta^{\mathrm{T}})$.

The experimental design was defined by the three matrices $\mathbf{X}$, $\mathbf{Z}$, and Δ. $\mathbf{X}$ and $\mathbf{Z}$ basically depend on the number of samples for each patient. Because the microarrays used were two-color assays, the total number of samples was twice the number of slides.

The important remaining choice was the comparison to be made—that is, the choice of Δ. Consideration of the differential expression between two patients analyzed on the same array eliminated the array effect and the corresponding constants. We were not interested in other technical effects, such as fluorochrome or interactions of fluorochrome with other effects. To eliminate DF_{im} and SF_{jm} effects, we proposed a balanced design for the fluorochrome effect. Finally, data were normalized across genes, to remove the dye effect F_m.

The last criterion was Δ—the precision of the estimated effects (gathered into the vector $\hat{\theta}$). According to the mixed linear model theory, this precision is given by its variance matrix,

$$V(\hat{\theta}_A) = [\Delta^T\mathbf{X}^T(\Delta\Sigma\Delta^T)^{-1}\Delta\mathbf{X}]^{-1} \ ,$$

which depends on Δ and the ratio σ^2/s^2. The diagonal contained all the information about the quality of the estimates. It gave the variance of the estimates of all effects of interest to D_i, S_j, and DS_{ij}. This matrix was the ultimate tool for comparing the designs.

We calculated the variance of the disease effect (which is of primary interest) and the determinant of the variance covariance matrix (which gave a global measure of the precision of the estimates) for a certain number of designs Δ. To do that, we used a value given a priori for the ratio σ^2/s^2 equal to 2.

We finally chose a design involving 40 arrays. Each patient appeared in two to eight different experiments, and samples from the same patient were marked the same number of times with each fluorescent dye (Cy3 or Cy5). On each array, a DS LCL and a control LCL were compared, to increase the precision of the disease effect. Ten arrays compared (i) a man with DS and an unaffected man, (ii) a female with DS and an unaffected female, (iii) a man with DS and an unaffected female, or (iv) a female with DS and an unaffected male. The design is described in table 1.

mRNA Extraction, HSA21 Oligoarray Hybridization, Data Filtering, and Normalization

mRNA was extracted from frozen individual cell samples by use of Fast Track 2.0 mRNA Isolation kit (Invitrogen) in accordance with the manufacturer's instructions. To eliminate DNA contamination, the appended DNase protocol of RNeasy mini kit (Qia-

gen) was used in accordance with the manufacturer's protocol. Samples were further tested for purity and quantity with RNA 6000 NanoChips by use of the Agilent 2100 Bioanalyzer (Agilent Technologies). By use of the Reverse-iT RTase Blend kit (ABGene), 2 μg of mRNA was converted into Cy3- or Cy5-labeled cDNA by incorporation of fluorescent dUTP (Amersham). Labeled targets were then purified on Qiaquick columns in accordance with the manufacturer's protocol (Qiagen). Hybridization of sample pairs on HSA21 oligoarrays (one DS sample and one control sample), according to the experimental design, was performed using hybridization buffer (50% formamide, 4× saline sodium citrate [SSC], 0.1% SDS, and 5× Denhart) at 42°C overnight. Slides were washed in 2× SSC and 0.1% SDS three times for 5 min, in 0.2× SSC for 1 min, and in 0.1× SSC for 2 min. Data were acquired with GenePix 4000B scanner and by use of the GenePix Pro 6.0 software (Axon). For each array, the raw data comprised the median feature pixel intensity at wavelengths 635 nm and 532 nm for Cy5 and Cy3 labeling, respectively. After subtraction of the background signal, LOWESS normalization[28] of the *M* values corresponding to Cy5/Cy3 signal ratios in $\log_2$ was applied to all oligonucleotides representing non-HSA21 genes and was used to calculate a correction factor applied to *M* values for HSA21 probes under The R Project for Statistical Computing. Normalized data from each slide was then filtered using two criteria: (i) for each oligonucleotide, at least two values among the four replicates had to be available, and (ii) SD of values corresponding to the geometric mean in $\log_2$ of Cy3 and Cy5 signal intensities (the *A* value) had to be <1. Arithmetic means of normalized and filtered *M* and *A* values were calculated for each oligonucleotide and were submitted to the statistical analysis. All microarray data used in this study were deposited in the Gene Expression Omnibus (GEO) database (accession number GSE6408).

Expression Data Analysis: Statistical Testing

Mixed model.—To determine differentially expressed genes for DS, sex, and DS × sex effects, we performed a mixed-model analysis of variance according to the experimental design. This method was chosen to distinguish between interindividual variability and experimental variability.[29] We used the mixed procedure of the SAS software with the restricted maximum likelihood (REML) method of estimation.[30] After the filtering and normalization steps, the number of observations per spot varied between 8 and 40, which was enough to calculate the variance for each gene.

We first tested the effects of the complete model (3). Since the sex and DS × sex effects were not significant for any gene, these two effects were dropped from the model. We finally analyzed the simplified model

$$y_{ii'jj'kk'l} = y_{ijkl} - y_{i'j'k'l} = (D_i - D_{i'}) + [I(DS)_{ijk} - I(DS)_{i'j'k'}] + (\varepsilon_{ijklm} - \varepsilon_{i'j'k'l'm'}) \ .$$

We deduced raw *P* values from comparison with 1 under the null hypothesis and adjusted them by the Benjamini-Hochberg procedure, which controls the false-discovery rate (FDR).[31] We then analyzed the simplified model by comparison with 1.5 under the null hypothesis and adjusted the raw *P* values by use of the method described by Storey et al.[32]

Principal-components analysis (PCA).—Results from microarray experiments were obtained as the differential expression between DS and control samples. *M* values corresponded to $\log_2(\text{DS}) - \log_2(\text{control})$, and *A* values to $[\log_2(\text{DS}) + \log_2(\text{control})]/2$. For each probe *p*, the mean value of its expression (in $\log_2$) in DS cell lines and in controls could thus be expressed as

$$\left\{ \begin{array}{ll} E_p^i = \frac{1}{N_i}\sum_{k\in S_i}\left(A_p^k + \frac{M_p^k}{2}\right) & \text{for } i = 1\text{–}10 \text{ (DS)} \\ E_p^i = \frac{1}{N_i}\sum_{k\in S_i}\left(A_p^k - \frac{M_p^k}{2}\right) & \text{for } i = 11\text{–}21 \text{ (controls)} \end{array} \right\} \ ,$$

where *i* denoted the index of the individual, S_i the set of slides on which all samples from the individual *i* have been hybridized, and N_i the size of this set. Since the overall expression level of the genes of one individual varied from one slide to another and to avoid normalization across slides, we made the simplification $A_p^k = 0$ and reconstructed relative mean values of expression *E* for each individual as

$$\left\{ \begin{array}{ll} E_p^i = \frac{1}{N_i}\sum_{k\in S_i}\left(\frac{M_p^k}{2}\right) & \text{for } i = 1\text{–}10 \text{ (DS)} \\ E_p^i = \frac{1}{N_i}\sum_{k\in S_i}\left(-\frac{M_p^k}{2}\right) & \text{for } i = 11\text{–}21 \text{ (controls)} \end{array} \right\} \ .$$

PCA of chromosome 21 genes and genes mapping to other chromosomes was performed separately using *E* values deduced from all expressed chromosome 21 probes and all expressed non–chromosome 21 probes, respectively.

PCR Experiments

To validate expression ratios between DS and control samples obtained from HSA21 oligoarray data, 100 ng of mRNA was reverse transcribed into cDNA by use of Reverse-iT RTase Blend kit (ABgene). Quantitative real-time PCR (QPCR) on diluted cDNA was conducted in the presence of 0.6 mM of each specific primer (designed by Primer3 software) and 1× Quantitect SYBR Green PCR master mix (Qiagen) containing 2.5 mM $MgCl_2$, Hotstart *Taq* polymerase, dNTP mix, and the fluorescent dye SYBR Green I. QPCR experiments were performed in a Lightcycler system (Roche Molecular Biochemicals) on 11 HSA21 genes: *CHAF1B, CSTB, DSCR1, DYRK1A, GART, H2BFS, MX1, SNF1LK, SOD1, STCH,* and *TMEM1*. The ubiquitin-activating enzyme E1 (NCBI Entrez accession number L13852) mRNA mapping to HSA3 and the zinc-finger protein (NCBI Entrez accession number AB000468) mRNA mapping to HSA4 were used as endogenous control genes, as described by Janel et al.[33] For each sample, the mean cycle threshold value, C_t, was corrected by subtracting the mean of the C_t obtained with the two reference genes. PCR primers are listed in table 2.

Results

Design of a Comprehensive HSA21 Oligoarray

The HSA21 oligoarray was designed for the exhaustive study of human chromosome 21 gene expression in DS. This microarray contained 664 sequences representing 145 genes, 58 ORFs, 118 predictions (plus the reverse sequences for 20 of them), and 18 antisense transcripts, allowing expression analysis of all putative genes mapping to chromosome 21 and related to DS. To increase the

Table 5. List of Genes Classified According to Their Expression in DS LCLs

Gene Symbol	GenBank Accession Number	Ratio	*A*	*M*	Var(*M*)	Class[a]
as-TTC3	BF979681.2	1.56	9.07	.64	.64	I
C21orf108 (exon 39)	AF231919.1	1.30	7.84	.38	.18	I
C21orf33	BI824121.1	1.52	10.54	.60	.20	I
C21orf59	AF282851.1	1.35	9.61	.44	.21	I
C21orf66	AY033903.1	1.51	9.50	.59	.12	I
C21orf7	AY171599.2	1.43	8.00	.52	.49	I
C21orf91	BC015468.2	1.59	9.50	.67	.26	I
CBS	AF042836.1	1.61	8.13	.69	.48	I
CCT8	BC095470.1	1.65	12.64	.72	.32	I
CRYZL1	BC033023.2	1.52	9.45	.60	.12	I
DONSON	AF232673.1	1.42	9.73	.50	.11	I
DYRK1A	D86550.1	1.41	10.20	.49	.17	I
HMGN1	M21339.1	1.38	11.82	.46	.10	I
IFNAR1	AY654286.1	1.47	8.10	.56	.18	I
IFNAR2	BC013156.1	1.67	8.38	.74	.14	I
IFNGR2	AY644470.1	1.45	10.35	.54	.24	I
IL10RB	BT009777.1	1.66	10.18	.73	.21	I
ITGB2	BC005861.2	1.61	11.34	.68	.40	I
MCM3AP	AY590469.1	1.45	11.69	.54	.19	I
MRPL39	AF109357.1	1.47	9.82	.55	.15	I
PFKL	X15573.1	1.50	12.44	.58	.23	I
PIGP	AF216305.1	1.60	8.14	.68	.25	I
PTTG1IP	NM_004339.2	1.52	9.89	.60	.20	I
SFRS15	AF023142.1	1.39	10.42	.47	.14	I
SLC5A3	L38500.2	1.57	8.92	.66	.18	I
SON	AY026895.1	1.51	12.64	.60	.14	I
SUMO3	BC008420.1	1.41	10.82	.50	.10	I
USP16	AY333928.1	1.46	9.94	.54	.10	I
USP25	AF170562.1	1.58	9.93	.66	.10	I
ZNF294	NM_015565.1	1.51	8.77	.60	.11	I
BTG3	D64110.1	1.82	10.63	.86	.21	II
C21orf57	AY040875.1	1.74	9.49	.80	.28	II
MRPS6	AB049942.1	1.64	10.26	.72	.13	II
PDXK	AY303972.1	1.71	10.06	.77	.23	II
SAMSN1	AF222927.1	2.27	10.47	1.18	.72	II
SLC37A1	AF311320.1	1.72	9.33	.78	.32	II
SNF1LK	AB047786.1	2.14	9.47	1.10	1.15	II
STCH	U04735.1	1.97	9.86	.98	.42	II
TTC3	D84296.1	1.79	10.59	.84	.20	II
aa071193	AA071193.1	.97	7.34	−.05	.48	III
AIRE	AB006682.1	.82	7.28	−.28	.29	III
AL041783	AL041783.1	1.06	7.10	.09	.36	III
as-C21orf56	BC084577.1	1.07	11.75	.10	.08	III
as-KIAA0179	AA425659.1	1.16	8.08	.22	.67	III
ATP5J	BC001178.1	1.35	7.99	.44	.11	III
B184	AL109967.2	1.06	9.10	.08	.62	III
B27 inverse	AP000034.1	.83	8.40	−.27	.31	III
C21orf108 (exon 26)	AF231919.1	1.09	8.86	.13	.09	III
C21orf12	AP001705.1	.97	7.49	−.05	.72	III
C21orf2	NM_004928.1	1.30	8.80	.38	.14	III
C21orf21	AA969880	1.16	7.33	.22	.28	III
C21orf25	AB047784.1	1.19	8.61	.25	.31	III
C21orf29	AJ487962.1	.98	7.27	−.03	.22	III
C21orf34	AF486622.1	.93	8.17	−.11	.28	III
C21orf42	AY035383.1	1.14	10.05	.19	.22	III
C21orf45	*AF387845.1*	1.23	9.39	.30	.12	III
C21orf49	BC117399.1	1.14	7.80	.19	.32	III
C21orf51	AY081144.1	1.26	9.47	.33	.09	III
C21orf54	AA934973.1	1.01	7.26	.01	.38	III
C21orf58	BC028934.1	1.11	7.73	.15	.21	III
C21orf6	BC017912.1	1.20	9.63	.26	.22	III
CHAF1B	U20980.1	1.29	8.35	.37	.13	III
CLIC6	AF448438.1	.74	9.26	−.43	.98	III

(*continued*)

Table 5. (continued)

Gene Symbol	GenBank Accession Number	Ratio	*A*	*M*	Var(*M*)	Class[a]
CXADR	AF200465.1	.90	8.16	−.15	.32	III
D21S2056E	U79775.1	1.33	10.55	.41	.18	III
DCR1-17.0	AJ001875.1	1.13	7.16	.17	.71	III
DCR1-19.0	AJ001906.1	.94	7.13	−.09	.19	III
DCR1-20.0-reverse	AJ001893.1	.95	7.14	−.07	.36	III
DCR1-25.0-reverse	AJ001905.1	.96	7.34	−.06	.31	III
DCR1-7.0	AJ001861.1	1.09	7.48	.12	.27	III
DCR1-7.0-reverse	AJ001861.1	1.21	7.13	.28	.28	III
DCR1-8.0	AJ001862.1	.98	7.58	−.03	.42	III
DCR1-8.0-reverse	AJ001862.1	1.00	8.61	.00	.20	III
DSCAM_Intronic_Model	BG221591.1	1.21	8.39	.28	1.39	III
DSCR1	AY325903.1	.93	7.13	−.10	.74	III
DSCR10	AB066291.1	.95	8.60	−.07	.25	III
DSCR2	AY463963.1	1.25	10.98	.32	.18	III
DSCR3	D87343.1	1.40	10.09	.48	.12	III
DSCR6	AB037158.1	1.03	7.37	.04	.41	III
DSCR9	BC066653.1	1.05	7.69	.07	.28	III
ETS2	J04102.1	1.40	7.18	.49	.43	III
GABPA	BC035031.2	1.40	8.88	.49	.08	III
GART	X54199.1	1.17	9.55	.23	.15	III
H2BFS	AB041017.1	1.13	13.19	.17	.57	III
HLCS	AB063285.1	1.32	8.48	.40	.18	III
ICOSLG	AF289028.1	1.23	9.99	.30	.28	III
JAM2	AY016009.1	.98	8.07	−.03	.82	III
KCNE1	BC046224.1	1.12	7.18	.16	.40	III
KIAA0179	D80001.1	1.21	10.15	.28	.24	III
MORC3	BC094779.1	1.34	8.51	.42	.15	III
n74695	N74695	.93	9.83	−.10	.18	III
PKNOX1	AY196965.1	1.04	7.81	.05	.12	III
POFUT2	NM_015227.3	1.35	7.46	.43	.49	III
PRED21	AP001693.1	.95	7.75	−.07	.66	III
PRED24	AP001695.1	1.00	7.41	.00	.65	III
PRED41	AP001726.1	.93	7.15	−.11	.64	III
PRED59	AL163301.2	.94	7.94	−.09	.17	III
PRED63	AP001759.1	.99	7.56	−.01	.18	III
PRED65	AL163202.2	1.06	7.10	.08	.25	III
PRMT2 (exon 5/6)	U80213.1	1.34	8.67	.42	.14	III
PWP2H	U56085.1	1.34	9.43	.42	.11	III
RUNX1	D43968.1	.85	7.91	−.23	.76	III
SETD4	AF391112.1	1.09	9.24	.13	.21	III
SH3BGR	X93498.1	1.07	7.19	.10	.83	III
SLC19A1	AF004354.1	1.20	7.80	.26	.14	III
SOD1[b]	AY835629.1	1.15	9.55	.21	.21	III
SYNJ1	AF009039.1	1.29	8.07	.37	.13	III
TFF3	BC017859.1	.97	8.00	−.04	.46	III
TMEM1	BC101728.1	1.27	10.52	.34	.14	III
TMEM50B	AF045606.2	1.38	10.07	.46	.20	III
U2AF1	M96982.1	1.27	12.15	.35	.06	III
UBASH3A	AJ277750.1	1.13	7.32	.17	.21	III
UBE2G2	AF032456.1	1.17	12.08	.22	.18	III
W90635	W90635.1	1.09	7.39	.12	.45	III
WRB	BC012415.1	1.21	8.21	.27	.29	III
ZNF295	BC063290.1	1.19	8.78	.25	.27	III
ABCG1	AY048757.1	1.25	8.11	.32	.79	IV
ADARB1	AY135659.1	1.26	7.28	.34	3.06	IV
as-MCM3AP-C21orf85	AW163084.1	1.41	7.71	.50	1.02	IV
BRWD1	AB080586.1	1.44	9.14	.53	.38	IV
C21orf22	AY040089.1	1.27	7.56	.34	.88	IV
C21orf8	AA843704.1	1.09	7.55	.12	1.32	IV
CBR1	AB124848.1	1.47	8.20	.55	.76	IV
COL6A1	NM_001848.2	1.60	7.34	.68	.62	IV
CSTB	AF208234.1	1.35	12.79	.43	.38	IV

(*continued*)

Table 5. (continued)

Gene Symbol	GenBank Accession Number	Ratio	*A*	*M*	Var(*M*)	Class[a]
DCR1-12.0	AJ001868.1	1.28	7.37	.36	.76	IV
DCR1-12.0-reverse	AJ001868.1	1.44	8.01	.53	1.06	IV
DCR1-13.0	AJ001869.1	1.25	9.20	.32	1.26	IV
DCR1-13.0-reverse	AJ001869.1	1.14	8.98	.19	.91	IV
DCR1-15.0	AJ001872.1	1.28	7.08	.36	1.00	IV
DSCR4	DQ179113.1	1.32	7.41	.40	.53	IV
MX1	AF135187.1	1.49	13.61	.58	.67	IV
MX2	M30818.1	1.33	11.94	.41	.48	IV
PRDM15	AF426259.1	1.38	8.88	.47	.51	IV
PRMT2 (exon 8/9)	U80213.1	1.46	8.73	.55	.58	IV
TRPM2	AY603182.1	1.08	7.50	.11	1.82	IV

NOTE.—The value *A* corresponds to $[\log_2(\text{DS}) + \log_2(\text{control})]/2$ for the corresponding gene across the 40 hybridizations, *M* corresponds to the mean of $\log_2(\text{DS}) - \log_2(\text{control})$ for the corresponding gene across the 40 hybridizations, Var (*M*) is the variance of *M*, and the DS/control ratio is equal to 2^M.

[a] Class I corresponds to genes expressed proportionally to the gene-dosage effect in DS cell lines, class II contains genes that are amplified, class III contains genes that are compensated, and class IV contains genes that are highly variable between individuals.

[b] Oligonucleotide probes mapped to the long isoform of *SOD1*. See details in the "Discussion" section.

strength of the results, where possible, at least two probes per gene were designed (~80% of the HSA21 oligoarray content). The description of the HSA21 oligoarray content according to BLAST results performed on the latest version of the human genome sequence (NCBI Gene Database build 36.2) is summarized in table 3. Oligonucleotide sequences spotted on the array have been designed on the basis of four main criteria: specificity for the represented sequence, GC content equal to 50%, melting temperature allowing an optimal match between probe and target at the hybridization temperature, and no stable predicted secondary structure.

By use of this new specific high-content HSA21 oligoarray, 40 differential hybridizations comparing DS and control LCLs were performed. The mean signal intensities (represented in $\log_2$ by the *A* value) of each array spot indicated the expression levels of chromosome 21 genes. A total of 134 genes gave signal intensities above the background cutoff (mean $A > 7$).

Biological Material from Patients with DS and Control Individuals

LCLs were obtained after immortalization by EBV of B lymphocytes collected from blood samples of individuals with DS and control individuals. To make sure that EBV transformation did not induce any chromosomal rearrangement, all cell lines were karyotyped after immortalization. Cell lines were always maintained in exponential growth phase. No significant difference in cell morphology or cell proliferation was observed between DS and control LCLs.

For three individuals with DS, transcriptome comparisons between fresh blood samples and LCLs obtained from the same individuals were conducted on pangenomic microarrays.[34] From these experiments, no major alteration of the transcriptome by the EBV transformation could be detected; only 0.5% of the genes exhibited significant differential expression ($P = .01$) (L.D., E.A.Y.-G., and M.-C.P., unpublished data).

Experimental Design and Statistical Analysis

The main objective was to detect differentially expressed genes between DS and control samples, taking into account the sex of and variability between individuals. The aim of the experimental design was to adapt to the experimental constraints of the study (see table 1 and the "Material and Methods" section). Forty experiments were thus programmed.

First, a mixed model was constructed to highlight the effects of DS, sex, and DS × sex and to take into account the gene-expression variability between individuals. We used the Benjamini-Hochberg procedure to adjust the *P* values obtained and to limit false-positive results due to multiple testing.[31] The FDR was set at 0.05. The list of significant genes was thus expected to contain 5% false-positive results.

Chromosome 21 genes did not have significant *P* values when sex or DS and sex combined (DS × sex) were tested. In other words, chromosome 21 gene expression was not significantly different between men and women. In addition, DS effects on gene expression were not dependent on sex. The effects of sex and DS × sex were thus dropped from the model, and a simplified mixed model testing the effects of DS on HSA21 gene expression was ultimately used. We first selected genes that had DS/control ratios significantly different from 1 (FDR 0.05), using the Benjamini-Hochberg procedure.[31] Among the 136 expressed transcripts (134 genes), about half (58) had DS/control

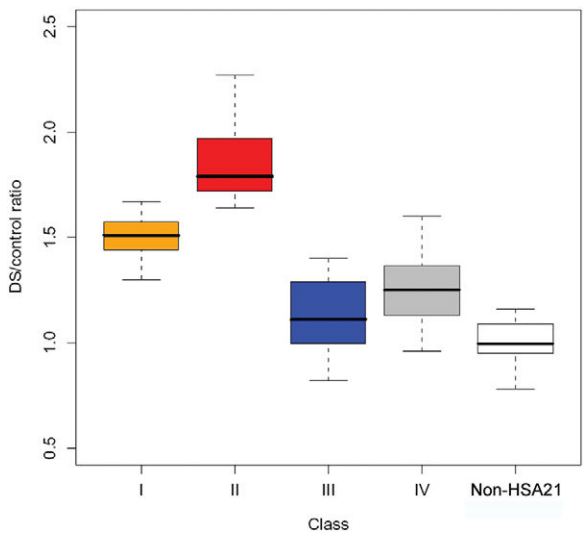


Figure 2. Distribution of DS/control ratios for class I, II, III, and IV genes and non-HSA21 reference genes. The plot represents the minimum and maximum values (*whiskers*), the first and third quartiles (*box*), and the median value (*midline*) of DS/control ratios for each class of genes.

ratios different from 1 and always >1, indicating that these genes were significantly overexpressed in DS LCLs. Expression ratios of these 58 genes ranged from 1.25 to 2.27, with a mean of 1.5, corresponding to the gene-dosage effect in DS. In parallel, among the 134 expressed genes, we selected those that deviated from this gene-dosage effect with a ratio significantly different from 1.5. Surprisingly, the majority of expressed transcripts (86) had DS/control ratios significantly different from 1.5 (FDR 0.05). Because of this high number, the method described by Storey et al.[32] for adjustment of the FDR had to be applied. Results showed that 86 genes had DS/control ratios significantly different from 1.5. Of these 86 genes, 9 had DS/control ratios >1.5, in the range 1.64–2.27, and 77 had DS/control ratios <1.5, in the range 0.74–1.4.

On the basis of this statistical analysis, we classified genes into four categories according to their variation of expression between DS and control LCLs, as described in table 4 and represented in figure 1. Class I contained 30 genes with DS/control ratios significantly different from 1 but not significantly different from 1.5, in the range 1.3–1.67. Class II contained nine genes that were significant in both statistical tests, with DS/control ratios significantly different from 1, significantly different from 1.5, and >1.5 (range 1.64–2.27). Class III comprised 77 genes that had DS/control ratios significantly different from 1.5 and <1.5 (range 0.74–1.4). The majority of gene predictions and antisense transcripts (77%) belonged to this class. In addition, gene expression levels of the transcripts belonging to class III were significantly lower than those belonging to class I ($P = 1.32 \times 10^{-5}$) and class II ($P = 5.7 \times 10^{-7}$). Class IV included the remaining 20 genes with DS/control ratios not significantly different from 1 or from 1.5, in the range 1.08–1.6. Table 5 gives the complete list of genes. Distributions (box plots) of DS/control ratios for each class and for non–chromosome 21 reference genes are shown in figure 2.

The goal of this study was also to demonstrate whether chromosome 21 gene-expression profiles could differentiate DS from control samples. We therefore performed two distinct PCAs on the 134 chromosome 21–expressed genes and the 39 non–chromosome 21 genes used as references (see the "Material and Methods" section). PCA could clearly distinguish individuals with DS from control individuals, suggesting that the effects of DS on chromosome 21 gene expression prevails over any other effect, including biological variability (fig. 3*A*). In addition, no distinction could be obtained between individuals with DS and control individuals when PCA was conducted on non–chromosome 21 genes (fig. 3*B*). Non–chromosome 21 genes had a mean DS/control expression ratio of 1 (fig. 2).

Most of the genes present on the HSA21 oligoarray were represented by two probes. When the two probes were found to be expressed, they belonged to the same class, except for *C21ORF108* and *PRMT2*. Concerning *C21ORF108*, one probe (*B1+KIAA0539.eri10102_a*) mapping to exon 39 of *C21ORF108* belonged to class I (DS/control ratio 1.3). The other probe (*KIAA0539.gff6561_b*), mapping to exon 26 of *C21ORF108*, was in class III (DS/control ratio 1.09). This difference could result from the existence of two alternative transcripts containing either exon 26 or exon 39. Similarly, the two probes representing *PRMT2* (*HRMT1L1.gff2216_a* and *HRMT1L1.gff2216_b*) belonged to classes IV and III, respectively. This difference could also be explained by the existence of two alternative transcripts containing either exons 5/6 or exons 8/9 described in the ENSEMBL database.

QPCR Validation Experiments

To confirm variations in gene expression and to validate the classification of chromosome 21 genes, we performed QPCR on 11 genes belonging to class I (1 gene), class II (2 genes), class III (6 genes), and class IV (2 genes). QPCR was conducted on all LCLs from individuals with DS and control individuals. Ratios obtained by QPCR confirmed the classification of chromosome 21 genes deduced from HSA21 oligoarrays, except for *SOD1*. *SOD1* belonged to class III and had a ratio of 1.57 by QPCR. However, this 1.57-fold difference between LCLs from individuals with DS and control individuals was not significant ($P = .15$). DS/control ratios from QPCR were in agreement with ratios obtained from HSA21 oligoarrays (table 6), with a correlation coefficient of 0.82. Our HSA21 oligoarray was thus a comprehensive, reproducible, and sensitive tool for studying gene expression in DS.

Discussion

The aim of the study was to analyze chromosome 21 gene expression in LCLs from individuals with DS and control individuals. Forty differential hybridizations comparing DS LCLs with control LCLs were performed on a dedicated HSA21 oligoarray designed from the complete human chromosome 21 gene catalogue (359 genes). Approximately one-third (134) of all chromosome 21 genes, ORFs, and predictions were expressed in LCLs.

On the basis of the expression levels of chromosome 21 genes, DS samples were clearly distinct from control samples, thus reflecting the prevalent effect of DS on chromosome 21 gene expression. On the contrary, reference genes mapping to chromosomes other than 21 could not distinguish DS LCLs from control LCLs. Using the mixed-model analysis, we were able to detect genes that are significantly overexpressed in DS cell lines (58) and also genes that deviate from the gene-dosage effect, with DS/control expression ratios significantly different from 1.5.

Classification of HSA21 Genes

By use of this new data analysis protocol, human chromosome 21 genes can now be ranked into four classes by their expression levels in DS cell lines relative to controls. This protocol could be applied to expression data obtained from other human tissues, to validate the classification.

Class I contains 30 genes with expression ratio of DS/control close to 1.5 (range 1.3–1.67), correlated to the presence of three genomic copies (table 4 and fig. 1). These class I genes could be responsible for the phenotype observed in DS, either directly or indirectly through a secondary effect of *cis*- or *trans*-acting genes.

Class II contains nine genes with expression ratio of DS/control >1.64, corresponding to an amplification of the initial gene dosage (table 4 and fig. 1). Among these genes, *SAMSN1, SNF1LK, STCH,* and *BTG3* show the highest expression ratio, in the range 1.67–2.27.

Gene-dosage amplification could result from a cascading effect through regulation networks involving *trans*- or *cis*-acting genes.[35] Pellegrini et al.[36] identified *in silico* a putative mitogen-activated kinase cascade with chromosome 21 kinases involved in various signaling pathways: *DYRK1A, SNF1LK, RIPK4,* and *DSCR3*. In our study, *DYRK1A, SNF1LK,* and *DSCR3* were expressed in LCLs, whereas *RIPK4* was not. Thus, four replicates were chosen for each patient.

DYRK1A is under the gene-dosage effect and *DSCR3* is compensated, whereas *SNF1LK* is amplified from the initial gene dosage. On the basis of this putative mitogen-

Table 6. Comparison between QPCR Results and Microarray Data

HSA21 Gene	GenBank Accession Number	Data from HSA21 Oligoarray		DS/Control Ratio by QPCR[a]
		DS/Control Ratio[b]	Class	
SNF1LK	AB047786.1	2.14	II	3.36
STCH	U04735.1	1.97	II	2.06
MX1	AF135187.1	1.49	IV	1.78
DYRK1A	D86550.1	1.41	I	1.77
CSTB	AF208234.1	1.35	IV	1.49
CHAF1B	U20980.1	1.29	III	1.38
TMEM1	BC101728.1	1.27	III	1.27
GART	X54199.1	1.17	III	1.38
SOD1	AY835629.1	1.15	III	1.57
H2BFS	AB041017.1	1.13	III	1.05
DSCR1	AY325903.1	.93	III	1.11

[a] The mean expression ratio for the corresponding gene between DS and control cell lines.

[b] DS/control ratio by QPCR was calculated from normalized C_t obtained for DS cell lines relative to control cell lines.

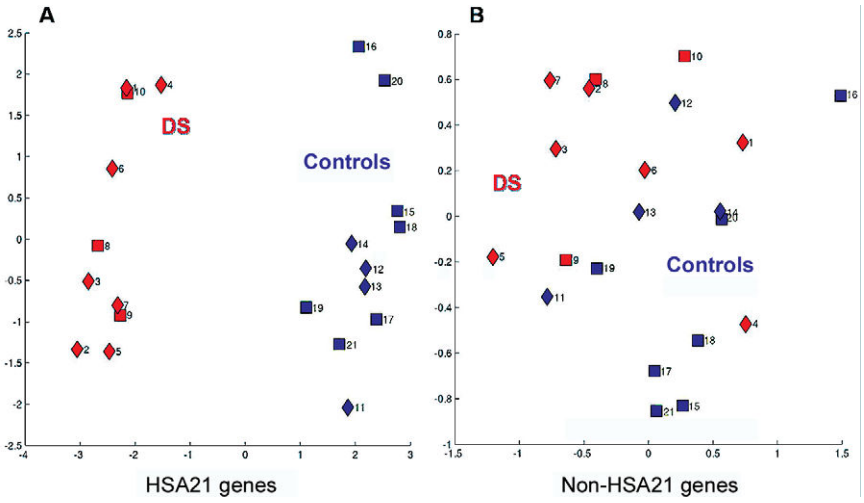


Figure 3. PCA of HSA21 genes (*A*) and non-HSA21 genes (*B*). Red and blue symbols represent DS and control samples, respectively. Squares represent samples extracted from females, and diamonds represent samples extracted from males.

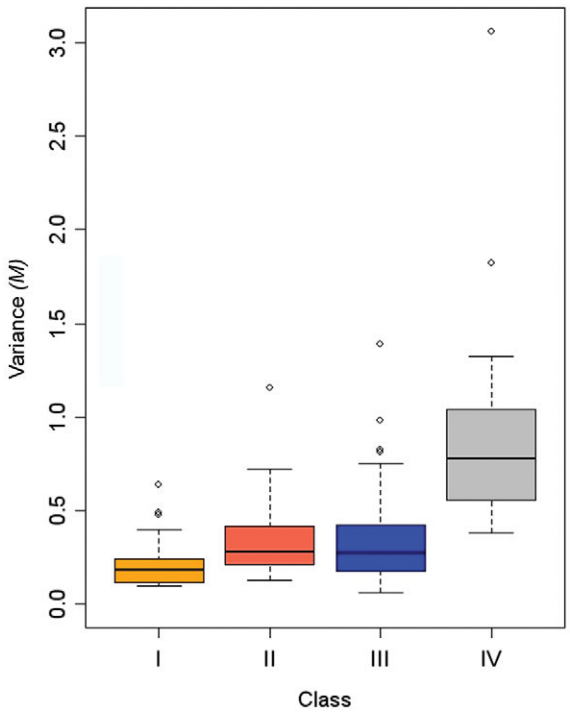


Figure 4. Distribution of the variance of *M* for class I, II, III, and IV genes. The plot represents the minimum and maximum values (*whiskers*), the first and third quartiles (*box*), and the median value (*midline*) of the variances of *M* for each class of genes, where *M* is the mean of $\log_2(\text{DS}) - \log_2(\text{control})$.

activated kinase cascade, amplification of *SNF1LK* gene expression could thus result from the overexpression of DYRK1A acting as a regulatory factor on *SNF1LK* in the cascade, and DSCR3 could act as a scaffolding protein.

Class III is the most abundant and contains 77 genes, with a large proportion of gene predictions and antisense transcripts with DS/control expression ratio <1.4 (table 4 and fig. 1). These class III genes are likely to be compensated in DS. Compensation mechanisms in trisomic conditions have been described in maize and *Drosophila*[37–39] and have been suggested in previous transcriptome studies, both in patients with DS[10,12,16] and in mouse models.[4,19–21] For example, Lyle et al.[20] found that 45% of the triplicated genes analyzed in their study were compensated. Compensation is most likely due to negative feedbacks that would modulate transcriptional activity or mRNA stability of class III genes. Thus, expression of compensated genes could be regulated by mechanisms that are not impaired in DS. For example, *trans*-inhibitors could act directly on the level of expression of these genes. Alternatively, *trans*-activators would activate inhibitors present in three copies on chromosome 21 and would reduce the expression level of target genes that could be also be present on chromosome 21.[40] However, the existence of polymorph alleles correlated to different levels of expression should not account for either gene compensation or amplification in a representative population of patients with DS. A recent study has demonstrated that two CpG islands from human chromosome 21 can be methylated monoallelically.[41] One of those maps to *DSCR3,* the other to *C21orf29.* Both are class III genes in LCLs.

Six class III genes were tested by QPCR, and all were validated, except *SOD1. SOD1* is a well-characterized gene that has been shown elsewhere to be overexpressed in DS tissues and cells at the RNA and protein levels.[13,16,42] In LCLs, the *SOD1* gene is transcribed into two variants, a long and a short isoform.[43] Since *SOD1* probes from the HSA21 oligoarray mapped to the long isoform only, the classification (class III compensated) (table 5) corresponded to this long isoform. The ratio deduced from the HSA21 oligoarray (1.15) was found to be slightly lower than the one obtained by QPCR for the long isoform (1.57). However, by use of QPCR primers amplifying both isoforms of *SOD1,* with the short isoform the most abundant in LCLs, we found that the ratio between DS and control LCLs was 1.96 (data not shown). These results suggest that, in DS LCLs, *SOD1* is overexpressed.

Class IV contains 15 genes and 5 gene predictions that have DS/control expression ratio not different from either 1 or 1.5. These class IV genes are thus highly variable between individuals with DS and control individuals. Indeed, figure 4 shows that the variance distribution of expression ratios is the highest for class IV genes. Three class IV genes (*CBR1, PRDM15,* and *ADARB1*) were shown elsewhere to be highly variable among unaffected individuals.[44]

Using the mixed-model analysis, we have been able to distinguish between gene expression differences resulting from DS and those from interindividual variations. Interindividual variations have been assessed in normal LCLs.[45–47] In the present study, we used lymphoblastoid cells established from individuals with DS and control individuals all belonging to Indo-European populations, thus limiting the variations due to ethnic groups.

Copy-number variations have also been described in LCLs.[48] They should not have an impact on the results unless their frequencies are different in individuals with DS and control individuals, which is unlikely. Moreover, we could not find any correlation between the type of copy-number variation (gain or loss) described for particular genes and their gene class. For example, two genes mapping to the same copy-number variant (variation 5162[48]) belonged to class II (*PDXK*) and class IV (*CSTB*).

Comparison with Expression Data Obtained from DS Tissues

Mao et al.[16] studied transcriptome modifications in DS fetal heart, cerebellum, and astrocyte cells, using a pangenomic Affymetrix U133A chip. Of the 200 genes assigned to HSA21, 23 were significantly changed in DS tissues and 17 were in common with the 58 HSA21 genes that were significantly changed in our study. Our results

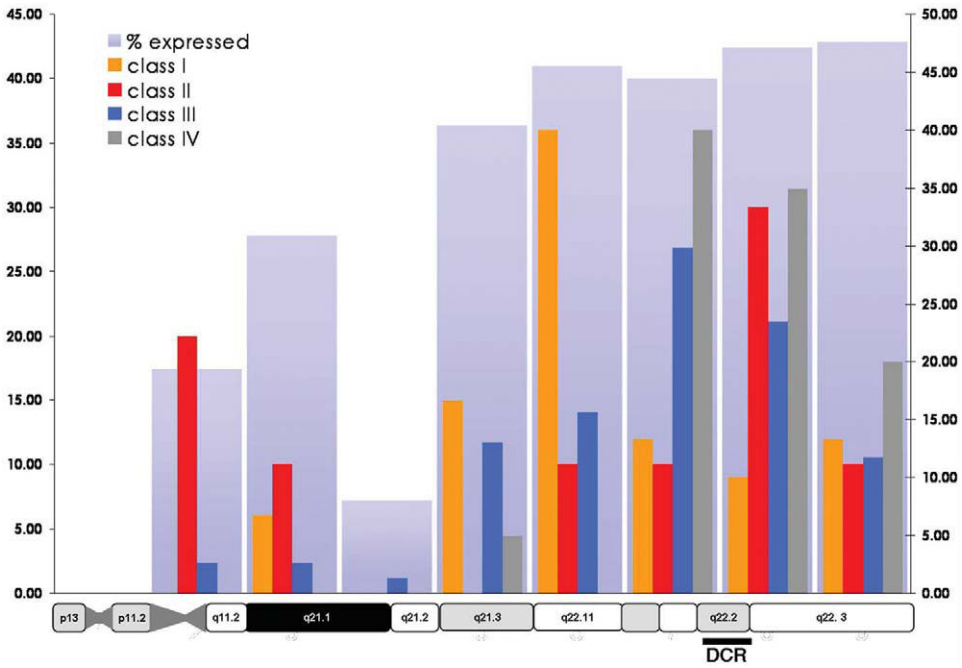


Figure 5. Distribution of expressed, class I, II, III, and IV genes along HSA21. The left *Y*-axis indicates the proportion of expressed genes in each 5-Mb interval, and the right *Y*-axis indicates the proportion of class I, II, III, and IV genes in each 5-Mb interval.

are also in agreement with another gene-expression study performed on DS fetal heart cells[17] that showed that 16 HSA21 genes are significantly overexpressed in fetal hearts. Among these 16 genes, 8 were significantly changed in our study. Class I, II, and III genes were present in all tissues, suggesting that gene-dosage effect, amplification, and compensation are general phenomena and that LCLs are a good model for studying gene-dosage effects. The differences observed between our study and the others suggest tissue-specific regulations that have been described elsewhere for the control of *GABPα* expression.[49]

Distribution of Gene-Expression Modifications along HSA21

We have analyzed the distribution of expressed genes, as well as individual gene classes along HSA21. Figure 5 shows that expressed genes map preferentially to the distal part of HSA21, reflecting the nonuniform gene density along HSA21q.[27] The most telomeric region of HSA21 has a high proportion of class III genes, perhaps because of the presence of a higher proportion of gene predictions that are localized inside gene introns.

DS Effects on Alternative Transcripts

To search for differential effects of DS on alternatively spliced transcripts, we analyzed genes for which oligonucleotide probes present on the HSA21 oligoarray could differentiate between alternative transcripts. Seventeen genes had probes specific to alternative transcripts (table 7). For seven of those genes, all the probes gave a very low signal, indicating that these transcripts are not expressed in LCLs. Three genes (*C21orf33, C21orf34,* and *MRPL39*) were expressed in LCLs as a unique transcript and belonged to class I genes, which are overexpressed with a ratio close to 1.5. For the last seven genes (*ADARB1, C21orf66, DYRK1A, GART, PKNOX1, RUNX1,* and *TMEM1*), oligonucleotide probes could distinguish between splicing variants that had very similar DS/control ratios. Only two of these genes (*C21orf66* and *DYRK1A*) were significantly overexpressed in DS LCLs (i.e., were class I genes), whereas the others were compensated. These results suggest that most of the transcripts belonging to the same gene and expressed in LCLs are similarly regulated in DS.

DS Effects on Antisense Transcripts

The HSA21 oligoarray was also designed to analyze the effects of DS on the expression of antisense transcripts. Fourteen antisense transcripts are present with their nesting genes on the HSA21 oligoarray (table 8). Among them, 10 have been extracted from the HSA21 database established by Kathleen Gardiner at the Eleanor Roosevelt Institute.[50] The four remaining antisense transcripts corresponded to transcribed sequences in the DCR that have been generated from various cDNA mapping and exon-trapping experiments.[8,51,52] Seven genes (*C21orf25, CHAF1B, DYRK1A, HLCS, KIAA0179, MCM3AP,* and *TTC3*) are expressed in LCLs. Four of the corresponding antisense transcripts (*as-DYRK1A, as-HLCS, as-KIAA0179,* and *as-TTC3*) were also found to be expressed in LCLs, thus confirming

Table 7. Alternative Transcripts of HSA21 Genes

Gene Symbol and Probe	GenBank Accession Number	*A*	DS/Control Ratio	*M*	Var(*M*)	Recognized Variants[a]	Class
ADARB1:							
ADARB1.alt23565_a	AY135659.1	5.23	NE	NE	NE	1, 2, 3, 4	NE
ADARB1.alt23565_b	AY135659.1	4.89	NE	NE	NE	1, 2, 3, 4	NE
ADARB1.alt3788_a	AY135659.1	7.36	1.22	.29	2.69	1, 2, 4	IV
ADARB1.alt3788_b	AY135659.1	7.21	1.3	.38	3.46	1, 2, 4	IV
C21orf33:							
HES1.gff1583_b	BC003587.1	5.46	NE	NE	NE	1	NE
bi824121	BI824121.1	10.2	1.5	.58	.16	1, 2	I
HES1.gff1583_a	BC003587.1	10.86	1.53	.62	.25	1, 2	I
C21orf34:							
orf34+35.eri594_a	AF486622.1	5.5	NE	NE	NE	1	NE
C21orf34.gff397_a	AF486622.1	5.87	NE	NE	NE	1, 2	NE
C21orf34.gff397_b	AF486622.1	5.83	NE	NE	NE	1, 2	NE
orf34+35.alt629_b	AF486622.1	6.15	NE	NE	NE	1, 2	NE
orf34+35.eri594_b	AF486622.1	5.69	NE	NE	NE	1, 2	NE
C21orf35.gff251_a	AF486622.1	5.95	NE	NE	NE	1, 2, 3	NE
orf34+35.alt2559_a	AF486622.1	8.15	.95	−.07	.30	1, 2, 3	III
orf34+35.alt629_a	AF486622.1	8.2	.9	−.15	.27	1, 2, 3	III
C21orf66:							
B3+GCFC.eri2361_a	AY033903.1	5	NE	NE	NE	1, 2, 3	NE
B3+GCFC.eri2361_b	AY033903.1	5.48	NE	NE	NE	1, 2, 3, 4	NE
GCFC.eri2361_a	AY033903.1	10.6	1.56	.64	.09	1, 2, 3	I
GCFC.eri2361_b	AY033903.1	8.93	1.51	.6	.04	1, 2, 3, 4	I
GCFC.gff1083_a	AY033903.1	10.19	1.48	.57	.22	1, 2, 3	I
GCFC.gff1083_b	AY033903.1	8.29	1.47	.56	.12	1, 2, 3	I
DYRK1A:							
DYRK1.alt2571_a	D86550.1	10.33	1.4	.48	.13	1, 2, 3, 4, 5	I
DYRK1.alt2571_b	D86550.1	10.71	1.4	.49	.22	1, 2, 3, 4, 5	I
DYRK1.gff5318_a	D86550.1	9.47	1.41	.5	.19	1, 2, 3, 4, 5	I
DYRK1.gff5318_b	D86550.1	11.02	1.41	.5	.22	1, 2, 3, 4, 5	I
GART:							
GART.gff3271_a	X54199.1	8.79	1.18	.24	.10	1	III
GART.gff3271_b	X54199.1	10.35	1.17	.22	.20	1, 2	III
MRPL39:							
PRED22.eri707_a	AF109357.1	9.46	1.41	.5	.14	1, 2	I
PRED22.eri707_b	AF109357.1	10.56	1.47	.55	.17	1, 2	I
PRED22.gff1072_a	AF109357.1	10.33	1.47	.56	.13	1, 2	I
PRED22.gff1072_b	AF109357.1	10.33	1.46	.55	.18	1, 2	I
PRED66.eri187_a	AF109357.1	10.28	1.51	.6	.12	1, 2	I
PRED66.eri187_b	AF109357.1	9.36	1.51	.59	.16	1, 2	I
PRED66.gff641_b	AF109357.1	8.31	1.42	.5	.17	1, 2	I
PRED66.gff641_a	AF270511.1	6.58	NE	NE	NE	2	NE
PKNOX1:							
PKNOX1.gff3279_a	AY196965.1	7.64	1.05	.07	.11	1	III
PKNOX1.gff3279_b	AY196965.1	7.98	1.03	.04	.14	1, 2	III
RUNX1:							
RUNX1.alt25714_a	D43968.1	7.23	.86	−.21	.84	1, 2	III
RUNX1.alt7267_a	D43968.1	7.37	.8	−.32	.74	1, 2	III
RUNX1.alt7267_b	D43968.1	5.56	NE	NE	NE	2	NE
RUNX1.gff2722_a	D43968.1	7.92	.85	−.23	.74	2	III
RUNX1.gff2722_b	D43968.1	9.03	.89	−.17	.76	2	III
TMEM1:							
TMEM1.gff5126_a	BC101728.1	10.53	1.25	.32	.14	1	III
TMEM1.gff5126_b	BC101728.1	10.5	1.28	.36	.14	1, 2	III

NOTE.—The value *A* corresponds to $[\log_2(\text{DS}) + \log_2(\text{control})]/2$ for the corresponding gene across the 40 hybridizations, *M* corresponds to the mean of $\log_2(\text{DS}) - \log_2(\text{control})$ for the corresponding gene across the 40 hybridizations, Var(*M*) is the variance of *M,* and the DS/control ratio is equal to 2^M. NE = not expressed.

[a] The number of transcript variants hybridizing to the oligonucleotide probe.

Table 8. Sense and Antisense Transcripts on Chromosome 21

Gene Symbol and Probe	GenBank Accession Number	Intragenic Location[a]	*A*	*M*	Var(*M*)	DS/Control Ratio	Class
C21orf56:							
C21orf56.gff691_a	BC084577.1	Exon 2	5.51	NE	NE	NE	NE
as-C21orf56:							
C21orf56.gff284_a	BC084577.1	Exon 4	10.68	.12	.08	1.08	III
C21orf56.gff284_b	BC084577.1	Exon 4	12.79	.08	.08	1.05	III
CHAF1B:							
CHAF1B.gff2194_a	U20980.1	Exon 14	8.66	.37	.12	1.29	III
CHAF1B.gff2194_b	U20980.1	Exon 14	8.03	.38	.14	1.30	III
as-CHAF1B:							
BF740066	BF740066.1	3′	5.08	NE	NE	NE	NE
HLCS:							
HLCS.gff6722_a	AB063285.1	Exon 12	6.37	NE	NE	NE	NE
HLCS.gff6722_b	AB063285.1	Exon 11	8.48	.40	.18	1.32	III
as-HLCS:							
DCR1-8.0_a	AJ001862.1	Intron 7	7.58	−.03	.42	.98	III
DCR1-8.0_b	AJ001862.1	Intron 7	5.83	NE	NE	NE	NE
TTC3:							
TTC3.gff9074_a	D84296.1	Exon 47	10.02	.84	.23	1.79	II
TTC3.gff9074_b	D84296.1	Exon 34	11.17	.83	.17	1.78	II
as-TTC3:							
bf979681	BF979681.2	Exon 41	9.07	.64	.64	1.56	I
DYRK1A:							
DYRK1.alt2571_a	D86550.1	Exon 13	10.33	.48	.13	1.40	I
DYRK1.alt2571_b	D86550.1	Exons 7/8	10.71	.49	.22	1.40	I
DYRK1.gff5318_a	D86550.1	Exon 13	9.47	.50	.11	1.41	I
DYRK1.gff5318_b	D86550.1	Exon 11	11.02	.50	.22	1.41	I
as-DYRK1A:							
DCR1-12.0_a	AJ001868.1	Intron 1	7.37	.36	.76	1.29	IV
DCR1-13.0-RC_a	AJ001869.1	Intron 1	8.33	.15	.95	1.11	IV
DCR1-13.0-RC_b	AJ001869.1	Intron 1	9.55	.22	.91	1.17	IV
KCNJ6:							
GIRK2(U52153)_a	U52153.1	Exon 3	5.60	NE	NE	NE	NE
GIRK2(U52153)_b	U52153.1	Exon 1	5.28	NE	NE	NE	NE
as-KCNJ6:							
DCR1-17DCR1-17_a	AJ001875.1	Intron 3	7.16	.17	.70	1.12	IV
ADAMTS5:							
ADAMTS5.gff5523_a	AF142099.1	Exon 8	5.39	NE	NE	NE	NE
ADAMTS5.gff5523_b	AF142099.1	Exon 8	5.39	NE	NE	NE	NE
as-ADAMTS5:							
r18879	R18879.1	Intron 3	5.69	NE	NE	NE	NE
as-C21orf25:							
aa575913	AA575913.1	3′	5.85	NE	NE	NE	NE
C21orf25:							
C21orf25.gff6217_a	AB047784.1	Exon 14	8.96	.17	.45	1.13	III
C21orf25.gff6217_b	AB047784.1	Exon 14	8.24	.32	.17	1.25	III
as-CBR3:							
bi836686	BI836686.1	Exon 3	5.90	NE	NE	NE	NE
CBR3:							
CBR3.gff878_a	AB124847.1	Exon 3	6.18	NE	NE	NE	NE
CBR3.gff878_b	AB124847.1	Exons 1/2	5.31	NE	NE	NE	NE
CLDN14:							
CLDN14.gff1693_a	AF314090.1	Exon 3	6.78	NE	NE	NE	NE
CLDN14.gff1693_b	AP001726.1	3′	6.05	NE	NE	NE	NE
as-CLDN14:							
w90592	W90592.1	3′	5.71	NE	NE	NE	NE
KIAA0179:							
KIAA0179.gff4984_a	D80001.1	Exon 16	5.26	NE	NE	NE	NE
KIAA0179.gff4984_b	D80001.1	Exon 16	10.15	.28	.24	1.22	III
as-KIAA0179:							
aa425659	AA425659.1	3′	8.08	.22	.67	1.17	III
MCM3AP:							
MCM3.gff6113_a	AY590469.1	Exon 27	11.69	.54	.19	1.45	I
MCM3APAS:							
af426262	AF426262.1	Introns 25–26	5.54	NE	NE	NE	NE
af426263	AF426263.1	Introns 20–21	5.73	NE	NE	NE	NE

NOTE—The value *A* corresponds to $[\log_2(\text{DS}) + \log_2(\text{control})]/2$ for the corresponding gene across the 40 hybridizations, *M* corresponds to the mean of $\log_2(\text{DS}) - \log_2(\text{control})$ for the corresponding gene across the 40 hybridizations, Var(*M*) is the variance of *M*, and the DS/control ratio is equal to 2^M. NE = not expressed.

[a] The exon or intron to which the oligonucleotide probe maps.

their existence. *TTC3* (class II) and its antisense transcript (class I) were overexpressed, whereas the other antisense sequences did not belong to the same class as their corresponding genes. In addition, two antisense sequences (*as-C21orf56* and *as-KCNJ6*) were expressed in LCLs, whereas their corresponding genes were not. The probe referred to as antisense transcript *as-KCNJ6* mapping in intron 3 of *KCNJ6,* on the opposite strand, corresponds to one of the transcribed sequences isolated in the DCR by exon-trapping experiments.[8] Since there is no evidence that this sequence is an antisense transcript of *KCNJ6,* it could thus belong to a gene locus that has not yet been identified and might map to the opposite orientation of *KCNJ6*.

According to the NCBI Gene Database, *C21orf56* (accession number 84221) currently maps on the negative strand of HSA21 but was previously annotated on the positive strand when the HSA21 oligoarray was designed. Thus, probe sequence representing the antisense transcript *as-C21orf56* could correspond to the actual sense transcript of *C21orf56*. Therefore, the expression of antisense transcripts is confirmed by our HSA21 oligoarray experiments. Sense and antisense transcripts are not always similarly changed in DS.

In conclusion, using our new high-content HSA21 oligoarrays combined with a new powerful statistical analysis protocol, we were able to classify HSA21 genes according to their level of expression in DS LCLs. We show that, among the expressed transcripts, 29% are sensitive to the gene-dosage effect or are amplified, 56% are compensated, and 15% are highly variable among individuals. Thus, most of the chromosome 21 genes are compensated for the gene-dosage effect. Gene-expression variations in DS are controlled by mechanisms involving *trans* and *cis* regulators acting either directly or through gene-regulation networks. Overexpressed genes are likely to be involved in the DS phenotype, in contrast to the compensated genes. Highly variable genes could account for phenotypic variations observed in patients. Finally, we show that alternative transcripts belonging to the same gene are similarly regulated in DS, whereas sense and antisense transcripts are not always similarly regulated. Studies of human tissues by use of the same analysis protocol will validate genes that are involved in the DS phenotype.

Acknowledgments

We thank the Banque de Cellules from the Cochin Hospital; Drs. N. Faucon and R. Meloni (CNRS UMR 9923, Hôpital Pitié Salpetrière, Paris), for advice; D. Leclerre, C. Mikonio, and F. Richard, for their help with spotting the HSA21 oligoarrays; I. Haddad, for the design of the first version of the HSA21 oligoarray; C. J. Epstein (University of California, San Francisco), for very important comments on the manuscript; and R. Veitia (Hôpital Cochin, Paris), J.-J. Daudin (Institut National d'Agronomie Paris-Grignon, Paris), Dr. Y. Hérault (CNRS, Orléans), and Dr. P. M. Sinet (Institut National de la Santé et de la Recherche Médicale, Paris), for helpful discussions about the results. E.A.G. had a fellowship from the Ministère de la Recherche, and G.G had a fellowship from the Fondation Jérôme Lejeune. This work was supported by the Fondation Jérôme Lejeune, European Economic Community grant T21 Targets, and AnEUpolidy.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

Eleanor Roosevelt Institute: Chromosome 21 Gene Function and Pathway Database, http://chr21db.cudenver.edu/

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for accession numbers in tables 5–8)

Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/ (for accession number GSE6408)

Max Planck Institute: Chromosome 21 Gene Catalog Based on the New AGP File July 2002, http://chr21.molgen.mpg.de/chr21_catalogs/chr21_mar_2002.html

NCBI Entrez, http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi (for accession numbers L13852 and AB000468)

NCBI Gene Database, http://www.ncbi.nlm.nih.gov/sites/entrez (for accession number 84221)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for DS)

The R Project for Statistical Computing, http://www.r-project.org/

References

1. Epstein CJ, Korenberg JR, Anneren G, Antonarakis SE, Ayme S, Courchesne E, Epstein LB, Fowler A, Groner Y, Huret JL, et al (1991) Protocols to establish genotype-phenotype correlations in Down syndrome. Am J Hum Genet 49:207–235
2. Antonarakis SE, Lyle R, Dermitzakis ET, Reymond A, Deutsch S (2004) Chromosome 21 and Down syndrome: from genomics to pathophysiology. Nat Rev Genet 5:725–738
3. Lejeune J, Gautier M, Turpin R (1959) Etudes des chromosomes somatiques de neuf enfants mongoliens. C R Hebd Seances Acad Sci 248:1721–1722
4. Saran NG, Pletcher MT, Natale JE, Cheng Y, Reeves RH (2003) Global disruption of the cerebellar transcriptome in a Down syndrome mouse model. Hum Mol Genet 12:2013–2019
5. Shapiro BL (2001) Developmental instability of the cerebellum and its relevance to Down syndrome. J Neural Transm Suppl 61:11–34
6. Schinzel A (2001) Catalogue of unbalanced chromosome aberrations in man. Walter de Gruyter, Berlin
7. Olson LE, Roper RJ, Sengstaken CL, Peterson EA, Aquino V, Galdzicki Z, Siarey R, Pletnikov M, Moran TH, Reeves RH (2007) Trisomy for the Down syndrome "critical region" is necessary but not sufficient for brain phenotypes of trisomic mice. Hum Mol Genet 16:774–782
8. Dahmane N, Ghezala GA, Gosset P, Chamoun Z, Dufresne-Zacharia MC, Lopes C, Rabatel N, Gassanova-Maugenre S, Chettouh Z, Abramowski V, et al (1998) Transcriptional map of the 2.5-Mb CBR-ERG region of chromosome 21 involved in Down syndrome. Genomics 48:12–23
9. Delabar J, Theophile D, Rahmani Z, Chettouh Z, Blouin J, Prieur M, Noel B, Sinet P (1993) Molecular mapping of twenty-four features of Down syndrome on chromosome 21. Eur J Hum Genet 1:114–124
10. FitzPatrick DR, Ramsay J, McGill NI, Shade M, Carothers AD,

Hastie ND (2002) Transcriptome analysis of human autosomal trisomy. Hum Mol Genet 11:3249–3256
11. Gross SJ, Ferreira JC, Morrow B, Dar P, Funke B, Khabele D, Merkatz I (2002) Gene expression profile of trisomy 21 placentas: a potential approach for designing noninvasive techniques of prenatal diagnosis. Am J Obstet Gynecol 187:457–462
12. Mao R, Zielke CL, Zielke HR, Pevsner J (2003) Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. Genomics 81:457–467
13. Giannone S, Strippoli P, Vitale L, Casadei R, Canaider S, Lenzi L, D'Addabbo P, Frabetti F, Facchin F, Farina A, et al (2004) Gene expression profile analysis in human T lymphocytes from patients with Down syndrome. Ann Hum Genet 68: 546–554
14. Tang Y, Schapiro MB, Franz DN, Patterson BJ, Hickey FJ, Schorry EK, Hopkin RJ, Wylie M, Narayan T, Glauser TA, et al (2004) Blood expression profiles for tuberous sclerosis complex 2, neurofibromatosis type 1, and Down's syndrome. Ann Neurol 56:808–814
15. Chung IH, Lee SH, Lee KW, Park SH, Cha KY, Kim YS, Lee S (2005) Gene expression analysis of cultured amniotic fluid cell with Down syndrome by DNA microarray. J Korean Med Sci 20:82–87
16. Mao R, Wang X, Spitznagel E, Frelin L, Ting J, Ding H, Kim JW, Ruczinski I, Downey T, Pevsner J (2005) Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. Genome Biol 6:R107
17. Li CM, Guo M, Salas M, Schupf N, Silverman W, Zigman W, Husain S, Warburton D, Thaker H, Tycko B (2006) Cell type-specific over-expression of chromosome 21 genes in fibroblasts and fetal hearts with trisomy 21. BMC Med Genet 7: 24
18. Amano K, Sago H, Uchikawa C, Suzuki T, Kotliarova SE, Nukina N, Epstein CJ, Yamakawa K (2004) Dosage-dependent over-expression of genes in the trisomic region of Ts1Cje mouse model for Down syndrome. Hum Mol Genet 13:1333–1340
19. Kahlem P, Sultan M, Herwig R, Steinfath M, Balzereit D, Eppens B, Saran NG, Pletcher MT, South ST, Stetten G, et al (2004) Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of Down syndrome. Genome Res 14:1258–1267
20. Lyle R, Gehrig C, Neergaard-Henrichsen C, Deutsch S, Antonarakis SE (2004) Gene expression from the aneuploid chromosome in a trisomy mouse model of Down syndrome. Genome Res 14:1268–1274
21. Dauphinot L, Lyle R, Rivals I, Dang MT, Moldrich RX, Golfier G, Ettwiller L, Toyama K, Rossier J, Personnaz L, et al (2005) The cerebellar transcriptome during postnatal development of the Ts1Cje mouse, a segmental trisomy model for Down syndrome. Hum Mol Genet 14:373–384
22. Gardiner K (2006) Transcriptional dysregulation in Down syndrome: predictions for altered protein complex stoichiometries and post-translational modifications, and consequences for learning/behavior genes ELK, CREB, and the estrogen and glucocorticoid receptors. Behav Genet 36:439–453
23. Wester U, Bondeson ML, Edeby C, Anneren G (2006) Clinical and molecular characterization of individuals with 18p deletion: a genotype-phenotype correlation. Am J Med Genet A 140:1164–1171
24. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, et al (2000) The DNA sequence of human chromosome 21. Nature 405:311–319
25. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919
26. Reymond A, Camargo AA, Deutsch S, Stevenson BJ, Parmigiani RB, Ucla C, Bettoni F, Rossier C, Lyle R, Guipponi M (2002) Nineteen additional unpredicted transcripts from human chromosome 21. Genomics 79:824–832
27. Gardiner K, Fortna A, Bechtel L, Davisson MT (2003) Mouse models of Down syndrome: how useful can they be? Comparison of the gene content of human chromosome 21 with orthologous mouse genomic regions. Gene 318:137–147
28. Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet Suppl 32:496–501
29. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637
30. SAS Institute (2000) SAS/STAT software release 8. Cary, NC
31. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300
32. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445
33. Janel N, Christophe O, Ait Yahya-Graison E, Hamelet J, Paly E, Prieur M, Delezoide AL, Delabar JM (2006) Paraoxonase-1 expression is up-regulated in Down syndrome fetal liver. Biochem Biophys Res Commun 346:1303–1306
34. Brigand KL, Russell R, Moreilhon C, Rouillard JM, Jost B, Amiot F, Magnone V, Bole-Feysot C, Rostagno P, Virolle V, et al (2006) An open-access long oligonucleotide microarray resource for analysis of the human and mouse transcriptomes. Nucleic Acids Res 34:e87
35. Potier MC, Rivals I, Mercier G, Ettwiller L, Moldrich RX, Laffaire J, Personnaz L, Rossier J, Dauphinot L (2006) Transcriptional disruptions in Down syndrome: a case study in the Ts1Cje mouse cerebellum during post-natal development. J Neurochem Suppl 1 97:104–109
36. Pellegrini-Calace M, Tramontano A (2006) Identification of a novel putative mitogen-activated kinase cascade on human chromosome 21 by computational approaches. Bioinformatics 22:775–778
37. Birchler JA (1979) A study of enzyme activities in a dosage series of the long arm of chromosome one in maize. Genetics 92:1211–1229
38. Devlin RH, Holm DG, Grigliatti TA (1982) Autosomal dosage compensation in *Drosophila melanogaster* strains trisomic for the left arm of chromosome 2. Proc Natl Acad Sci USA 79: 1200–1204
39. Guo M, Birchler JA (1994) Trans-acting dosage effects on the expression of model gene systems in maize aneuploids. Science 266:1999–2002
40. Birchler JA, Riddle NC, Auger DL, Veitia RA (2005) Dosage balance in gene regulation: biological implications. Trends Genet 21:219–226
41. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, Mukai T, Sakaki Y, Ito T (2004) A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res 14:247–266
42. Sinet PM, Lavelle F, Michelson AM, Jerome H (1975) Super-

oxide dismutase activities of blood platelets in trisomy 21. Biochem Biophys Res Commun 67:904–909
43. Sherman L, Levanon D, Lieman-Hurwitz J, Dafni N, Groner Y (1984) Human Cu/Zn superoxide dismutase gene: molecular characterization of its two mRNA species. Nucleic Acids Res 12:9349–9365
44. Deutsch S, Lyle R, Dermitzakis ET, Attar H, Subrahmanyan L, Gehrig C, Parand L, Gagnebin M, Rougemont J, Jongeneel CV, et al (2005) Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet 14:3741–3749
45. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33:422–425
46. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39:226–231
47. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Key JM (2007) Gene-expression variation within and among human populations. Am J Hum Genet 80:502–509
48. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al (2006) Global variation in copy number in the human genome. Nature 444:444–454
49. O'Leary DA, Pritchard MA, Xu D, Kola I, Hertzog PJ, Ristevski S (2004) Tissue-specific overexpression of the HSA21 gene *GABPα*: implications for DS. Biochim Biophys Acta 1739:81–87
50. Nikolaienko O, Nguyen C, Crinc LS, Cios KJ, Gardiner K (2005) Human chromosome 21/Down syndrome gene function and pathway database. Gene 364:90–98
51. Chen H, Chrast R, Rossier C, Morris MA, Lalioti MD, Antonarakis SE (1996) Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. Genome Res 6: 747–760
52. Ohira M, Seki N, Nagase T, Suzuki E, Nomura N, Ohara O, Hattori M, Sakaki Y, Eki T, Murakami Y, et al (1997) Gene identification in 1.6-Mb region of the Down syndrome region on chromosome 21. Genome Res 7:47–58

Annexe I
Article en préparation présenté dans la section 6.2

Translatome analysis at oocyte-to-embryo transition in sea urchin

Héloïse Chassé[1,2], Sandrine Boulben[1,2], Julie Aubert[3,4], Gildas Le Corguillé[5], Erwan Corre[5], Patrick Cormier[1,2], Julia Morales[1,2*]

[1] CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff cedex, France

[2] Sorbonne Universités, UPMC Univ Paris 06, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff cedex, France

[3] AgroParisTech, UMR518 MIA-Paris, F-75231 Paris Cedex 05, France

[4] INRA, UMR518 MIA-Paris, F-75231 Paris Cedex 05, France

[5] CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France.

* To whom correspondence should be addressed:

Tel: (33) 298292369; Email: morales@sb-roscoff.fr

Keywords: Translational control / polysome profiling / Fertilization / mTOR

Introduction

Protein synthesis is, in all living organisms, one of the most fundamental processes required for cell proliferation, cell differentiation, fast response to stress and environmental cues. Deciphering the machinery underlying protein synthesis as well as the biological functions of the produced proteins thus represent a key step towards a better comprehension of the functional mechanisms of the cell or an organism. Production of proteins from the genome relies on two steps: transcription (the DNA code is transcribed into messenger RNA) and translation (the messenger RNA is translated into protein). The level of a protein within a cell depends on the regulation of messenger RNA (mRNA) abundance by transcriptional regulation, but also on the rates of protein synthesis by the translational machinery as well as on the rate of protein degradation by the lysosome or proteasome pathways. Importantly, recent studies have established that almost half of the variation of protein concentration within a cell is due to the efficiency of translation (Schwanhäusser et al., 2011). Furthermore, translational control has been shown to be critical for protein production in response to a number of physiological and pathological situations, including development (Hershey et al., 2012). The term "translatome" that emerged in 2011 characterizes the subset of mRNAs present in the cell that are actively translated, *i.e.* that are associated with polysomes, thereby providing a very accurate reflection of the functional protein readout of the genome. Translatome analysis is made possible by carrying out polysome profiling coupled with high-throughput sequencing technologies (Kuersten et al., 2013; Larsson et al., 2012), which allows the identification of the sets of proteins that are produced under specific developmental, physiological or pathological conditions (1), deciphering the processes of selection and recruitment of mRNAs to polysomes (2), and establishing translational regulatory networks (3). Early embryonic development is an appropriate model to address the question of translational control. In sea urchins, translational activity is repressed in unfertilized eggs and a stockpile of maternal mRNAs is present in the eggs. From fertilization to the onset of zygotic transcription, maternal mRNAs drive cell cycles and the establishment of cell fates during the first cell divisions, independently of mRNA transcription and ribosome biogenesis. Fertilization triggers the activation of the translation machinery and the recruitment of mRNAs into polysomes, leading to an increase of protein synthesis. Upon fertilization, the stimulation of protein synthesis depends on the release of the cap-binding protein eIF4E from its translational repressor 4E-BP, and its association to the scaffolding protein eIF4G, depending on mTOR signaling (Cormier et al., 2001; Salaun et al., 2003;

Oulhen et al., 2007). Moreover, screens of sea urchin cDNA libraries prepared from polysomal mRNA followed by analyses of their translational status at various developmental stages, revealed the unmasking of a few maternally stored mRNAs and their integration into polysomes (*e.g.* (Alexandraki and Ruderman, 1985; Kelso-Winemiller et al., 1993; Le Breton et al., 2003). Although many studies focus on the analysis of the transcribed mRNAs, no data is currently available on the translationally active dataset of maternal mRNAs at fertilization, which should reflect the functional output of gene expression. To date no large-scale analysis of translated mRNAs in sea urchin is available.

We focused on the oocyte-to-embryo transition, by comparing the translated mRNAs in unfertilized eggs and in 1-cell post-fertilization embryo. We analyzed the recruitment of maternal mRNAs triggered by fertilization, using RNA-sequencing of polysomal fractions. The determination of the embryo translatome allowed us to demonstrate that only a fraction of the maternal mRNAs is recruited after fertilization, with enrichment of biological functions such as “cell cycle”, “signaling” and “RNA-binding proteins”. Furthermore, we showed that all mRNAs translated after fertilization are not equally impacted by mTOR signaling inhibition, and the translation of some mRNAs were not impacted at all, suggesting the existence of a selective translation initiation at fertilization in sea urchin.

Material and methods

Handling and treatment of eggs and embryos

Sea urchin (*Paracentrotus lividus)* were collected in the bay of Crozon (Brittany, France), and maintained in the CRBM facility (Marine Biological Station, Roscoff). Animals were induced to spawn by intracœlomic injection of 0.1M acetylcholine; gametes were collected in filtered sea water (FSW) and kept as a 5% dilution in FSW. Oocytes were dejellied in pH5 water for 45 seconds, washed in fresh FSW and fertilized. Embryos were cultured under constant stirring. For polysome preparation, emetine [100µM] was added to eggs or embryos suspension 5 minutes before collection to freeze elongating polysomes on translated mRNAs, and samples were processed for polysomes analysis. Puromycin [300µg/ml] was added 20 minutes before collection when noted. PP242 [10µM] were added to eggs or embryos suspension when noted (Chassé et al. 2016).

In vivo protein synthesis and western blot analysis

Embryos (5% suspension in seawater) were incubated in presence of [^{35}S]-L-methionine at the final concentration of 5µCi/ml and equivalent aliquots were removed at the indicated times. For pulse-labeling, eggs or embryos taken one hour after fertilization were incubated in presence of [^{35}S]-L-methionine for 15 minutes at a final concentration of 10µCi/ml. [^{35}S]-L-methionine incorporation into proteins was measured on duplicate aliquots after 10% TCA precipitation as described (Costache et al., 2012). Proteins were separated on denaturing SDS-Polyacrylamide gel, and exposed to PhosphorImager screen. Western blot analysis was done after electrotransfer onto nitrocellulose membranes (Amersham). Antibodies directed against sea urchin 4E-BP (1:5000; Oulhen et al., 2010), eIF4E (1:1000; Transduction laboratories) were used in TBS-T / 1% bovine serum albumin. Secondary peroxidase conjugated antibodies (DAKO) were used at 1:10000 dilution in TBS-T / 1% BSA. Revelation was done using Pierce ECL 2 Western blotting substrate (Thermo Scientific) according to the manufacturer's instructions on a Typhoon Imager (GE Healthcare Life Sciences).

Polysome preparation

Equivalent number of eggs or embryos were collected at the indicated time points, and resuspended in polysome lysis buffer (10mM Tris pH7.4, 250mM KCl, 10mM $MgCl_2$, 25mM EGTA, 0.4% Igepal, 5% sucrose, 1mM DTT, 10µg/ml aprotinin, 2µg/ml leupeptin, 40U/ml

RNasin, 100µM emetine), lysed by 10 strokes in Dounce homogenizer, clarified by centrifugation at 16000g for 10 minutes. Lysates were then centrifuged through a 15-40% sucrose gradient in gradient buffer (10mM Tris pH7.4, 250mM KCl, 10mM $MgCl_2$, 25mM EGTA, 1mM DTT) for 2.5 hours at 38000 rpm in a SW41Ti rotor (Beckman). Puromycin-treated samples were incubated with KCl 0.5M 15 minutes at 4°C then 15 minutes at 37°C, before applying onto sucrose gradient. Gradients were collected with an ISCO gradient fractionator, coupled with an optical density recorder, in 21 equivalent fractions. RNA was extracted from each fraction using one volume of phenol pH4 / chloroform (v/v), and isopropanol precipitated in presence of glycogen carrier. RNA pellet was resuspended in RNAse-free water. One tenth of the RNA was used for quality control on agarose gel. RNA was kept at -80°C until further use.

RNA-Seq libraries and sequencing

RNA-Seq data were generated from three independent polysome profiling experiments, each comprising eight different samples as follow: cytoplasm and polysomal fractions from unfertilized eggs and 1-hour (1-cell) post-fertilization embryos, in absence or presence of puromycin (see supplementary **Table. S1**). All eight samples from a biological replicate are taken from eggs or embryos arising from a single female and male pair. The independent experiments correspond to three independent sets of parents. For polysomal RNA samples, RNA from heavy polysomal fractions (prepared as above) were pooled and precipitated with 2vol. ethanol 100% / $1/10^{th}$ vol. sodium acetate 3M, and resuspended in RNAse-free water. For cytoplasmic RNA samples, an aliquot of lysate before polysome fractionation was taken and RNA was extracted and precipitated twice as described for polysomal fraction. Quality and quantity of RNA were assessed by Agilent Bioanalyzer to obtain the RNA integrity numbers (RIN) and concentration. Libraries were constructed with the Illumina Truseq mRNA-stranded kit, and sequenced using 100-base length read chemistry in a paired-end flow cell on Illumina HiSeq 2000. Both library building and sequencing were performed by McGill University and Genome Québec Innovation Centre. Description of the twenty-four libraries and corresponding data generated for this study are provided in Supplementary **Table. S1**.

Assembly of a maternal transcriptome by de novo pipeline

The maternal transcriptome was generated from the cytoplasmic RNAs corresponding to unfertilized and fertilized eggs (three independent samples for each condition). Raw reads were filtered from low-quality sequences, low-complexity sequences and trimmed using

FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Reads were trimmed and filtered using a quality threshold of 25 (base calling) and a minimal size of 60bp. Only reads in which more than 75% of nucleotides had a minimal quality threshold of 20 were retained. Reads were then cleaned from rRNA contaminant using riboPicker (Schmieder et al., 2012) and cleaned from adapter ends using Cutadapt version 1.01. Finally the cleaning process was checked using FastQC (version 0.10.01 http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Sequences from cytoplasmic RNAs were assembled *de novo* using Trinity package (release 2013-02-25; Grabherr et al., 2011). Reads were remapped on the full transcriptome using Bowtie (version 0.12.8; (Langmead et al., 2009) and relative abundances were estimated using RSEM (version 1.2.0) to get the FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) values and thus identify low coverage contigs and rare isoforms (<1%) that were excluded later from the analysis (both software programs were launched through the Trinity package Wrapper filter_fasta_by_rsem_values.pl). For further analysis, the *de novo* transcriptome was filtered to remove transcripts with FPKM <5, leading to a final number of 43823 transcripts. The raw read counts for each of the 6 Illumina libraries corresponding to unfertilized and fertilized eggs cytoplasmic RNAs (2 conditions, three biological replicates per condition) were used as input to the DESeq and EdgeR packages (Anders and Huber, 2010; Robinson et al., 2010) to perform pairwise differential expression analysis cytoplasmic mRNA content before and after fertilization. Transcriptome GO term search was performed using the Trinotate pipeline (Haas et al. 2013). Annotation was done using the *Strongylocentrotus purpuratus* genes Echinobase database (genome version 3.1) (http://echinobase.org; Cameron et al., 2009). The transcriptome was filtered using best blastn hit values $<10^{-5}$ and a final set of 14002 transcripts, corresponding to 7685 Trinity 'genes' was further used for translatome analysis.

Translatome analysis

The raw read counts from the 24 libraries (**Suppl. Table. S1**) were used to analyze the translation profile of each transcript, from the final 14002 transcript set, generating three $\log_2$ fold change values to evaluate the translation status in unfertilized eggs (UnF_*vs*_UnFpuro), in fertilized eggs (F_*vs*_Fpuro) and the recruitment into or exit from polysomal fractions induced by fertilization (F_*vs*_UnF). Normalization and differential analysis were carried out using the generalized linear model framework according to the EdgeR model and package (Robinson et al., 2010). For each gene g, we assumed that the observed count y_{ijg} from polysomal fraction of female i in group j follows a negative binomial distribution with a mean

parameter μ_{ijg}. This mean parameter depends on the sequencing depth for sample corresponding to female i in condition j s_{ij}, on the number of pooled polysome gradients nb_{ij}, on the relative abundance of enriched polysomal mRNA g π_{ijg} and on the observed counts in the associated cytoplasmic mRNA y_{ijg}^{c} normalized by the corresponding sequencing depth s_{ij}^{c} and the number of gradients nb_{ij}^{c}.

$$E(y_{ijg}) = \mu_{ijg} = s_{ij} nb_{ij} \pi_{ijg} \frac{y_{ijg}^{c}}{s_{ij}^{c} nb_{ij}^{c}}$$

The model is the following one :

$$\log(\mu_{ijg}) = \kappa_{ijg} + (F)_i + (G)_j + \varepsilon_{ijg}$$

where κ_{ijg}= log (s_{ij}) + log(nb_{ij}) + log (y_{ijg}^{c})- log (s_{ij}^{c}) − log (nb_{ij}^{c})

$(G)_j$ is the main effect of Group j (j= UnF, UnFpuro, F, Fpuro) , $(F)_i$ is the main effect of female i (i=1,2,3) and ε_{ijg} is a random error term that is assumed to be independent between observations.

The scaling factors s_{ij} and s_{ij}^{c} were calculated using the Trimmed Mean of M-values (TMM) (Robinson and Oshlack, 2010). We used the EdgeR Bioconductor package with a matrix of offsets $[\kappa_{ijg}]$ to fit a negative binomial model per gene with a genewise dispersion as calculated in (Chen et al., 2014; McCarthy et al., 2012). The normalization as pointed out in (Dillies et al., 2013) has an important impact on all the downstream analyzes. The matrix of offsets was assumed to account for all normalization issues (here sequencing depth, numbers of polysome gradients and cytoplasmic values). We computed GLM likelihood ratio tests for UnF *vs* UnFpuro differences, F *vs* Fpuro differences and F *vs* UnF differences within female, to select respectively for mRNAs translated in unfertilized eggs, in fertilized eggs and for mRNAs which translation is modified by fertilization.

Raw p-values were adjusted for multiple comparisons by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) which controls the false discovery rate. Genes with an adjusted p-value lower than 0.05 were considered significant. All statistical analyses were performed using the R software (R Development Core Team, 2011) with Bioconductor (Gentleman et al., 2004) packages.

RT-PCR analysis

PCR primer pairs were designed from the component sequences of the *de novo* maternal

transcriptome using web interface Primer3 (v. 0.4.0), and synthesized by Eurogentec (sequences available in **Table. S2**). The primer pairs were chosen within the ORF and amplicons were between 70 and 250 bp long. All primer pairs were tested for amplification efficiency by standard curves using 5 points of 2-fold range dilution of reverse transcription reaction using cytoplasmic RNA. Only primers with amplification efficiency above 95% were retained for further analyses.

The relative amounts of each mRNAs in polysomal pool (fractions 18-21) and in cytoplasmic samples in unfertilized eggs and one-hour embryos were determined by quantitative RT-PCR (qRT-PCR), on the same 3 biological replicates as the ones used for the RNA-Seq analysis, and were run in 3 experimental replicates. qRT-PCR reactions were performed in 5µl containing 2.1µl of cDNA (1:300); 0.4µl of a 10µM primers solution containing forward and reverse primers; and 2.5µl of SYBR Green I Master Mix (Roche). Thermal cycling parameters were 95°C for 15min and 55 cycles (95°C for 10sec, 60°C for 30sec) followed by a denaturation step to verify the amplification of a single product. Data were analyzed with the LightCycler480 software. The metallothionein 1 (MT1) transcript was chosen as reference. The polysomal recruitment ratio $(polysomal_F/cytoplasmic_F)/(polysomal_{UnF}/cytoplasmic_{UnF})$ was calculated for each mRNA as the average of three biological replicates, and expressed as a log_2FC for comparison with RNA-Seq data.

mRNA distribution along the polysome gradient was analyzed by RT-PCR using an equal volume of RNAs from each fraction, as described in (Chassé et al. 2016). Briefly, reverse transcription (RT) was performed using the reverse transcriptase SuperScript II (Invitrogen) following the manufacturer's instructions. Semi-quantitative RT-PCR was done using a cDNA dilution (1:300) within the linear range of amplification for 30 cycles, with GoTaq Flexi kit (Promega) and 5µM primers. PCR were carried out as follows: 95°C for 2min; 30 cycles (95°C for 30s, 60°C for 30s, 72°C for 1min); 72°C for 5min. PCR products were analyzed on 2% agarose/TBE gels labeled with SybrSafe DNA stain (Invitrogen) and scanned on a Typhoon Trio (GE Healthcare Life Sciences). Quantification was done using the public domain ImageJ program (written by Wayne Rasband at the US National Institutes of Health).

Results

Fertilization induces polysomal recruitment of cyclin B and ribonucleotide reductase small subunit mRNAs, but not of eIF4A mRNA

In the sea urchin *Paracentrotus lividus*, fertilization induces an increase in protein synthesis, as measured by incorporation of labeled methionine into protein, and the first cell division occurs as early as 70 minutes (**Fig. 1A**). Neosynthesized proteins increased shortly after fertilization, but the profile observed by SDS-PAGE did not vary dramatically (**Fig. 1B**), as noted by others in different sea urchin species (Brandhorst, 1976; Winkler et al., 1985). We addressed the question of the identification of these neo-synthesized proteins by using polysome profiling, which allows the separation of an mRNA according to its translational status. Preliminary experiments were conducted to select polysomal fractions for subsequent RNA sequencing and translatome analysis. Polysome profile between unfertilized eggs and 1-hour post-fertilization eggs showed little difference, and no significant changes in heavy fractions of the gradient were detected (Chassé et al., in revision). We tested the polysomal recruitment of maternal mRNAs translated after fertilization, such as cyclin B and the small subunit of ribonucleotide reductase (R2) mRNAs (Kelso-Winemiller et al., 1993; Standart et al., 1985). In unfertilized eggs, both mRNAs were associated to top fractions (1-7) of the polysome gradient. Fertilization induced a strong association of cyclin B and R2 mRNAs with heavy polysomes fractions (18-21). In contrast, eIF4A mRNA repartition on polysome gradient was not strikingly modified after fertilization (**Fig. 1C**). An important bias of the analysis could be that untranslated mRNAs in large RNA-protein complexes (mRNPs) might co-migrate with polysomal mRNAs on a sucrose gradient (Kelso-Winemiller and Winkler, 1991). Control gradients were done in the presence of puromycin, a polysome disrupter to distinguish between translated mRNAs and co-migrating mRNPs. The antibiotic puromycin is incorporated into and disrupt only elongating polysomes (Alexandraki and Ruderman, 1985; Blobel and Sabatini, 1971). As shown in **Fig. 1C**, the pool of the cyclin B and R2 mRNAs associated to polysomal fractions in fertilized embryos was shifted towards lighter fractions upon puromycin treatment, indicating that these mRNAs were efficiently translated after fertilization. We therefore focused on the heavy polysome fractions of the sucrose gradient in order to identify new maternal mRNAs strongly recruited into polysomes after fertilization, with a recruitment behavior similar to cyclin B and ribonucleotide reductase small subunit mRNAs.

Identification of translated mRNAs following fertilization

To assess globally the translational changes and identify the mRNAs translated following fertilization (i.e, the translatome), heavy polysome-associated mRNAs were prepared and sequenced from unfertilized and 1-hour fertilized embryos. RNAs remaining in heavy polysomes fractions after *in vivo* puromycin treatment were sequenced in parallel, to account for mRNPs-associated mRNAs. The full set of mRNAs present in the eggs or embryos was also sequenced from cytoplasmic lysates (**Suppl. Table. S1)**.

As good genomic data was not currently available for *P. lividus*, RNA-Seq was generated from unfertilized and fertilized cytoplasmic RNAs samples and was *de novo* assembled using the Trinity suite to generate a maternal reference transcriptome. A good correlation was obtained between the three independent biological replicates (Pearson correlation R^2 above 0.80; **Supplementary Fig. S1**). The assembly generated 164256 transcripts, corresponding to 76110 genes. Only transcripts with a FPKM>5 were considered for the subsequent analysis, reducing the dataset to 43823 transcripts. Differential comparison between unfertilized eggs and 1-hour post fertilization embryos showed no significant differences (**Suppl. Fig. S2**), suggesting that there was no variation in the abundance of the transcripts between these two stages. An estimation of the repertoire of transcripts present in the maternal transcriptome was done in comparison with data available in the American sea urchin, for which a genome is available (Cameron et al., 2009; Sodergren et al., 2006). An annotation against the 23000 gene predictions of the *S. purpuratus* genome assigned a best blastn hit with an e-value$<10^{-5}$ to 14002 transcripts, corresponding to 7685 Trinity's unique genes and to 6518 *S. purpuratus* gene prediction. The maternal transcriptome thus represented a rough estimate of 28% of total gene number, which is in the range of other species (20-45% in mouse, 55% in drosophila; Horner and Wolfner, 2008).

For the translatome analysis, the reads generated from the 24 samples (cytoplasmic and polysomal samples; see **Suppl. Table. S1**) were mapped independently against the filtered maternal transcriptome comprising 14002 transcripts. The translatome pipeline is presented on **Fig. 2A**. In order quantify each transcript in equivalent number of eggs or embryos, the counts obtained in polysome samples were corrected by the number of pooled gradients used in each condition. A generalized linear model derived from EdgeR, on three independent biological replicates, was fitted to analyze the translation efficiency of each individual transcript, taking into account the cytoplasmic abundance of the transcript. A female effect was added in the model in order to consider the paired design. Pair-wise

comparison of the translation efficiencies of each individual transcript in unfertilized and fertilized samples allowed to select transcripts which translation is modified by fertilization. Transcripts with significant differences (BH adjusted p-value <0.05) between unfertilized eggs and fertilized embryos were retained. Data were then filtered for translated mRNAs in both stages: the translation efficiencies in polysomes were compared with samples treated by puromycin, to distinguish between translated mRNAs and co-migrating mRNPs. Transcripts with a positive fold change and significant difference (BH adjusted p-value <0.05) between treated and untreated polysomes were considered translated. The polysome data obtained from fertilized samples showed a good correlation between the three independent replicates (R^2 above 0.60; **Suppl. Fig. S1**), and no $\log_2$FC(F/Fpuro) threshold was set. In contrast, we observed a lower correlation (R^2 above 0.40) between the three biological replicates of the unfertilized polysomes samples, the translation activity being lower in unfertilized eggs (as shown fig 1A). Therefore a threshold for the unfertilized data was set at $\log_2$FC(UnF/UnFpuro)>1, to insure that the relevant translated transcripts were selected. Applying these filters on the translatome, we found that 2514 transcripts (18% of the maternal set) were significantly recruited into polysomes at fertilization (**Fig. 2B,** "recruitment") from which a majority was translated de novo (**suppl. Fig. 3**).

A subset of 15 genes (translated and untranslated, high and low abundance) was analyzed by qRT-PCR (see primers **Table. S2**), using the polysomal and cytoplasmic RNAs used for sequencing. Quantification for each mRNA was done relative to reference mRNA encoding MT1 (metallothionein 1). Polysomal recruitment was then calculated as the ratio $(polys_F/cyto_F)/(polys_{UnF}/cyto_{UnF})$. The fold-change recruitment obtained by qRT-PCR of pooled polysomal fractions *versus* cytoplasmic was consistent with the recruitment fold-change found by RNA-Seq analysis (**Fig. 2C**).

The maternal mRNAs were further classified according to their polysomal change and their translational efficiencies in unfertilized eggs and fertilized embryos (**Suppl. Fig. 3**). Interestingly, we noticed that a proportion of transcripts (1851/14002; 13%) showed no statistically significant differences between unfertilized and fertilized polysomes but were efficiently translated after fertilization ($\log_2$FC (F/Fpuro)>0, p<0.05); these transcripts probably corresponded to mRNAs stored in heavy mRNPs or stalled polysomes in unfertilized eggs, and that were activated for translation after fertilization (**Fig. 2D**, "masked recruitment"). In contrast, transcripts showing a significant difference between fertilized and

unfertilized polysomes without being actively translated at any stage corresponded to the ones entering into or leaving from heavy mRNPs at fertilization. 52% of the maternal transcripts (7319/14002) were never translated, whereas 7% (967/14002) were translated both before and after fertilization, but showed no change in their polysomal repartition (**Fig. 2D,** "no change"). Finally, 1351 transcripts (10%) were found significantly released of polysomes at fertilization. These data strongly suggest that corrections with puromycin-treated data are determinant to fully identify actively translated mRNAs.

Distribution of newly identified translated mRNAs on polysomes gradients

To further analyze the proportion of each mRNA that enters polysomes after fertilization, we monitored the repartition on sucrose gradients of the new mRNAs that were identified in the translatome analysis. We compared the repartition along the gradient in unfertilized eggs, 1 hour after fertilization, and in puromycin-treated fertilized embryos. All conditions for each biological replicate were obtained from sibling eggs and embryos of a single female and male pair and were processed in parallel. The repartitions for 13 mRNAs identified as recruited into polysomes at fertilization are shown in at least five biological replicates **(Fig. 4)**. We chose to test mRNAs ranging a wide spectrum of abundance and fold-change (see **Table 1**). The polysome profile showed that in unfertilized eggs, the mRNAs were mainly present in the light fractions (3-7) of the gradient. Fertilization triggered a shift of the mRNAs towards the heavy fractions (18-21) of the gradient, displaced in puromycin-treated embryos, which demonstrated the recruitment of these mRNAs into polysome fractions. As negative controls, two mRNAs encoding respectively MT1 and ribosomal protein rps3, identified as untranslated following fertilization (see **Table S3**) were also tested. The polysome profile showed no differences between fertilized and puromycin-treated fertilized samples, indicating that the MT1 and rps3 mRNAs were not translated. These data showed that the translatome analysis by RNA-Seq as described above is therefore validated to uncover newly translated mRNAs.

Entry into polysomes is not correlated to transcript abundance in the maternal transcriptome

It was first suggested that fertilization triggering the activation of the translation machinery would lead to translation of all mRNAs according to the bulk amount in the maternal stock (Brandhorst, 1976). Our translatome analysis showed that only a fraction of

the maternal mRNAs entered polysomes at fertilization. We then asked whether there was a bias towards abundant mRNAs. Comparing mRNAs abundance with their corresponding recruitment index (log_2FC(F/UnF)) showed no correlation between these two parameters, and the mean FPKM value of recruited mRNAs was similar to the value obtained for the maternal mRNAs. We further analyzed abundant maternal transcripts with a FPKM>500 from the RNAseq data and checked for their translational status. Among the 38 most abundant transcripts, only 9 of them were significantly recruited into polysomes after fertilization (**Table S4**). Interestingly, the abundant cleavage histone mRNAs detected in the maternal transcriptome were not translated in our analysis. This is in agreement with previous data (Wells et al., 1981) showing that histone H3 mRNA moves into polysomes only after first cleavage division. These data suggest that the abundance of an mRNA is not correlated with its translational status at fertilization.

Functional enrichment of translated mRNAs at fertilization

As expected, the mRNAs encoding cyclin B and the small subunit of ribonucleotide reductase were identified by our RNA-Seq analysis of polysomal mRNAs in the set of 2514 transcripts significantly recruited at fertilization, We focused our analysis on this set in order to identify new biological actors strongly regulated at the oocyte-to-embryo transition. Cyclin B mRNAs was ranked first when translated mRNAs were classified from lowest to highest p-values, indicating a strong biological constraint on the translation of this mRNAs. **Table 1** presents a short list of translated mRNAs at fertilization (the complete data set can be found in supplementary excel file). Strikingly, many other mRNAs encoding for proteins in cell cycle regulation were found. To estimate the enriched functions in the translatome, GOterms were retrieved from the Trinotate annotation of the maternal transcriptome, and functional classes in the translated mRNAs set were compared to the maternal set. We used a binomial test to compare the observed number of translated genes with the expected number for each functional class, assuming random representation of the maternal mRNAs in the translated set. We found that maternal transcripts implicated in mRNA processing, metabolic processes, cell cycle and signal transduction were the most over-represented biological processes; for molecular processes, RNA-binding, transporter and kinase activities were over represented (binomial test, pvalue<0,05; **Fig. 4**). We further classified maternal mRNAs according to their translational status at fertilization (recruited, translated from mRNPs stocks, release from polysomes and unchanged; supplementary **Fig. S4**). For cell cycle mRNAs and RNA-BP functional categories, half of the maternal mRNAs were in the translated categories. A

striking feature of the mRNAs from the signaling functional class was their abundance in the translated pools from mRNPs stores. The transcripts coding for the ribosomal proteins were significantly under-represented in the translated set of mRNAs, and were shown to be mainly untranslated.

Newly translated mRNAs are differentially sensitive to the mTOR pathway

mTOR signaling pathway is implicated in the protein synthesis increase after fertilization in sea urchin. We showed previously that PP242, an ATP-competitor mTOR inhibitor (Apsel et al., 2008; Benjamin et al., 2011) delays progression through cell cycle and that cyclin B translation was partially dependent on mTOR pathway (Chassé et al. 2016). Since we identified newly recruited RNAs at fertilization (**Fig. 4**), we asked whether the mTOR pathway impacted on the translation of these mRNAs. Eggs were treated with PP242 mTOR inhibitor 10 minutes before fertilization, incubation of sea urchin embryos inhibited 4E-BP degradation and protein synthesis increase triggered by fertilization (**Fig 5**). We then compared mRNA recruitment into polysomes in PP242-treated and untreated control embryos one hour post-fertilization, in 5 independent experiments. The mRNAs which translation we tested could be grouped into three categories regarding the inhibition of their translation in presence of PP242. The first one comprises mRNAs which fertilization-induced translation is completely inhibited by PP242 (Cyclin A, ribonucleotide reductase small subunit R2, CDC6, and RNA-binding protein Musashi). The polysome profile in presence of PP242 matched the one obtained after puromycin treatment, showing no residual translation in presence of the mTOR inhibitor (**Fig. 5, left panel**). In the second group, a fraction of mRNAs remained present in the polysome fractions after PP242 treatment, and addition of puromycin dissociated the remaining polysomes, demonstrating that the fraction of the mRNAs present in polysomes were still translated despite protein synthesis inhibition (**Fig. 5, right panel**). We confirmed that translation of Cyclin B mRNA was partially dependent on mTOR pathway (as described in (Chassé et al. 2016)). In addition the mRNA encoding its kinase partner CDK1 also exhibited the same translation pattern, as well as eIF4B and RNA-binding protein RBM4 (Fig. 5, right panel). Interestingly, a third category of mRNAs was revealed by this analysis: we showed the existence of mRNAs which translation completely escaped the inhibition of cap-dependent translation by PP242 mTOR inhibition: DAP5, CUGBP, RKHD, cdt1, geminin, SoxB1 and gustavus (**Fig. 6**). In all cases, puromycin-treated embryos showed the same distribution profile on polysome gradients (**Suppl. Fig. 5**).

Discussion

Protein synthesis changes at fertilization relies on maternal mRNA use only

No difference in the transcript repertoire was found between eggs and embryos following fertilization, showing that no new transcripts were produced during this short time frame. Although histone mRNAs transcription (Brandhorst, 1980) were shown to occur rapidly at fertilization, these mRNAs were not selected and therefore do not appear in our study because RNA-Seq libraries were made after poly(A) selection. The observed protein synthesis increase at fertilization therefore relies only on the translation of stored maternal mRNAs.

Translation of a subset of maternal mRNAs at fertilization

Fertilization triggers a large increase in protein synthesis due to the activation of cap-dependent translation machinery. To date, only quantitative changes have been reported, because most proteins detected after fertilization were already present in the egg (Brandhorst, 1976). Nonetheless, some reports described new proteins synthesized soon after fertilization, among which are the cyclins driving the cell cycle (Evans et al., 1983; Kelso-Winemiller and Winkler, 1991). To re-examine the question of the qualitative translational changes occurring at fertilization, we coupled polysome sucrose gradient technique to isolate translated mRNAs with RNA-sequencing technology, allowing a large-scale translatome analysis. In this study we have analyzed the set of mRNAs that were recruited one hour post-fertilization, in the 1-cell embryo before the occurrence of first cell division. We therefore focused on the oocyte-to-embryo transition, a key developmental stage going from a differentiated oocyte into a totipotent embryo (Horner and Wolfner, 2008). We also focused our analysis on the mRNAs that entered the heavy polysomal fractions, in other words, that were very efficiently translated. For the first time in sea urchin, we showed that a limited subset of the maternal stored mRNAs was efficiently translated, whereas the majority was not translated at this early time-point after fertilization. Interestingly, some mRNAs (2%) strongly recruited at fertilization were also translated before fertilization, albeit at a lower rate, suggesting the existence of a cap-independent mechanism of translation before fertilization. We also found mRNAs that were translated only before fertilization, and mRNAs for which ongoing translation did not change at fertilization, they were beyond the scope of this study but would need further study. The contribution of maternal mRNAs unmasking from mRNPs has been

underlined in this study by the fact that the translation of some mRNAs was only revealed by comparison to puromycin-treated embryos. 13% of the maternal mRNAs were found to display such compartment changes. Our study thus demonstrated the selective recruitment of a subset of stored maternal mRNAs at fertilization, in an invertebrate deuterostome, the sea urchin

Translated maternal mRNAs at fertilization are involved in several biological processes

The translated maternal mRNAs at fertilization were enriched in several biological processes, involved in sea urchin early development. It is noteworthy that functional classes were also enriched in sea urchin fertilization, as shown in mouse oocyte maturation and fertilization (Chen et al., 2011), or in drosophila egg activation (Kronja et al., 2014); but functional classes may be different according to whether the egg has to complete meiosis or the timing of zygotic re-activation.

Cell cycle regulators

The mRNAs encoding cell cycle regulators were enriched in the maternal set of translated mRNAs. It is well established that accumulation of some key components of the cell cycle, such as the cyclins, are regulated at the level of translation oocytes and early embryos, as well as in somatic cells (Tarn and Lai, 2011). In addition to the already described cyclin A and B mRNAs, we surprisingly found the mRNAs encoding their partner CDK1 in the maternal set of translated mRNAs. The CDK1 protein is present as an abundant maternal protein and its level is not modified by protein synthesis inhibition (Meijer et al., 1991). The amount of neosynthesized CDK1 resulting from the mRNA entry into polysome may be negligible compared to the maternal amount, as already observed for other proteins in different organisms (Kronja et al., 2014), but the small amount of unphosphorylated neosynthesized CDK1 associated to Cyclin B could have a role in the auto-amplification loop of the complex. Furthermore, we found that a large number of cell cycle regulators are translationally activated, giving an additional layer of complexity in the regulation of cell cycle progression. Three mRNAs encoding proteins of the pre-replication complex (cdt1, cdc6 and geminin) were equally recruited in polysomes 1 hour after fertilization. Geminin protein accumulated at 30 minutes post-fertilization, whereas Cdt1 and CDC6 did not in sea urchin (Aze et al., 2010). This apparent discrepancy may be explained by the small proportion of the neosynthesized proteins compared to large amount of maternal proteins, leading to

apparent unchanged steady state level. It is noteworthy that the three mRNAs were not impacted equally by mTOR inhibition.

Maternal determinants of development patterning

Axis specification and endomesoderm formation rely on determinants which mRNAs are expressed maternally in sea urchin. Among these maternal determinants, SoxB1 mRNA was strongly recruited and translated in one-cell embryo, very early with respect to its role in transcription reactivation at early blastula. This early translation is in agreement with data shown in *S. purpuratus*, where the SpSoxB1 protein was first detected by western blot after the 2-cell stage and increases in abundance during cleavage stages (Kenny et al., 1999). Other mRNAs acting in the same gene regulatory network were not recruited into polysomes at fertilization (for example Otx or Ets1/2), suggesting that SoxB1 may have an additional role in early cleavage stages. Interestingly, work from A. Giraldez's group in zebrafish showed recently that SoxB1 was one of the most strongly and early translated maternal mRNAs and was implicated in maternal mRNAs clearance, by regulating the transcription of microRNA mir-430 (Lee et al., 2013).

During sea urchin germ line development, the E3 ubiquitin ligase specificity receptor Gustavus regulates the accumulation of Vasa protein in small micromeres, and the protein accumulates in the embryo between egg and 4-cells stage (Gustafson et al., 2011). This increase in protein level can be explained by the strong polysomal recruitment of the Gustavus mRNA that we have detected after fertilization in our study.

In sea urchin, the establishment of the oral-aboral axis is highly regulated by the spatially localized zygotic expression of Nodal, which depends on maternal factors implicated in the TGF-β signaling (Haillot et al., 2015; Range and Lepage, 2011). Several members of the TGF-β pathway such as the activin receptor-like kinase receptor ALK2, Vg1/univin and SMAD4 were present among the maternal mRNAs strongly recruited at fertilization. Interestingly, other mRNAs from TGF-β pathway were mostly enriched in the fraction of maternal mRNAs that were in mRNPs or stalled polysomes before fertilization, since there recruitment into polysomes was only revealed by the comparison with puromycin-treated eggs and embryos.

RNA-binding proteins

We observed an enrichment of mRNA encoding RNA-binding proteins in the translated set of maternal mRNAs. RNA-binding protein RBM4 was found in the top five mRNAs recruited at fertilization ($log_2FC>3$; pvalue$<10^{-5}$; **Table 1**). RBM4 is involved in specific translation in hypoxia (Uniacke et al., 2012), and in IRES (*Internal Ribosome Entry Site*)-dependent translation in stress condition in mammalian cells (Lin and Tarn, 2009). The drosophila RBM4 homolog, LARK is required for development (McNeil et al., 1999). Several of the translated RNA-binding proteins are involved in translation repression of specific mRNAs. For example, CUGBP is an RRM-domain containing RNA-binding protein first identified in Xenopus embryo for its ability to bind specifically to a GU-rich element (Embryonic deadenylation element EDEN) located in the 3'UTRs of some mRNAs that are rapidly deadenylated and translationally repressed after fertilization in early development (Paillard et al., 1998). RKHD is a KH- and ring-domain containing RNA-binding protein identified in worm, ascidians and sea urchin genomes, with four paralogs in vertebrates (MEX3A-D), MEX3B being the closest to the invertebrate homologs. MEX3 has been shown to be a translation repressor, involved in embryonic cell fate (reviewed in Pereira et al., 2013). In sea urchin RKHD has been identified as a ubiquitous maternal and localized zygotic mRNA (Röttinger et al., 2006), its function in sea urchin has not been investigated yet. CUGBP and RKHD were strongly recruited into polysomes after fertilization. These two mRNAs were translated in some individuals also before fertilization. Interestingly, they were translated in PP242-treated embryos. In both condition, the translation inhibitor protein 4E-BP was present and inhibited cap-dependent translation, suggesting an alternative translation initiation for CUGBP and RKHD mRNAs. No data is available concerning the maternal presence of the proteins encoded by these mRNAs, but finding that they were neosynthesized after fertilization suggests that control of mRNA fate (stability, localization, translation, *etc*.) could be set-up or modified in the early embryo.

Identification of mTOR independent translation in sea urchin translated mRNAs

In this study we have shown that impairing the mTOR pathway activated at fertilization did not impact similarly all mRNAs. When embryos are treated with PP242 or rapamycin, the global protein synthesis is inhibited but some proteins are still neosynthesized (Chassé et al, 2016, fig. 5). For the first time, we have identified mRNAs which translation is independent of mTOR translation in sea urchin. On the survey of 16 mRNAs we found surprising that a high number of mRNAs (44%) is still translated in PP242-treated embryos. It was shown that 3-5% of mRNAs remained associated to polysomes in somatic cells when

cap-dependent translation was impaired (Johannes et al., 1999). This discrepancy could be due to the nature of the mRNAs tested, which encode for proteins involved in highly regulated processes, or to the embryo status compared to somatic cells. A first clue may be that mRNAs encoding ribosomal proteins were not translated in fertilized sea urchin embryos, whereas activation of the mTOR pathway in mammalian cells increases their translation (Fonseca et al., 2014). A translatome analysis of PP242-treated embryos will help apprehend the full spectrum of residual translation when cap-dependent translation is impaired.

We have shown for the first time in sea urchin that several mRNAs are translated in PP242-treated embryos, potentially in an IRES-dependent mechanism. These mRNAs encode for DAP5, CUGBP, RKHD, Cdt1, Geminin, SoxB1 and Gustavus, and can now be used to uncover the *cis*- and *trans*- regulatory mechanisms responsible for their specific translation regarding the mTOR pathway. Among these mRNAs, DAP5 (an eIF4G homolog unable to bind eIF4E) is translated through an IRES located in its 5'UTR in mammalian cells (Henis-Korenblit et al., 2000), suggesting an evolutionary conservation among deuterostomes.

Sequences databases

RNA-Seq raw data have been deposited in NCBI BioProject database under the accession PRJNA288758.

Additional files

The complete list of maternal and translated mRNAs can be found in the supplementary excel file.

Funding

This work was supported by research grants from the French Cancer League (*La Ligue contre le Cancer*, *comités Finistère, Côtes d'Armor, Deux-Sèvres et Morbihan*), the Brittany Regional Council (*Région Bretagne*), the Finistère Departmental Council (CG29); and by an equipment grant from the ITMO AVIESAN CNRS/INSERM. H. Chassé is a Ph.D. fellow supported by the Brittany Regional Council (*Région Bretagne*).

Acknowledgements

We thank the SMO and M3 services of the Station Biologique of Roscoff for collection and maintenance of sea urchins.

References

Alexandraki, D. and Ruderman, J. V. (1985). Expression of α- and β-tubulin genes during development of sea urchin embryos. *Dev. Biol.* **109**, 436–451.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Apsel, B., Blair, J. A., Gonzalez, B., Nazif, T. M., Feldman, M. E., Aizenstein, B., Hoffman, R., Williams, R. L., Shokat, K. M. and Knight, Z. A. (2008). Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat. Chem. Biol.* **4**, 691–699.

Aze, A., Fayet, C., Lapasset, L. and Geneviève, A. M. (2010). Replication origins are already licensed in G1 arrested unfertilized sea urchin eggs. *Dev. Biol.* **340**, 557–570.

Benjamin, D., Colombi, M., Moroni, C. and Hall, M. N. (2011). Rapamycin passes the torch: a new generation of mTOR inhibitors. *Nat. Rev. Drug Discov.* **10**, 868–880.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300.

Blobel, G. and Sabatini, D. (1971). Dissociation of mammalian polyribosomes into subunits by puromycin. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 390–394.

Brandhorst, B. P. (1976). Two-dimensional gel patterns of protein synthesis before and after fertilization of sea urchin eggs. *Dev. Biol.* **52**, 310–317.

Brandhorst, B. P. (1980). Simultaneous synthesis, translation, and storage of mRNA including histone mRNA in sea urchin eggs. *Dev. Biol.* **79**, 139–148.

Cameron, R. A., Samanta, M., Yuan, A., He, D. and Davidson, E. (2009). SpBase: The sea urchin genome database and web site. *Nucleic Acids Res.* **37**, 750–754.

Chassé, H., Mulner-Lorillon, O., Boulben, S., Glippa, V., Morales, J. and Cormier, P. (2016). Cyclin B translation depends on mTOR activity after fertilization in sea urchin embryos. *PLoS One*, **11**, e0150318.

Chen, J., Melton, C., Suh, N., Oh, J. S., Horner, K., Xie, F., Sette, C., Blelloch, R. and Conti, M. (2011). Genome-wide analysis of translation reveals a critical role for deleted in azoospermia-like (Dazl) at the oocyte-to-zygote transition. *Genes Dev.* **25**, 755–766.

Chen, Y., Lund, A. H. and Smyth, G. K. (2014). Differential expression analysis of complex RNA-Seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequence Data* (ed. Datta, S. and Nettleton, D.), New York: Springer.

Cormier, P., Pyronnet, S., Morales, J., Mulner-Lorillon, O., Sonenberg, N. and Bellé, R. (2001). eIF4E Association with 4E-BP Decreases Rapidly Following Fertilization in Sea Urchin. *Dev. Biol.* **232**, 275–283.

Costache, V., Bilotto, S., Laguerre, L., Bellé, R., Cosson, B., Cormier, P. and Morales, J. (2012). Dephosphorylation of eIF2α is essential for protein synthesis increase and cell cycle progression after sea urchin fertilization. *Dev. Biol.* **365**, 303–309.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683.

Evans, T., Rosenthal, E. T., Youngblom, J., Distel, D. and Hunt, T. (1983). Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* **33**, 389–396.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30.

Fonseca, B. D., Smith, E. M., Yelle, N., Alain, T., Bushell, M. and Pause, A. (2014). The ever-evolving role of mTOR in translation. *Semin. Cell Dev. Biol.* 1–12.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

Gildor, T. and Ben-Tabou de-Leon, S. (2015). Comparative Study of Regulatory Circuits in Two Sea Urchin Species Reveals Tight Control of Timing and High Conservation of Expression Dynamics. *PLOS Genet.* **11**, e1005435.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Gustafson, E. A., Yajima, M., Juliano, C. E. and Wessel, G. M. (2011). Post-translational regulation by gustavus contributes to selective Vasa protein accumulation in multipotent cells during embryogenesis. *Dev Biol* **349**, 440–450.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512.

Haillot, E., Molina, M. D., Lapraz, F. and Lepage, T. (2015). The Maternal Maverick/GDF15-like TGF-β Ligand Panda Directs Dorsal-Ventral Axis Formation by Restricting Nodal Expression in the Sea Urchin Embryo. *PLOS Biol.* **13**, e1002247.

Henis-Korenblit, S., Strumpf, N. L., Goldstaub, D. and Kimchi, A. (2000). A novel form of DAP5 protein accumulates in apoptotic cells as a result of caspase cleavage and internal ribosome entry site-mediated translation. *Mol. Cell. Biol.* **20**, 496–506.

Hershey, J. W. B., Sonenberg, N. and Mathews, M. B. (2012). Principles of Translational Control: An Overview. *Cold Spring Harb. Perspect. Biol.* **4**, a011528–a011528.

Horner, V. L. and Wolfner, M. F. (2008). Transitioning from egg to embryo: Triggers and mechanisms of egg activation. *Dev. Dyn.* **237**, 527–544.

Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. and Sarnow, P. (1999). Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc Natl Acad Sci U S A* **96**, 13118–13123.

Kelso-Winemiller, L. C. and Winkler, M. M. (1991). "Unmasking" of stored maternal mRNAs and the activation of protein synthesis at fertilization in sea urchins. *Development* **111**, 623–633.

Kelso-Winemiller, L., Yoon, J., Peeler, M. T. and Winkler, M. M. (1993). Sea urchin maternal mRNA classes with distinct development regulation. *Dev. Genet.* **14**, 397–406.

Kenny, A. P., Kozlowski, D., Oleksyn, D. W., Angerer, L. M. and Angerer, R. C. (1999). SpSoxB1, a maternally encoded transcription factor asymmetrically distributed among early sea urchin blastomeres. *Development* **126**, 5473–5483.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. . (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Kronja, I., Yuan, B., Eichhorn, S. W., Dzeyk, K., Krijgsveld, J., Bartel, D. P. and Orr-Weaver, T. L. (2014). Widespread Changes in the Posttranscriptional Landscape at the Drosophila Oocyte-to-Embryo Transition. *Cell Rep.* **7**, 1495–1508.

Kuersten, S., Radek, A., Vogel, C. and Penalva, L. O. F. (2013). Translation regulation gets its "omics" moment. *Wiley Interdiscip. Rev. RNA* **4**, 617–630.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Larsson, O., Tian, B. and Sonenberg, N. (2013). Toward a genome-wide landscape of translational control. *Cold Spring Harb. Perspect. Biol.* **5**, a012302.

Le Breton, M., Bellé, R., Cormier, P., Mulner-Lorillon, O. and Morales, J. (2003). M-phase regulation of the recruitment of mRNAs onto polysomes using the CDK1/cyclin B inhibitor aminopurvalanol. *Biochem. Biophys. Res. Commun.* **306**, 880–886.

Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming, E. S. and Giraldez, A. J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature* **503**, 360–364.

Lin, J. C. and Tarn, W. Y. (2009). RNA-binding motif Protein 4 translocates to cytoplasmic granules and suppresses translation via argonaute2 duringmuscle cell differentiation. *J. Biol. Chem.* **284**, 34658–34665.

McCarthy, D. J., Chen, Y. and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297.

McNeil, G. P., Zhang, X., Roberts, M. and Jackson, F. R. (1999). Maternal function of a retroviral-type zinc-finger protein is essential for Drosophila development. *Dev. Genet.* **25**, 387–396.

Meijer, L., Azzi, L. and Wang, J. Y. (1991). Cyclin B targets p34cdc2 for tyrosine phosphorylation. *Embo J* **10**, 1545–1554.

Oulhen, N., Salaün, P., Cosson, B., Cormier, P. and Morales, J. (2007). After fertilization of sea urchin eggs, eIF4G is post-translationally modified and associated with the cap-binding protein eIF4E. *J. Cell Sci.* **120**, 425–434.

Oulhen, N., Mulner-Lorillon, O. and Cormier, P. (2010). eIF4E-binding proteins are differentially modified after ammonia versus intracellular calcium activation of sea urchin unfertilized eggs. *Mol. Reprod. Dev.* **77**, 83–91.

Paillard, L., Omilli, F., Legagneux, V., Bassez, T., Maniey, D. and Osborne, H. B. (1998). EDEN and EDEN-BP, a cis element and an associated factor that mediate sequence-specific mRNA deadenylation in Xenopus embryos. *EMBO J.* **17**, 278–287.

Pereira, B., Le Borgne, M., Chartier, N. T., Billaud, M. and Almeida, R. (2013). MEX-3 proteins: recent insights on novel post-transcriptional regulators. *Trends Biochem. Sci.* **38**, 477–479.

Petersen, T. N., Brunak, S., von Heijne, G. and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786.

R Development Core Team (2011). *R: A language and environment for statistical computing*.

Range, R. and Lepage, T. (2011). Maternal Oct1/2 is required for Nodal and Vg1/Univin expression during dorsal–ventral axis specification in the sea urchin embryo. *Dev. Biol.* **357**, 440–449.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40.

Röttinger, E., Besnardeau, L. and Lepage, T. (2006). Expression pattern of three putative RNA-binding proteins during early development of the sea urchin Paracentrotus lividus. *Gene Expr. Patterns* **6**, 864–872.

Salaun, P., Pyronnet, S., Morales, J., Mulner-Lorillon, O., Bellé, R., Sonenberg, N. and Cormier, P. (2003). eIF4E/4E-BP dissociation and 4E-BP degradation in the first mitotic division of the sea urchin embryo. *Dev. Biol.* **255**, 428–439.

Salaün, P., Pyronnet, S., Morales, J., Mulner-Lorillon, O., Bellé, R., Sonenberg, N. and Cormier, P. (2003). eIF4E/4E-BP dissociation and 4E-BP degradation in the first mitotic division of the sea urchin embryo. *Dev. Biol.* **255**, 428–439.

Schmieder, R., Lim, Y. W. and Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* **28**, 433–435.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* **473**, 337–342.

Sodergren, E., Weinstock, G., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C. and Angerer, L. M. (2006). The genome of the sea urchin Strongylocentrotus purpuratus. *Science.* **314**, 941–952.

Standart, N. M., Bray, S. J., George, E. L., Hunt, T. and Ruderman, J. V (1985). The small subunit of ribonucleotide reductase is encoded by one of the most abundant translationally regulated maternal RNAs in clam and sea urchin eggs. *J Cell Biol* **100**, 1968–1976.

Swartz, S. Z., Reich, A. M., Oulhen, N., Raz, T., Milos, P. M., Campanale, J. P., Hamdoun, A. and Wessel, G. M. (2014). Deadenylase depletion protects inherited mRNAs in primordial germ cells. *Development* **141**, 3134–3142.

Tarn, W.-Y. and Lai, M.-C. (2011). Translational control of cyclins. *Cell Div.* **6**, 5.

Tu, Q., Cameron, R. A., Worley, K. C., Gibbs, R. A. and Davidson, E. H. (2012). Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome Res.* **22**, 2079–2087.

Uniacke, J., Holterman, C. E., Lachance, G., Franovic, A., Jacob, M. D., Fabian, M. R., Payette, J., Holcik, M., Pause, A. and Lee, S. (2012). An oxygen-regulated switch in the protein synthesis machinery. *Nature* **486**, 126–130.

Wells, D. E., Showman, R. M., Klein, W. H. and Raff, R. A. (1981). Delayed recruitment of maternal mRNA in sea urchin embryos. *Nature(London)* **292**, 477–478.

Winkler, M. M., Nelson, E. M., Lashbrook, C. and Hershey, J. W. (1985). Multiple levels of regulation of protein synthesis at fertilization in sea urchin eggs. *Dev. Biol.* **107**, 290–300.

Figures and Tables Legends

Table 1: Identification of recruited and translated mRNAs at fertilization by RNA-Seq analysis of polysomal mRNAs. Each mRNA was identified through a transcript ID corresponding to the Trinity *de novo* assembly of the maternal transcriptome, translation index (log_2FC(F/Fpuro)) and entry into polysomes fractions (log_2FC(F/UnF)) after fertilization are indicated with their respective adjusted pvalues. Abundance expressed as FPKM and best blast hit to *S. purpuratus* gene models (Echinobase; http://echinobase.org; Cameron et al, 2009) are indicated. In bold are shown the mRNAs which have been further verified by polysome gradient repartition in **Fig. 4**.

Figure 1: A- Cell division and [^{35}S]-methionine incorporation into proteins after fertilization. The data is the mean of two independent experiments, error bars represent standard deviation. **B -** Profile of [^{35}S]-methionine labeled neo-synthesized proteins separated on SDS-PAGE gel up to 90 min. post-fertilization. Arrow indicates the position of neo-synthesized cyclin B. **C-** Cyclin B (CycB), ribonucleotide reductase small subunit (R2) and initiation factor 4A (eIF4A) mRNAs repartition along 15-40% sucrose gradient before (UnF) and after (F) fertilization. mRNA was quantified by semi quantitative RT-PCR in each fraction, and represented as a percent of mRNA in all fractions (% of total), errors bars represent SEM on 5 biological replicates (UnF *vs* F: * pval<0.05). Presence of the mRNA into polysomes was assessed by treating embryos *in vivo* with puromycin before polysome gradient fractionation (F+puro *in vivo*, n=3).

Figure 2 : **A-** Outline of the translatome analysis, performed on three independent polysome profiling dataset. The colored section corresponds to the set of mRNAs actively translated and recruited into polysomes at fertilization (2514 transcripts). **B-** Log_2FC between F and UnF polysomal contents were plotted against FPKM for each mRNAs, maternal mRNAs are in grey, recruited mRNAs at fertilization are in red. **C**- Comparison of log_2FC between unfertilized and fertilized polysomal mRNA for 13 genes obtained by RNA-Seq analysis and by qRT-PCR. The line is a linear regression line. **D-** Pie chart repartition of maternal mRNAs according to their polysomal behavior at fertilization as determined by the log_2FC (F*vs*UnF), corrected by the puromycin control at both time-points.

Figure 3: Functional categories and GOterms associated to maternal mRNAs. The numbers of transcripts obtained in each category were compared to expected numbers assuming

random representation of maternal mRNAs in the translated set in a binomial test (p-value<0,05). Barcharts represent the enriched biological (top) and molecular (bottom) processes associated with translated mRNAs.

Figure 4: Translation analysis of selected mRNAs by polysome purification on sucrose gradient. mRNAs were detected by RT-PCR amplification in each fraction of the polysome gradient from unfertilized eggs (UnF), 1-hour post-fertilization embryos (F) or puromycin-treated embryos (F+puro). Amplicons were run on agarose gels, quantified using ImageJ software, repartition is shown along the gradient as a percentage of total mRNA. Fraction #1 corresponds to the top of the gradient (free mRNAs) and #21 corresponds to the bottom of the gradient. Values are shown as a mean of at least 5 independent biological replicates, error bars represent SEM (UnF *vs* F: * pval<0.05).

Figure 5: mTOR inhibition by PP242 impacts protein synthesis activity and 4E-BP degradation. **A-** top: Neosynthesized proteins labeled with [^{35}S]-Methionine separated on SDS-PAGE gel and autoradiography, in unfertilized eggs, 1 hour post-fertilization embryos, and fertilized embryos in presence of PP242 inhibitor. Bottom: Western blot analysis of 4E and 4E-BP in the above sample. 4E-BP degradation triggered by fertilization is inhibited by PP242 (3 independent experiments). **B-** Protein synthesis activity measured by incorporation of [^{35}S]-Methionine in TCA precipitated proteins, normalized to the values in control fertilized embryos (representative of two independent experiments).

Figure 6:Translation analysis by polysome purification on sucrose gradient in presence of PP242 inhibitor. mRNAs were detected as described in Fig. 4 in each fraction of the polysome gradient from 1-hour post-fertilization embryos (F), in presence of PP242 inhibitor (F+PP242) or in presence of PP242 and puromycin (F+PP242+puro). Left panel shows mRNAs exhibiting mTOR dependent translation. Right panel shows mRNAs exhibiting mTOR partially dependent translation. Values are shown as a mean of at least 5 independent biological replicates, error bars represent SEM (F *vs* F+PP242: * pval<0.05 and F+PP242 *vs* F+PP242+puro: † pval<0.05).

Figure 7: mRNAs exhibit mTOR independent translation, as described in **Figure. 5**. Values are shown as a mean of 5 independent biological replicates, error bars represent SEM (F+PP242 *vs* F+PP242+puro: † pval<0.05).

Table I: Translated mRNAs at fertilization

Transcript ID	gene	F_vs_UnF		F_vs_Fpuro		FPKM	SPU Best Blast Hit
		logFC	padj	logFC	padj		
Cell cycle related genes							
comp77341_c2_seq1	**Cyclin B**	2.964	2.14E-19	1.631	2.33E-07	3202.66	SPU_015285
comp79240_c1_seq1	**Cyclin A**	2.339	5.65E-13	1.513	1.81E-06	649.47	SPU_003528
comp80046_c1_seq2	**Ribonucleotide reductase small subunit R2**	2.645	1.67E-12	1.195	7.15E-04	4679.83	SPU_024933
comp79495_c0_seq1	**Cyclin dependent kinase 1 CDK1**	1.583	8.41E-07	1.933	8.15E-10	62.15	SPU_002210
comp79997_c3_seq5	**Geminin**	0.970	1.73E-03	1.211	5.63E-05	150.56	SPU_005762
comp79170_c0_seq2	**Cell division control protein 6 cdc6**	0.944	5.07E-03	1.482	6.35E-06	154.35	SPU_010595
comp78987_c0_seq3	**DNA-replication factor cdt1**	1.393	5.09E-07	1.192	1.30E-05	72.55	SPU_002046
comp79482_c0_seq1	Cyclin B3	1.071	3.16E-04	1.400	1.25E-06	190.16	SPU_006444
comp76547_c0_seq2	regulator of chromosome condensation 1 RCC1	2.426	9.07E-17	1.687	2.51E-09	103.06	SPU_023992
comp69502_c0_seq1	14-3-3 epsilon	2.223	1.96E-10	0.864	1.17E-02	64.30	SPU_003825
comp77014_c0_seq1	Polo-like kinase 1 PLK1	1.342	1.22E-06	1.033	1.51E-04	77.47	SPU_017949
comp78419_c0_seq1	Early mitotic inhibitor EMI1	1.417	2.56E-05	0.953	4.28E-03	14.66	SPU_008889
Maternal determinant of development patterning							
comp77686_c1_seq1	**SoxB1**	1.713	3.74E-06	1.700	1.87E-06	1320.41	SPU_022820
comp77921_c0_seq1	**Gustavus**	2.582	2.03E-13	1.533	5.46E-06	75.98	SPU_004717
comp79094_c0_seq1	Alk2	2.474	1.01E-12	1.859	2.63E-08	17.77	SPU_016008
comp73990_c1_seq1	beta-catenine	1.490	1.49E-04	1.172	2.00E-03	194.50	SPU_009155
comp79473_c0_seq3	transforming growth factor beta receptor	1.131	1.64E-03	1.284	3.16E-04	6.38	SPU_027380
comp76027_c0_seq8	Transforming growth factor beta Univin	2.311	1.16E-11	1.799	7.47E-08	18.28	SPU_000668
comp79158_c0_seq1	Smad4	1.447	2.21E-06	1.401	3.58E-06	14.88	SPU_004287
RNA-binding proteins							
comp70206_c1_seq4	**RBM4**	3.427	3.15E-16	1.597	3.84E-05	100.80	SPU_022878
comp66342_c1_seq4	**DAZAP Musashi**	1.757	6.76E-08	1.294	3.29E-05	85.10	SPU_024306
comp79981_c0_seq9	**CUG-BP**	1.401	3.26E-07	1.189	1.24E-05	7.74	SPU_015850
comp76987_c1_seq1	**RKHD/Pem-3/Mex3B homolog**	1.072	4.34E-04	1.164	9.54E-05	854.27	SPU_003290
comp78371_c0_seq1	Nova	1.103	2.74E-04	1.592	8.14E-08	10.04	SPU_003114
comp68546_c0_seq1	hnRNP K	1.849	2.21E-09	1.666	3.54E-08	90.53	SPU_008011
comp62631_c0_seq1	hnRNP A	1.191	2.06E-04	1.825	6.83E-09	115.70	SPU_015676
comp76265_c0_seq4	Histone RNA hairpin-binding protein SLBP	2.121	3.72E-08	1.246	8.67E-04	62.95	SPU_009593
Translation regulation							
comp73250_c0_seq1	**eIF4E binding protein Neuroguidin**	1.862	3.19E-04	1.366	6.29E-03	4.79	SPU_019210
comp79103_c1_seq1	**DAP5**	1.339	4.26E-04	1.398	1.68E-04	87.12	SPU_023932
comp78411_c2_seq1	**eIF4B**	1.111	6.02E-04	0.958	2.71E-03	94.72	SPU_004840
comp78490_c0_seq1	termination factor eRF1	1.376	1.55E-05	1.847	1.61E-09	11.40	SPU_023948
comp64074_c0_seq1	initiation factor eIF6	1.267	2.71E-02	1.554	4.25E-03	7.94	SPU_012909
mRNA processing							
comp78669_c0_seq3	pre-mRNA splicing factor SF3a	2.227	1.60E-14	1.828	5.33E-11	32.04	SPU_007675
comp77748_c1_seq1	splicing factor p54	1.723	5.97E-07	1.692	4.21E-07	17.86	SPU_023789
comp80399_c5_seq2	Splicing factor. arginine/serine-rich 6	1.529	1.41E-06	1.550	5.32E-07	81.34	SPU_002681
comp77662_c0_seq8	THO complex 4	2.473	2.88E-13	1.447	8.34E-06	16.07	SPU_010685
comp80045_c0_seq2	THO complex 5	1.203	7.08E-04	1.562	7.83E-06	4.62	SPU_024067
Metabolism. transport							
comp78749_c0_seq1	Glutamine synthase	1.632	5.11E-07	1.221	1.22E-04	1066.31	SPU_023123
comp76283_c0_seq1	Vacuolar ATP synthase subunit B	2.431	1.65E-11	1.588	5.15E-06	35.06	SPU_016414
comp76633_c0_seq1	solute carrier family 11	2.085	4.14E-08	1.162	1.90E-03	6.17	SPU_023546
comp64211_c0_seq1	GDP-fucose transporter 1	1.783	1.07E-07	2.086	1.95E-10	28.70	SPU_006903
comp74935_c0_seq4	Spermidine synthase	2.147	4.64E-07	0.908	2.92E-02	6.46	SPU_018922
Proteases							
comp75368_c0_seq2	Nocturnin. CCR4 protein homolog	2.455	3.88E-07	2.293	1.23E-06	2.75	SPU_020125
comp72397_c0_seq1	ubiquitin-conjugating enzyme E2S	1.445	1.48E-04	0.821	2.05E-02	63.68	SPU_005257

Fig. 1

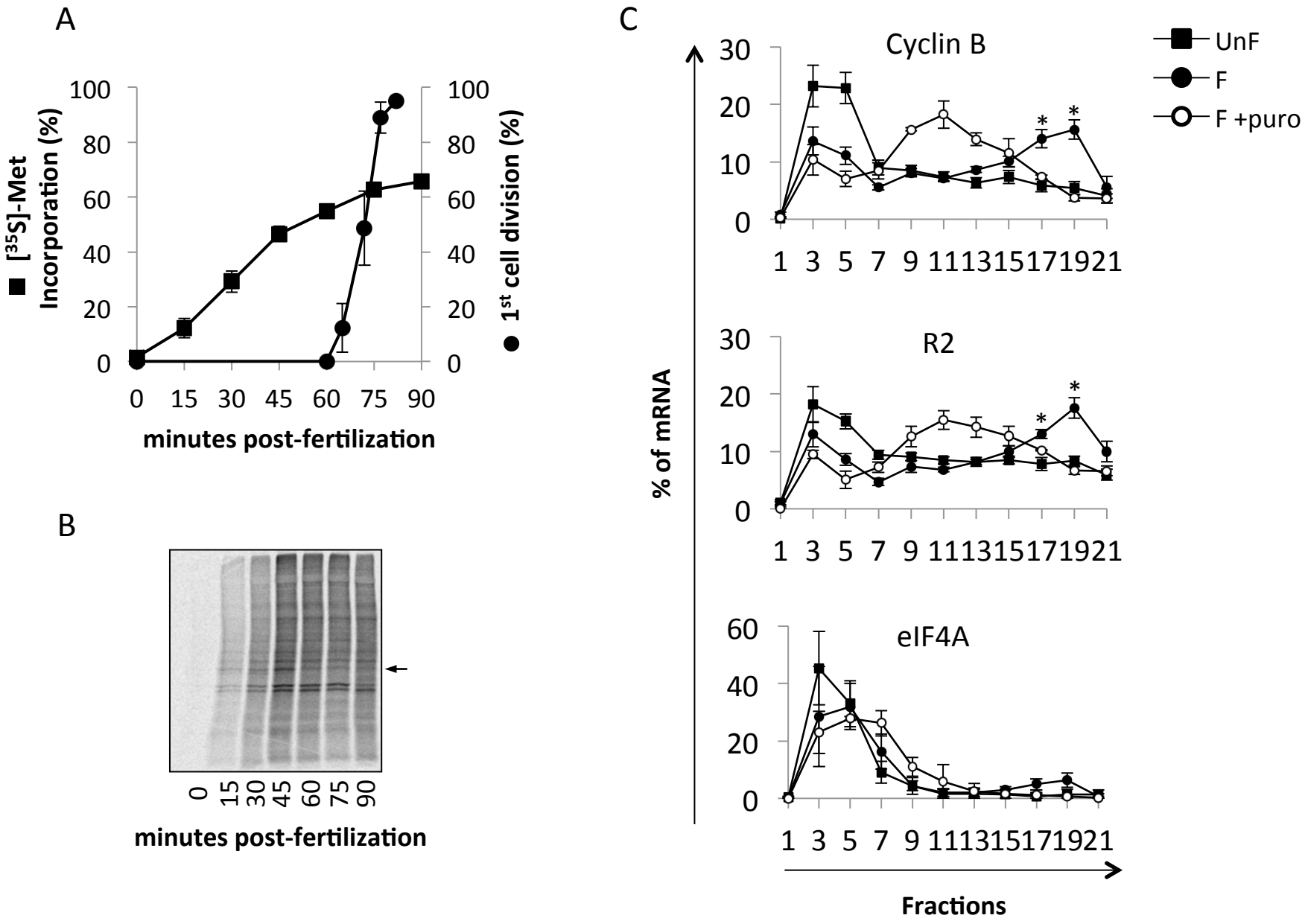
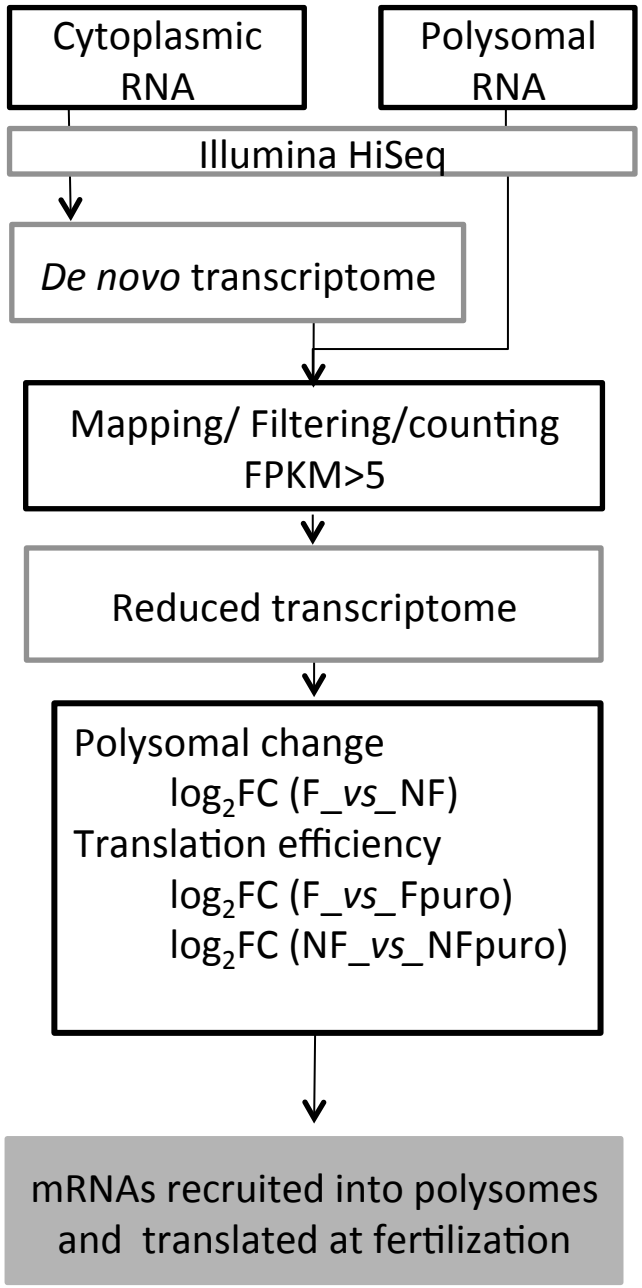
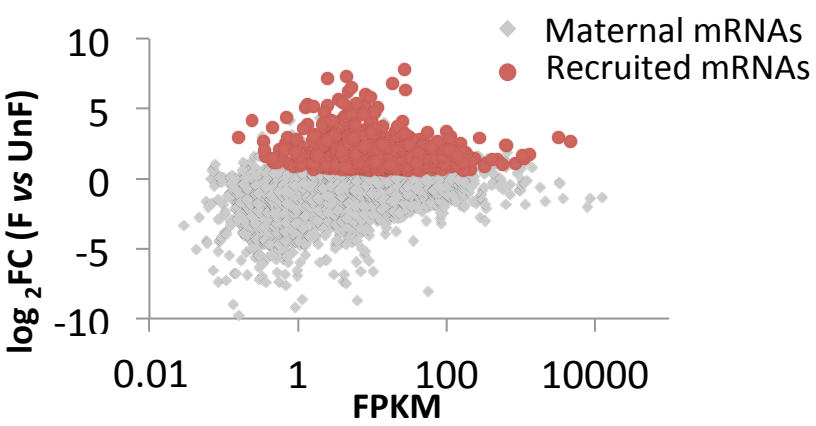

A
■ [35S]-Met Incorporation (%)
● 1st cell division (%)
minutes post-fertilization
B
minutes post-fertilization
C
Cyclin B
R2
eIF4A
UnF
F
F +puro
% of mRNA
Fractions

Fig. 2

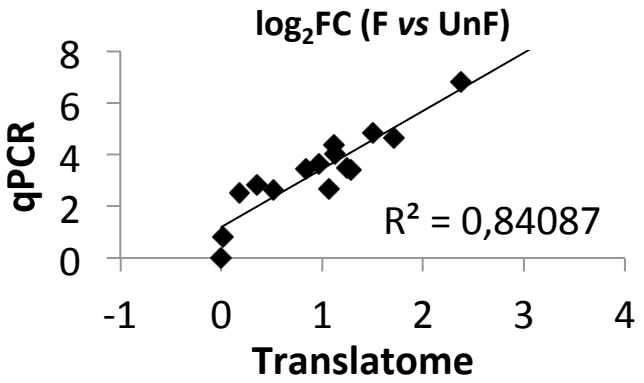
A


Cytoplasmic RNA
Polysomal RNA
Illumina HiSeq
De novo transcriptome
Mapping/ Filtering/counting FPKM>5
Reduced transcriptome
Polysomal change
log2FC (F_vs_NF)
Translation efficiency
log2FC (F_vs_Fpuro)
log2FC (NF_vs_NFpuro)
mRNAs recruited into polysomes and translated at fertilization

B


Maternal mRNAs
Recruited mRNAs
log2FC (F vs UnF)
10
5
0
-5
-10
0.01
1
100
10000
FPKM

C


log2FC (F vs UnF)
qPCR
8
6
4
2
0
R² = 0,84087
-1
0
1
2
3
4
Translatome

D

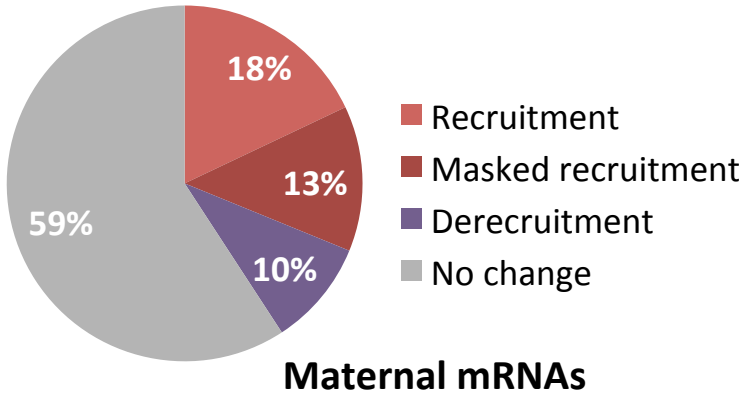

18%
13%
10%
59%
Recruitment
Masked recruitment
Derecruitment
No change
Maternal mRNAs

Fig. 3

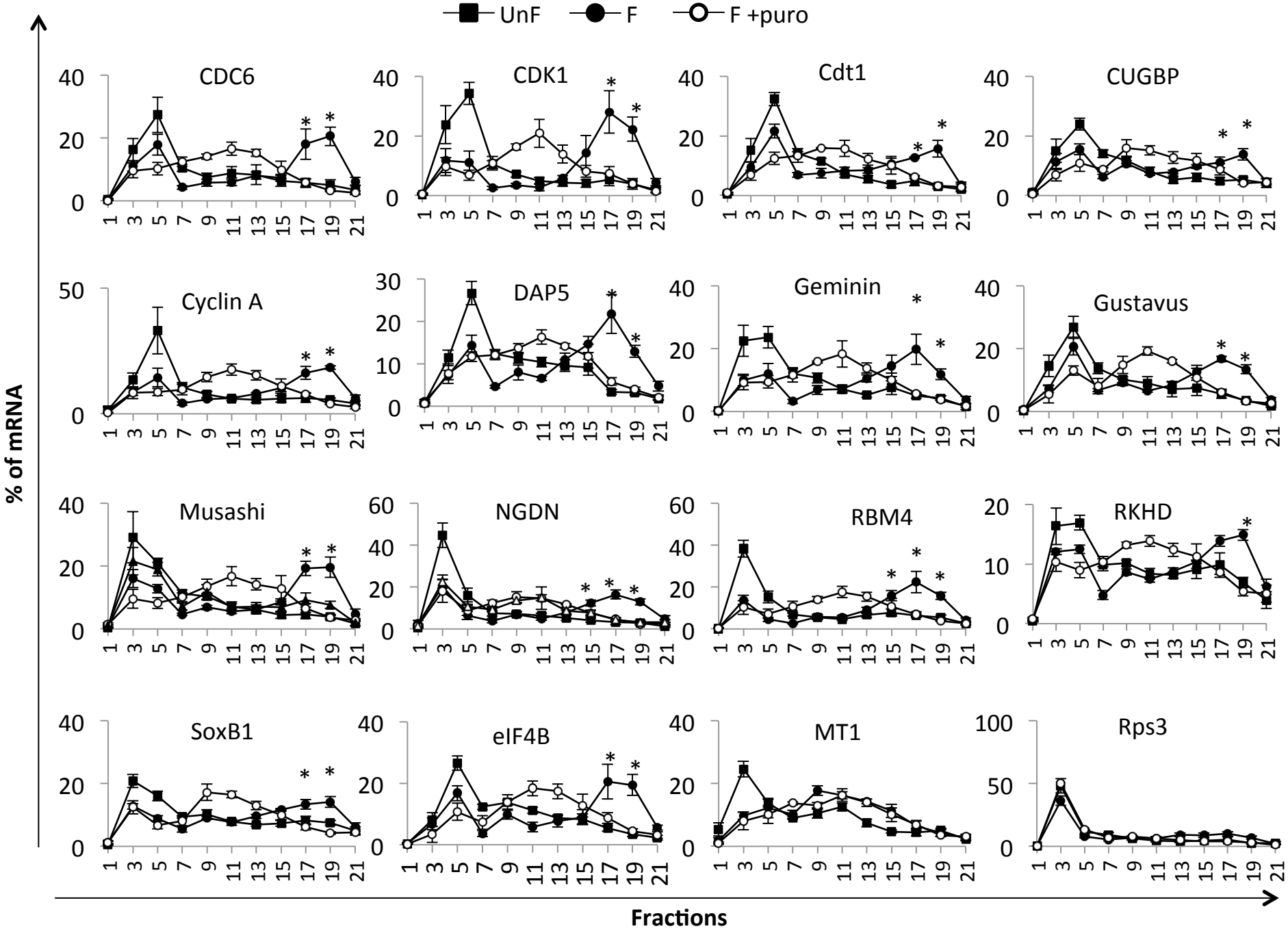

UnF
F
F +puro
CDC6
CDK1
Cdt1
CUGBP
Cyclin A
DAP5
Geminin
Gustavus
Musashi
NGDN
RBM4
RKHD
SoxB1
eIF4B
MT1
Rps3
% of mRNA
Fractions

Fig. 4

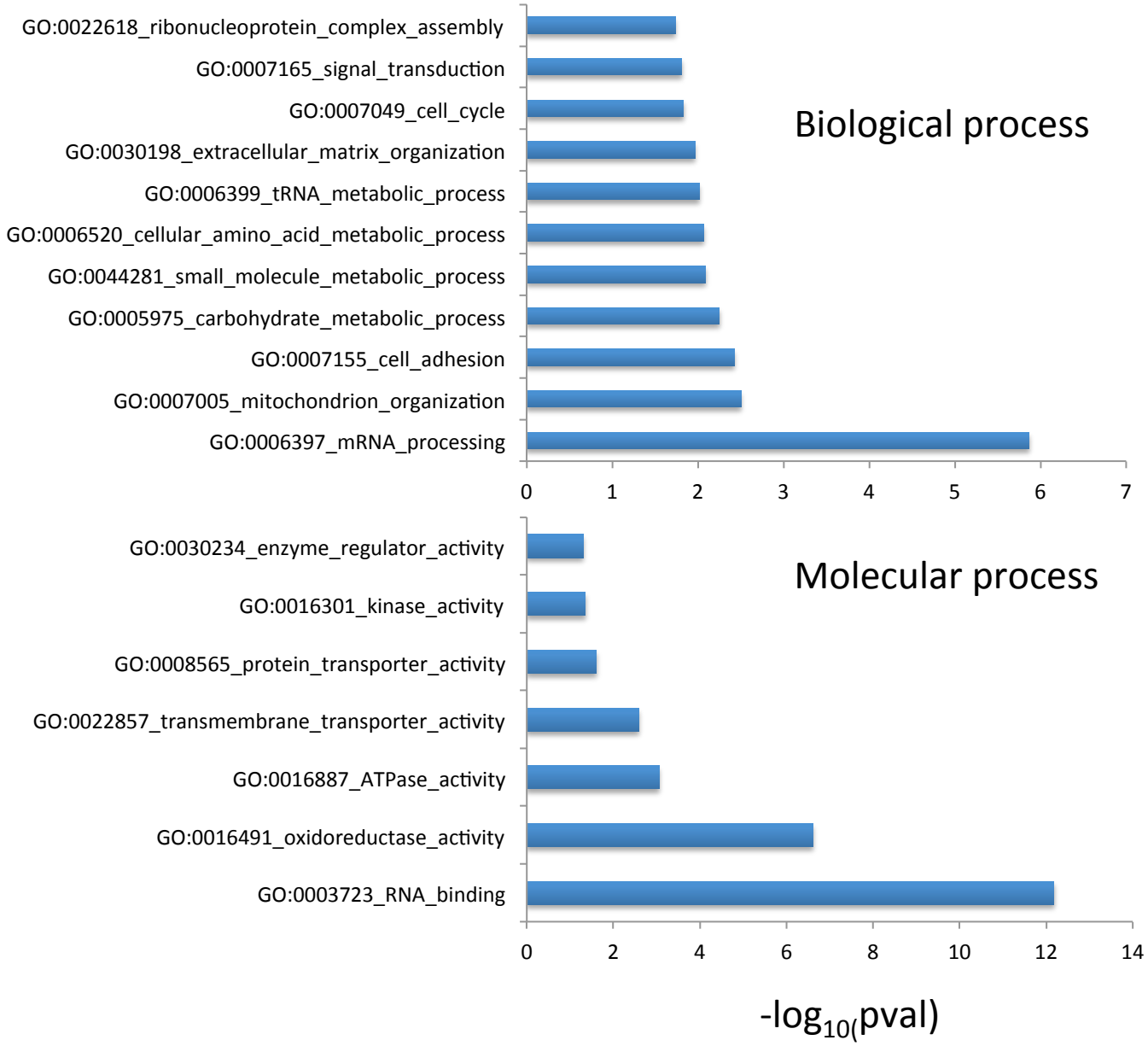

GO:0022618_ribonucleoprotein_complex_assembly
GO:0007165_signal_transduction
GO:0007049_cell_cycle
GO:0030198_extracellular_matrix_organization
GO:0006399_tRNA_metabolic_process
GO:0006520_cellular_amino_acid_metabolic_process
GO:0044281_small_molecule_metabolic_process
GO:0005975_carbohydrate_metabolic_process
GO:0007155_cell_adhesion
GO:0007005_mitochondrion_organization
GO:0006397_mRNA_processing
0
1
2
3
4
5
6
7
Biological process
GO:0030234_enzyme_regulator_activity
GO:0016301_kinase_activity
GO:0008565_protein_transporter_activity
GO:0022857_transmembrane_transporter_activity
GO:0016887_ATPase_activity
GO:0016491_oxidoreductase_activity
GO:0003723_RNA_binding
0
2
4
6
8
10
12
14
Molecular process
-log10(pval)

Fig. 5

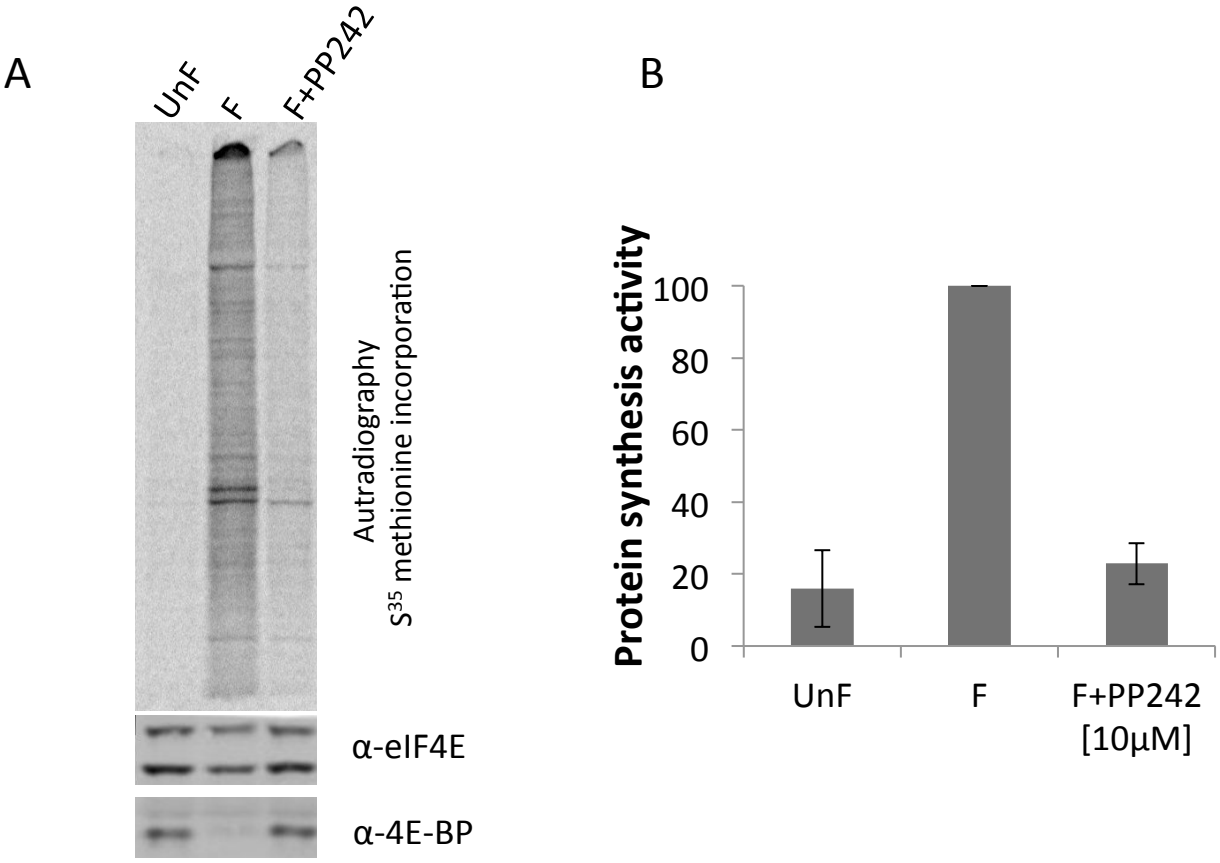

A
UnF
F
F+PP242
Autradiography
S35 methionine incorporation
α-eIF4E
α-4E-BP
B
Protein synthesis activity
100
80
60
40
20
0
UnF
F
F+PP242
[10µM]

Fig. 6

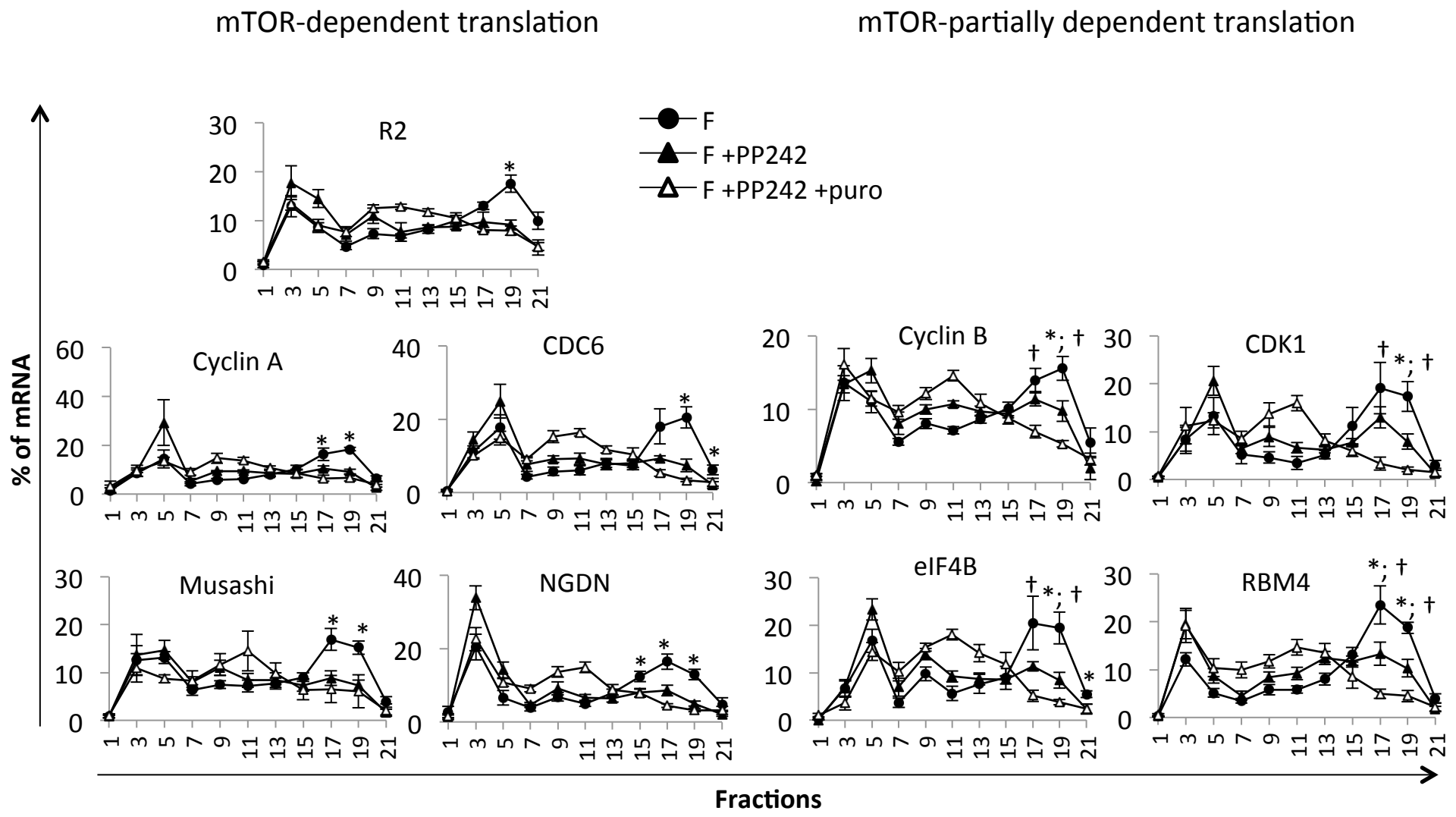
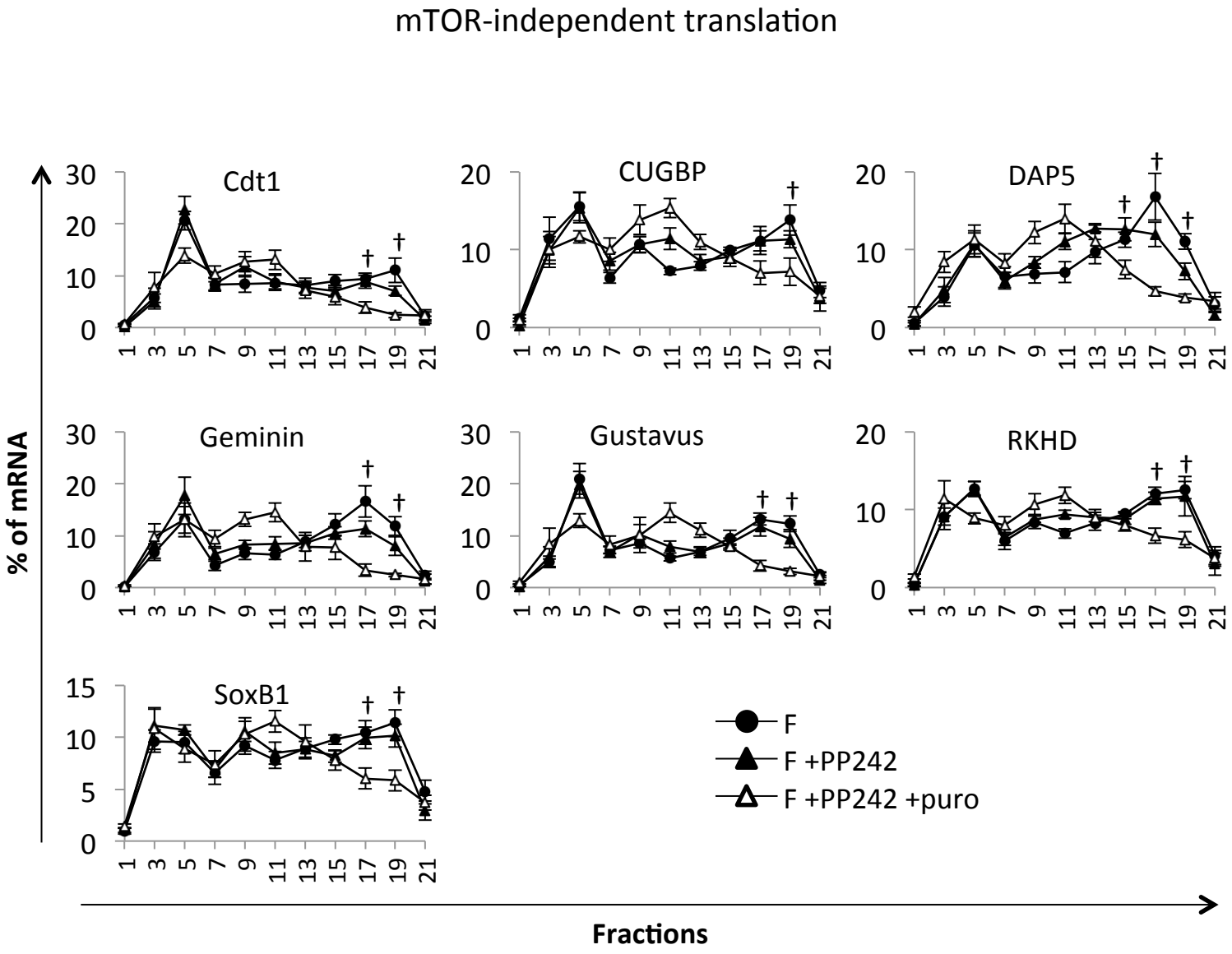

mTOR-dependent translation
mTOR-partially dependent translation
F
F +PP242
F +PP242 +puro
% of mRNA
Fractions
R2
Cyclin A
CDC6
Cyclin B
CDK1
Musashi
NGDN
eIF4B
RBM4

Fig. 7


mTOR-independent translation
Cdt1
CUGBP
DAP5
Geminin
Gustavus
RKHD
SoxB1
% of mRNA
Fractions
F
F +PP242
F +PP242 +puro

Supplementary figures and tables legends

Table S1: Illumina generated data on the twenty-four libraries of the study. 3 independent experiments were done, with four different conditions: before and 1 hour post-fertilization without (UnF; F) or with puromycin treatment (UnF+puro; F+puro). For each condition, polysomal RNA and cytoplasmic RNA were purified. Polysomal fractions were pooled from 3-9 gradients to obtain enough polysomal RNA for libraries construction (in parenthesis, the number of pooled gradients per condition). Cytoplasmic RNA was purified from equal volume of cytoplasmic lysates from each condition, leading to equivalent amount of RNA. Numbers for each library are indicated for the input (paired end 100bp reads), cleaned and mapped reads.

Table S2: Primers used in this study for RT-PCR analyses.

Table S3: Untranslated mRNAs at fertilization by RNA-Seq analysis of polysomal mRNAs, used as negative controls for mRNA distribution on polysome gradients. See table 1 for details

Table S4: List of the most abundant maternal mRNAs (FPKM>500), with their respective translation and recruitment index (as explained in Table. 1). Shadowed lines indicate recruited mRNAs at fertilization. Bold mRNAs indicate mRNAs that were also tested by polysome gradient.

Figure S1: Pearson correlation coefficients between the three independent biological replicates. The red lines show the expected correlations of 1 and the blue lines show the linear models, which fit with the data.

Figure S2: EdgeR differential analysis (MA plot and volcano plot representations) of the P. lividus maternal transcriptome before and after fertilization. Similar result was obtained using DESeq analysis (not shown). No significant difference (FDR <0.05) was detected.

Figure S3: Repartition of maternal mRNAs according to their translation and recruitment behavior as determined by RNA-Seq data analysis.

Figure S4: Repartition of mRNAs encoding for cell cycle, RNA-binding proteins (RBP), signaling and ribosomal proteins according to their recruitment behavior.

Figure S5: Puromycin treatment of control embryos and PP242-treated embryos led to similar polysome profile. mRNAs were detected by RT-PCR amplification as described in **Fig. 4**

Table S1: Summary of Illumina generated data

Sample type		Independent samples	Input read pairs	cleaned read	mapped reads
Polysomal fractions	UnF	A1 (8)	34 030 616	32 902 935	70.28%
		A2 (7)	33 704 324	32 418 813	74.28%
		A3 (9)	35 499 748	34 408 317	68.94%
	F	B1 (4)	33 285 471	32 136 980	66.00%
		B2 (4)	37 618 772	36 343 907	66.21%
		B3 (3)	37 194 989	36 171 153	63.96%
	UnF+puro	C1 (4)	33 883 652	32 823 787	64.22%
		C2 (4)	27 565 328	26 661 893	66.89%
		C3 (8)	29 567 387	28 502 813	69.13%
	F+puro	D1 (4)	28 867 787	28 026 934	64.15%
		D2 (4)	34 750 950	33 560 209	65.57%
		D3 (4)	33 714 918	32 715 942	62.77%
Cytoplasmic RNA	UnF	E1	31 155 992	30 228 804	72.11%
		E2	35 472 877	34 347 325	71.77%
		E3	30 658 576	29 710 431	66.81%
	F	F1	33 950 363	32 767 073	67.14%
		F2	26 870 053	26 064 389	67.95%
		F3	30 809 143	29 899 105	61.48%
	UnF+puro	G1	37 256 715	36 071 092	69.99%
		G2	34 317 375	33 229 198	70.21%
		G3	35 483 092	34 374 527	63.15%
	F+puro	H1	37 936 744	36 783 064	60.17%
		H2	33 916 426	32 741 477	60.08%
		H3	36 282 397	35 039 346	57.84%

Table S2: primers used in this study

Gene	Symbol	Primer sequence	Comp#	Size bp
Ribonucleotide reductase small subunit	R2	5'-TTCGCTGCCAGTGATGGA-3' 5'-TCGGCAACCTGGACTTCCT-3'	comp80046_c1_seq2	70
Cyclin B	CycB	5'-CAAAGAGCATGGCTGTTCAA-3' 5'-CCATTGTATCCATCGCCTCT-3'	comp77341_c2_seq1	234
Cyclin A	CycA	5'- CCAACCATGGCCCACTATAC-3' 5'- ACCCCATCTCCCATCCTTAC-3'	comp79240_c1_seq1	243
Cyclin-dependent kinase 1	CDK1	5'-CTTTCCAAAGTGGACGAACC -3' 5'-AAGGACGGGAACGAAAGACT -3'	comp79495_c0_seq1	198
Transcription factor SoxB1	SoxB1	5'-TTAGCCATTTGTGCAGCTTG -3' 5'-ACCAACACCTGAACGGCTAC-3'	comp77686_c1_seq1	108
Geminin	Gem	5'- TCAAAGACACGTCCATCAGC-3' 5'-CAACGACCTTGAGCTTGTCA -3'	comp79997_c3_seq5	189
Cell division control protein 6	CDC6	5'-GCAGGAGATGTCAGGAAAGC -3' 5'-TGGAGGAGGGAAACCTTCTT -3'	comp79170_c0_seq2	199
DNA replication factor Cdt1	Cdt1	5'-TGGAACTGGTTGCCAAGAAT -3' 5'- CTCCTCAGGGGTTCTCAACA-3'	comp78987_c0_seq3	199
CUG binding protein, CELF2	CUGBP	5'- CCTCGCTCAGAGTCAAGCATT-3' 5'-CCTGTTGAGCCCGGTGTAAC -3'	comp79981_c0_seq9	74
Ring and KH-domain protein, Mex3B-like	RKHD	5'-GCTACCCGAGCTGATGCTAC -3' 5'-CAACCGTACGAAGGTCCACT -3'	comp76987_c1_seq1	146
RNA binding protein 4	RBM4	5'-TCAGAGGTGGAAGAGGAGGA -3' 5'-CCTTGCTCTGTAGGGGTCAC -3'	comp70206_c1_seq4	196
Euc. Initiation Factor 4B	eIF4B	5'- GGAGGAGCAAAGCCTGTAGA-3' 5'- ACGCGTTCTGCTTTCTCTTC-3'	comp78411_c2_seq1	200
Gustavus, Spry	Gus	5'- CGTGAACTCGCATACAGTGG-3' 5'-ACATGCAGTCCTCTCGTGAA -3'	comp77921_c0_seq1	229
neuroguidin	NGDN	5'- AATCAATGGAGACCCTGCTG-3' 5'-GCATCTCCTTCCTCGTCTTG -3'	comp73250_c0_seq1	206
RNA-binding protein Musashi homolog / DAZAP	Mus	5'- AGCCACACCTGATGATCTCC-3' 5'- TGGGCATGGCTTTCTTAATC-3'	comp66342_c1_seq4	199
DAP5, euc. Initiation Factor 4 gamma 2	DAP5	5'-AGACGAGCAGGACCAGAGAG -3' 5'-GTCGGCCTACAGTGGTGATT -3'	comp79103_c1_seq1	208
Euc. Initiation Factor 4A	eIF4A	5'-TGGTCAAGAAGGAAGAAC-3' 5'-CGTCTCATACAAGTCACA -3'	comp73316_c0_seq1	103
Methallothionein	MT1	5'-ATGTTGCCAAGATGGAAAGC -3' 5'-GGGTCCGTGCTAACGTCTAA -3'	comp75920_c1_seq1	196
Ribosomal protein 3 small subunit	RPS3	5'-GGACGCATGATCTTCACCTT -3' 5'-GGTGGTGGTCTCTGGTAAGC -3'	comp61766_c0_seq1	168

Table S3: Untranslated mRNAs at fertilization, used as negative control in fig.4

Transcript ID	gene	F_vs_UnF		F_vs_Fpuro		FPKM	SPU Best Blast Hit
		logFC	padj	logFC	padj		
Untranslated mRNAS							
comp73316_c0_seq1	**translation initiation factor eIF4A**	0.342	4.90E-01	1.206	3.44E-03	59.53	SPU_023083
comp61766_c0_seq1	**ribosomal protein rps3**	-2.317	1.55E-02	-1.398	1.04E-01	8.27	SPU_013662
comp75920_c1_seq1	**metallothionein MT1**	-1.656	2.86E-05	-3.614	2.49E-19	2152.71	SPU_017989

Table S4: List of the most abundants maternal mRNAs (FPKM>500)

		F_vs_UnF		F_vs_Fpuro			
Transcript ID	gene	logFC	padj	logFC	padj	FPKM	SPU Best Blast Hit
comp80046_c1_seq2	**Ribonucleotide reductase small chain**	2,645	1,67E-12	1,195	7,15E-04	4679,83	**SPU_024933**
comp77341_c2_seq1	**CycB cyclin B**	2,964	2,14E-19	1,631	2,33E-07	3202,66	**SPU_015285**
comp77686_c1_seq1	**SoxB1**	1,713	3,74E-06	1,700	1,87E-06	1320,41	**SPU_022820**
comp78749_c0_seq2	***Glutamine synthetase***	1,421	9,61E-06	1,142	2,28E-04	1128,45	***SPU_023123***
comp78749_c0_seq1	***Glutamine synthetase***	1,632	5,11E-07	1,221	1,22E-04	1066,31	***SPU_023123***
comp76987_c1_seq1	**RKHD/Mex3B/Pem3**	1,072	4,34E-04	1,164	9,54E-05	854,27	**SPU_003290**
comp79240_c1_seq1	**CycA cyclin A**	2,339	5,65E-13	1,513	1,81E-06	649,47	**SPU_003528**
comp79240_c1_seq2	**CycA cyclin A**	2,381	1,67E-12	1,649	5,30E-07	619,43	**SPU_003528**
comp70565_c3_seq1	***hnRNP-like***	1,012	2,30E-02	1,316	1,48E-03	565,84	***SPU_000477***
comp73815_c0_seq1	Cleavage Histone H1	0,710	5,52E-02	-0,768	3,05E-02	1026,35	SPU_024567
comp73815_c0_seq2	Cleavage Histone H1	0,375	3,37E-01	-0,858	1,60E-02	912,56	SPU_024567
comp75920_c1_seq3	MT1 metallothionein	-1,364	1,12E-03	-3,716	6,39E-19	12773,51	SPU_017989
comp75270_c0_seq1	cytochrome b	-1,384	1,06E-02	-2,110	8,25E-05	8698,29	SPU_016012
comp75920_c1_seq1	MT1 metallothionein	-1,656	2,86E-05	-3,614	2,49E-19	2152,71	SPU_017989
comp75920_c1_seq4	MT1 metallothionein	-1,813	1,42E-04	-4,218	1,00E-17	1144,19	SPU_017989
comp78526_c0_seq5	Histone H3.3	-0,831	2,96E-02	-1,735	2,21E-06	1004,26	SPU_015647
comp61924_c0_seq1	Histone H3.3	-0,822	1,68E-02	-1,906	1,19E-08	978,07	SPU_028062
comp78526_c0_seq3	Histone H3.3	-1,175	6,74E-04	-1,742	8,31E-08	720,65	SPU_015647
comp60951_c0_seq1	Selenoprotein W	-1,855	1,61E-05	-3,378	1,20E-14	635,00	SPU_026490
comp56639_c0_seq1	Histone H4	-0,793	4,47E-02	-2,659	7,83E-12	586,77	SPU_018670
comp69555_c0_seq2	Cleavage Histone H2b	-0,344	4,53E-01	-1,988	1,27E-06	3404,22	SPU_001312
comp69555_c0_seq1	Cleavage Histone H2b	-0,280	4,83E-01	-2,308	9,45E-11	3279,07	SPU_001312
comp66449_c1_seq1	Histone H3.3	-0,651	8,73E-02	-1,779	3,54E-07	1873,98	SPU_008231
comp69718_c0_seq2	Cleavage Histone H2a	-0,349	4,14E-01	-1,843	1,29E-06	1131,21	SPU_022075
comp69718_c0_seq3	Cleavage Histone H2a	-0,373	3,87E-01	-1,742	5,52E-06	722,51	SPU_022075
comp69718_c0_seq1	Cleavage Histone H2a	-0,491	2,27E-01	-1,844	1,02E-06	625,60	SPU_022075
comp69718_c0_seq4	Cleavage Histone H2a	-0,112	8,10E-01	-1,797	1,67E-06	545,33	SPU_022075
comp67174_c0_seq1	translation initiation factor eIF1	-0,518	1,54E-01	-3,208	4,72E-20	530,38	SPU_016208
comp78178_c2_seq1	HMG-Box prot	1,148	7,16E-04	-0,452	1,88E-01	1108,79	SPU_027981
comp77245_c0_seq1	receptor accessory protein REEP5	1,870	4,61E-08	-0,204	5,83E-01	666,91	SPU_018377
comp70206_c0_seq2	Actin CyIIB	1,288	2,07E-03	0,608	1,43E-01	632,17	SPU_009483
comp67883_c0_seq1	dual oxidase maturation factor DuoxA	0,813	4,51E-02	0,497	2,11E-01	628,98	SPU_025513
comp70206_c0_seq1	Actin CyIIB	1,037	1,36E-02	0,607	1,44E-01	573,94	SPU_009483
comp76059_c1_seq4	DEAD box polypeptide Ddx5	0,027	9,51E-01	-0,645	5,84E-02	534,48	SPU_010343
comp59090_c0_seq1	Ubiquitin	-1,389	6,99E-05	-0,057	8,84E-01	3765,37	SPU_021496
comp59002_c0_seq1	Tubulin alpha5	-0,178	6,91E-01	0,089	8,35E-01	1208,05	SPU_026444
comp75920_c2_seq1	Tubulin	-0,138	7,91E-01	-0,563	2,05E-01	1056,82	SPU_012679
comp75920_c2_seq3	Tubulin	-0,484	2,15E-01	-0,123	7,57E-01	1044,92	SPU_012679

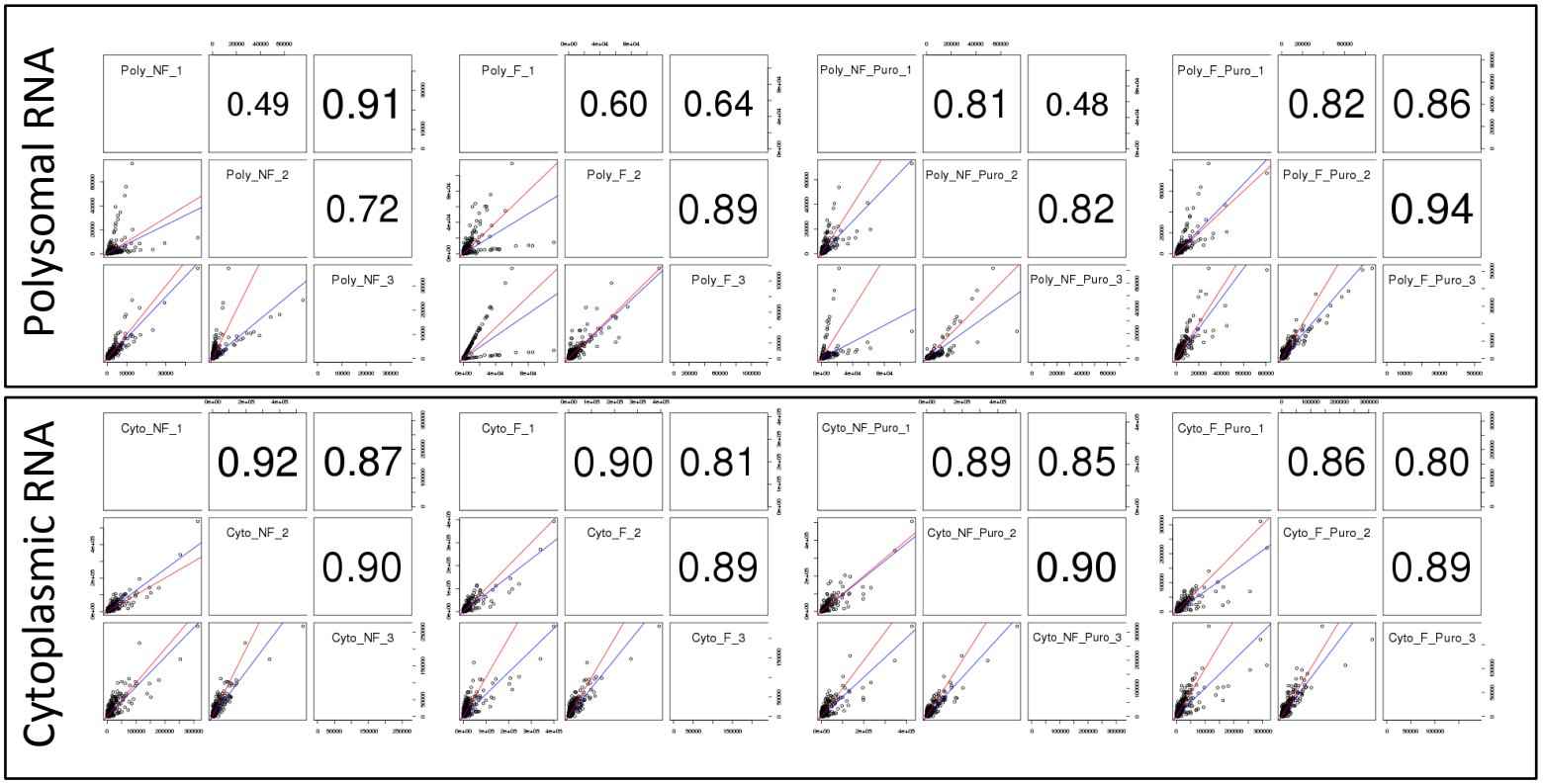
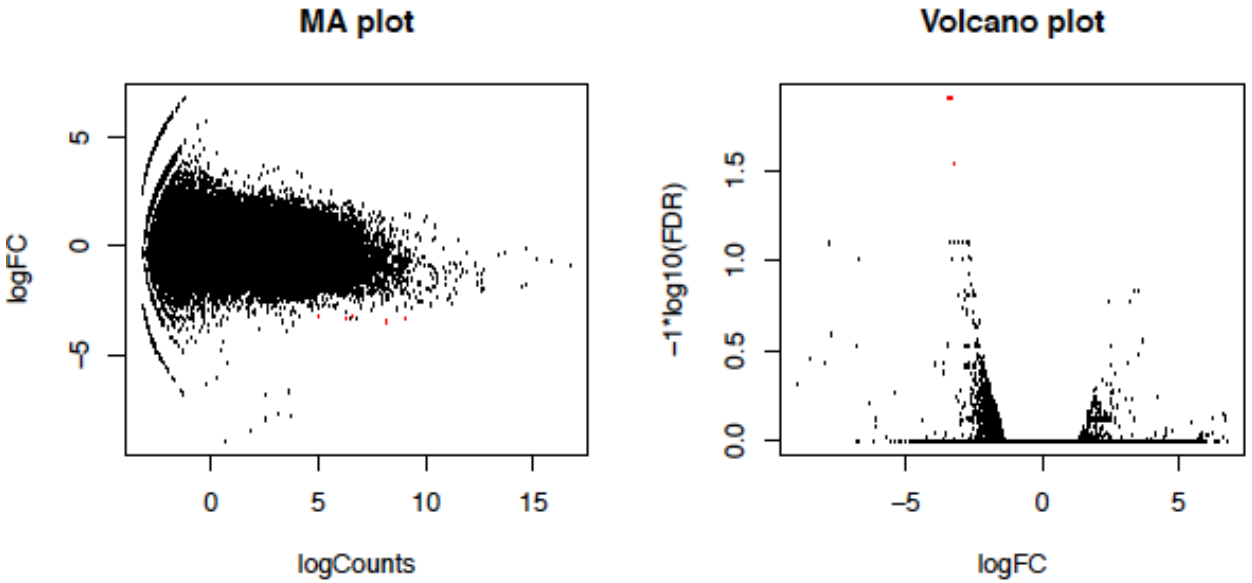

Polysomal RNA
Poly_NF_1
Poly_NF_2
Poly_NF_3
0.49
0.91
0.72
Poly_F_1
Poly_F_2
Poly_F_3
0.60
0.64
0.89
Poly_NF_Puro_1
Poly_NF_Puro_2
Poly_NF_Puro_3
0.81
0.48
0.82
Poly_F_Puro_1
Poly_F_Puro_2
Poly_F_Puro_3
0.82
0.86
0.94
Cytoplasmic RNA
Cyto_NF_1
Cyto_NF_2
Cyto_NF_3
0.92
0.87
0.90
Cyto_F_1
Cyto_F_2
Cyto_F_3
0.90
0.81
0.89
Cyto_NF_Puro_1
Cyto_NF_Puro_2
Cyto_NF_Puro_3
0.89
0.85
0.90
Cyto_F_Puro_1
Cyto_F_Puro_2
Cyto_F_Puro_3
0.86
0.80
0.89

Fig S1


MA plot
logFC
5
0
−5
0
5
10
15
logCounts
Volcano plot
−1*log10(FDR)
0.0
0.5
1.0
1.5
−5
0
5
logFC

FigS2

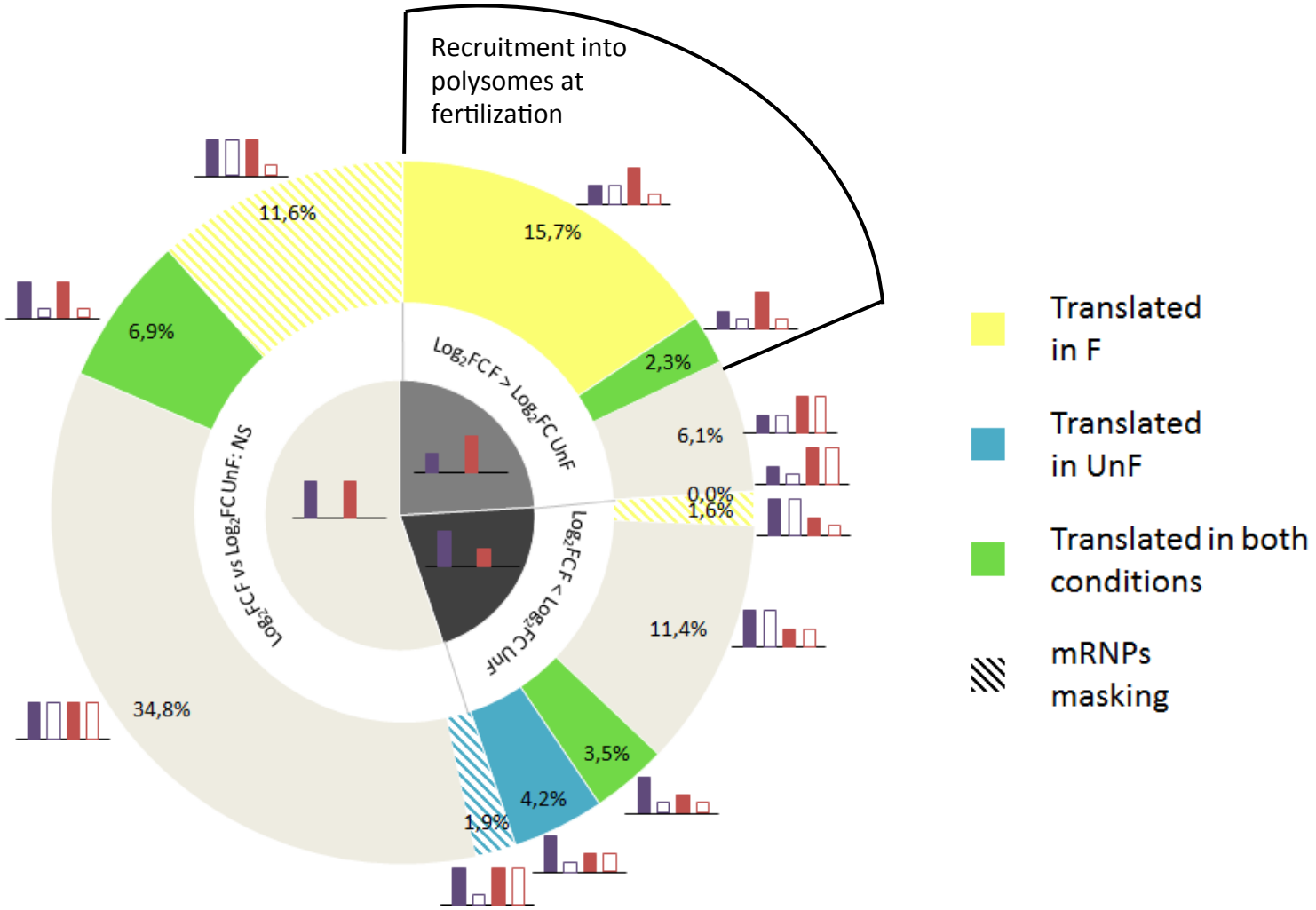

Recruitment into polysomes at fertilization
11,6%
15,7%
6,9%
2,3%
6,1%
0,0%
1,6%
11,4%
34,8%
3,5%
4,2%
1,9%
Log2FCF > Log2FC UnF
Log2FCF < Log2FC UnF
Log2FCF vs Log2FC UnF: NS
Translated in F
Translated in UnF
Translated in both conditions
mRNPs masking

Fig. S3

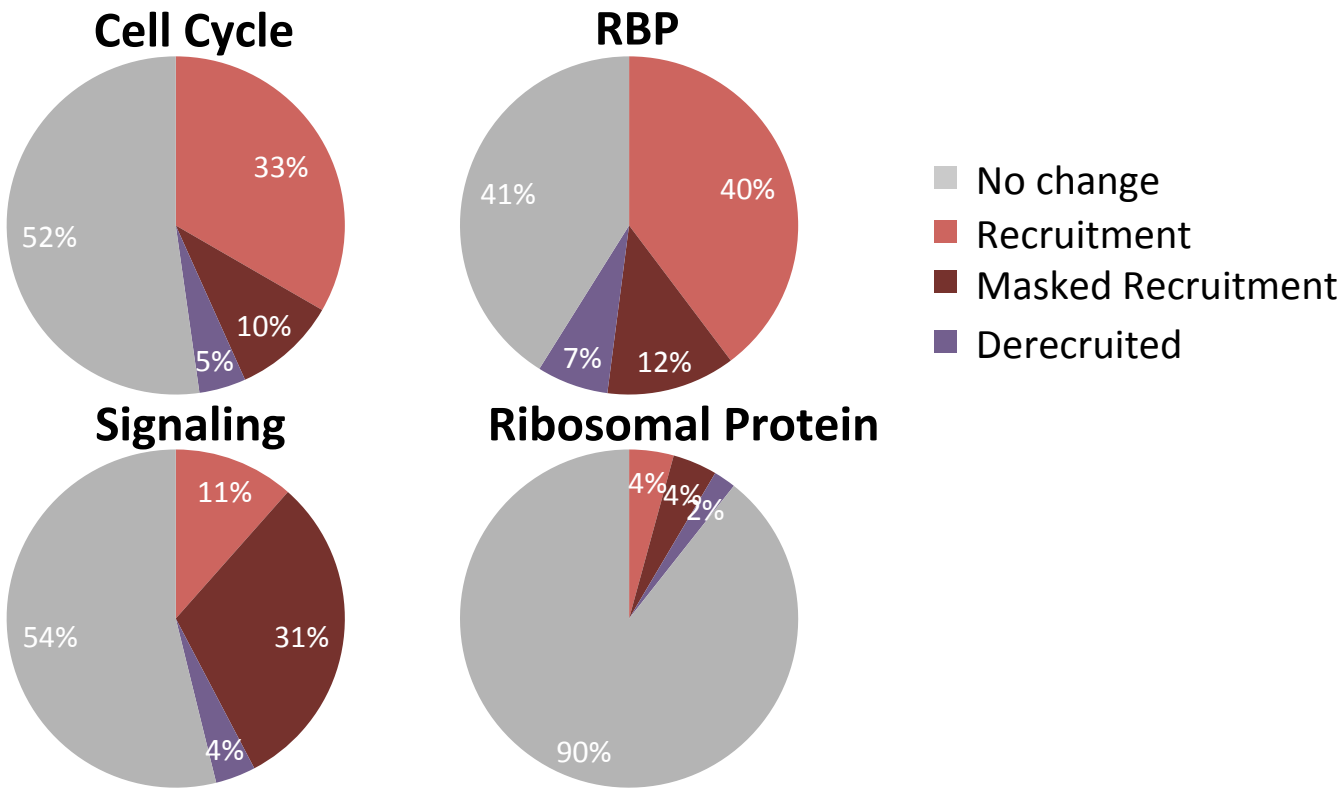

Cell Cycle
33%
52%
10%
5%
RBP
40%
41%
12%
7%
Signaling
11%
54%
31%
4%
Ribosomal Protein
4%
4%
2%
90%
No change
Recruitment
Masked Recruitment
Derecruited

Fig. S4

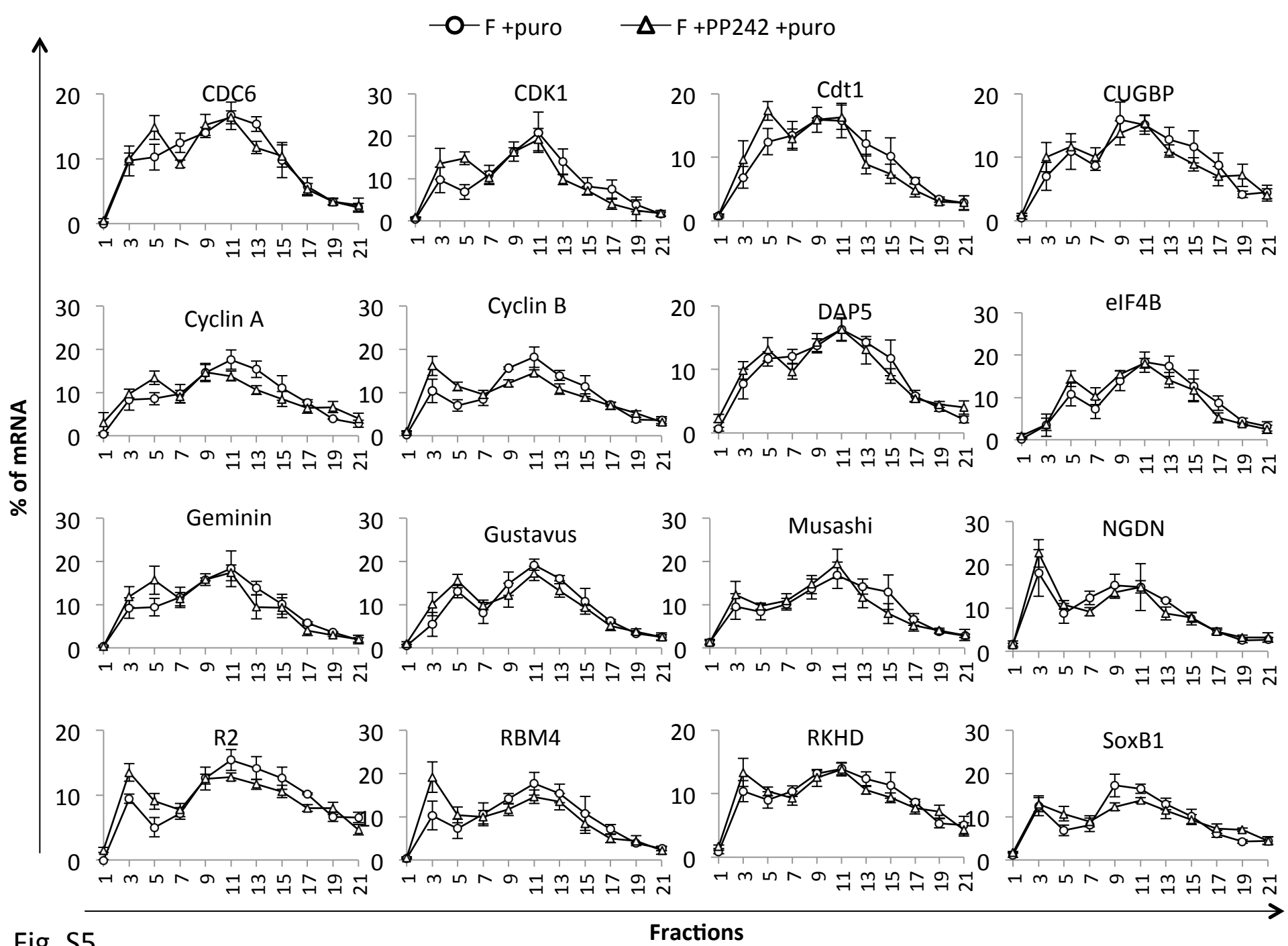

F +puro
F +PP242 +puro
CDC6
CDK1
Cdt1
CUGBP
Cyclin A
Cyclin B
DAP5
eIF4B
Geminin
Gustavus
Musashi
NGDN
R2
RBM4
RKHD
SoxB1
% of mRNA
Fractions

Fig. S5