Université de Montréal

Comparaison des méthodes d'analyse de l'expression différentielle basée sur la dépendance des niveaux d'expression

par
François Lefebvre

Département de biochimie
Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de M.Sc.
en bio-informatique

Mars, 2011

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

Comparaison des méthodes d'analyse de l'expression différentielle basée sur la dépendance des niveaux d'expression

Présenté par :
François Lefebvre

a été évalué par un jury composé des personnes suivantes :

Nicolas Lartillot Ph.D., président-rapporteur
Sébastien Lemieux Ph.D., directeur de recherche
Nadia El-Mabrouk Ph.D., membre du jury

Résumé

La technologie des microarrays demeure à ce jour un outil important pour la mesure de l'expression génique. Au-delà de la technologie elle-même, l'analyse des données provenant des microarrays constitue un problème statistique complexe, ce qui explique la myriade de méthodes proposées pour le pré-traitement et en particulier, l'analyse de l'expression différentielle. Toutefois, l'absence de données de calibration ou de méthodologie de comparaison appropriée a empêché l'émergence d'un consensus quant aux méthodes d'analyse optimales. En conséquence, la décision de l'analyste de choisir telle méthode plutôt qu'une autre se fera la plupart du temps de façon subjective, en se basant par exemple sur la facilité d'utilisation, l'accès au logiciel ou la popularité. Ce mémoire présente une approche nouvelle au problème de la comparaison des méthodes d'analyse de l'expression différentielle.

Plus de 800 pipelines d'analyse sont appliqués à plus d'une centaine d'expériences sur deux plateformes Affymetrix différentes. La performance de chacun des pipelines est évaluée en calculant le niveau moyen de co-régulation par l'entremise de scores d'enrichissements pour différentes collections de signatures moléculaires. L'approche comparative proposée repose donc sur un ensemble varié de données biologiques pertinentes, ne confond pas la reproductibilité avec l'exactitude et peut facilement être appliquée à de nouvelles méthodes. Parmi les méthodes testées, la supériorité de la sommarisation FARMS et de la statistique de l'expression différentielle TREAT est sans équivoque. De plus, les résultats obtenus quant à la statistique d'expression différentielle corroborent les conclusions d'autres études récentes à propos de l'importance de prendre en compte la grandeur du changement en plus de sa significativité statistique.

Mots-clés : microarrays, puces à ADN, expression différentielle, fold-change, Affymetrix.

Abstract

Microarrays remain an important tool for the measurement of gene expression, and a myriad of methods for their pre-processing or statistical testing of differential expression has been proposed in the past. However, insufficient and sometimes contradictory evidence has prevented the emergence of a strong consensus over a preferred methodology. This leaves microarray practitioners to somewhat arbitrarily decide which method should be used to analyze their data. Here we present a novel approach to the problem of comparing methods for the identification of differentially expressed genes.

Over eight hundred analytic pipelines were applied to more than a hundred independent microarray experiments. The accuracy of each analytic pipeline was assessed by measuring the average level of co-regulation uncovered across all data sets. This analysis thus relies on a varied set of biologically relevant data, does not confound reproducibility for accuracy and can easily be extended to future analytic pipelines. This procedure identified FARMS summarization and the TREAT gene ordering statistic as algorithms significantly more accurate than other alternatives. Most interestingly, our results corroborate recent findings about the importance of taking the magnitude of change into account along with an assessment of statistical significance.

Keywords : microarrays, differential expression, fold-change, Affymetrix.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
ARN	Acide ribonucléique
ARNc	Acide ribonucléique complémentaire
ARNm	Acide ribonucléique messager
ARNr	Acide ribonucléique ribosomal
ARNt	Acide ribonucléique de transfert
AUC	*Area under the curve*
avgdiff	*Average difference*
CRE	*cAMP response-element*
CREB	*cAMP response-element binding*
DEG	*Differentially expressed gene*
FARMS	*Factor Analysis for Robust Microarray Summarization*
FC	*Fold-change*
FDR	*False discovery rate*
GEO	*Gene Expression Omnibus*
GO	*Gene Ontology*
GSEA	*Gene set enrichment analysis*
lm	modèle linéaire
loess	*Locally weighted scatterplot smoothing*
MAQC	*Microarray Quality Control Project*
MM	*Mismatch*
nolm	sans modèle linéaire
PLM	*Probe-level model*
PM	*Perfect match*
pmonly	*Perfect Match only*
RMA	*Robust multi-array average*
SNP	*Single nucleotide polymorphism*
SNR	*Signal-to-noise ratio*
TREAT	*t-tests relative to a threshold*
UV	Ultraviolet
VSNRMA	*Variance stabilizing normalization,* suivie de *RMA*

Remerciements

Je tiens tout d'abords à remercier mon directeur de recherche, Sébastien Lemieux pour l'encadrement stimulant et chaleureux qu'il m'a offert lors de mon passage à dans son laboratoire à l'IRIC. Ensuite, je remercie Slim Fourati et Mathieu Courcelles qui furent de formidables camarades de parcours. Finalement, je remercie de tout cœur ma famille et ma copine Nancy pour le soutien moral apporté, particulièrement à l'occasion de la rédaction de ce mémoire.

Avant-propos

La biologie moléculaire est l'étude du vivant à l'échelle de la molécule. Dans le passé, cette science nous a donné par exemple l'insuline synthétique, les tests d'empreinte génétique en médecine légale et a accéléré le développement de nombreux vaccins. Plus récemment, on a assisté à l'émergence de technologies à *haut débit* qui rendent possible la mesure de milliers de variables biologiques à la fois. La technologie des *microarrays* est une addition récente à l'arsenal de la biologie moléculaire qui s'inscrit dans la foulée des techniques rendues possibles grâce au séquençage de génomes entiers. Essentiellement, un microarray permet de quantifier l'*expression* de plusieurs gènes simultanément. Pour diverses raisons qui seront amplement rapportées dans ce mémoire, l'analyse des données brutes provenant des microarrays est tout sauf triviale, ce qui se reflète par un déluge de méthodes proposées dans la littérature, chaque auteur vantant évidemment les mérites de sa proposition. Ce mémoire porte sur les méthodes d'analyse des données, en particulier sur leur comparaison.

Introduction

1. Biologie

La présente section se veut une présentation hautement vulgarisée des notions de base de la biologie moléculaire qui sont en principe nécessaires pour bien mettre en contexte la technologie des microarrays. Les mécanismes et la nature des molécules impliquées sont éminemment plus complexes. Une description plus complète est disponible dans les manuels tels que celui de Voet et Voet[1].

Introduction à l'ADN, l'ARN et les protéines

À quelques exceptions près, les plans pour la construction et le maintien des formes de vie sur terre sont encodés sous forme d'*acide désoxyribonucléique* (ADN). L'ADN est un polymère (une longue chaîne) formé par la concaténation de *nucléotides*, petites molécules simples composées d'un sucre, d'une base azotée et d'un phosphate (Figure 1). Les nucléotides de l'ADN se retrouvent généralement en quatre variétés ne différant que par le type de base, **A** (adénine), **C** (cytosine), **T** (thymine) ou **G** (guanine). Deux caractéristiques chimiques des nucléotides sont remarquables :

1. En formant des liens covalents (phosphodiesters), les nucléotides peuvent être concaténés en longs polymères, ou chaînes. Ces polymères se présentent donc comme une séquence de A, C, G, T et sont potentiellement porteurs de ce qui pourrait être qualifié d'*information*.

2. La formation de ponts hydrogène est possible entre les bases azotées de C et G ainsi qu'entre celles de A et T qui sont dît*es paires de bases complémentaires*. En conséquence, la formation d'un duplex de deux brins complémentaires, appelée *hybridation*, est énergétiquement favorable. De plus, ce duplex a une tendance

naturelle à s'enrouler sous la forme bien connue de double hélice, caractérisée par Watson et Crick en 1953.

La complémentarité des bases azotées implique qu'un polymère de nucléotides puisse servir de patron pour sa propre réplication, guidant la synthèse d'un brin fille. Ce phénomène est non seulement à la base de la réplication de l'ADN des organismes contemporains, mais expliquerait aussi avec élégance l'émergence de l'ancêtre ultime. Cet ancêtre ne fut peut-être qu'un simple polymère dont la séquence a permis l'autoréplication en quantité suffisante pour assurer sa perpétuation, mettant ainsi en marche l'évolution darwinienne par sélection naturelle. On peut alors facilement s'imaginer comment des erreurs de réplications auraient pu ainsi générer des séquences plus aptes à la réplication, par exemple capables de recruter des molécules protectrices ou de catalyser plus efficacement la réplication. Après plusieurs générations, certaines variétés en seraient venues à dominer la population originelle, pour finalement aboutir, des centaines de millions d'années plus tard, à un primate en costard élevant le bras en l'air pour appeler un taxi dans un décor urbain... .

Cette brève incursion dans le sujet de l'origine de la vie et de l'évolution n'était qu'une mise en contexte avant d'introduire quelques notions de base de la biologie moléculaire.

La *cellule* est l'unité structurelle et fonctionnelle de base de tous les organismes vivants connus, des bactéries, unicellulaires jusqu'à *Homo sapiens*, multicellulaire. Elle est un milieu relativement fermé à l'intérieur duquel se produit toute une panoplie de réactions chimiques entre différentes molécules. Au sein de la cellule, les *protéines* tiennent décidément le premier rôle. En fait, la cellule peut être vue comme une machine constituée de milliers de types de protéines interagissant entre elles, avec des acides nucléiques et autres molécules moins complexes (lipides, sucres, etc.). Comme l'ADN, une protéine est un polymère, construit cette fois par la concaténation de 20 différents *acides aminés* (abrév. *aa*). Ces polymères ont la particularité de se *replier* en une structure tridimensionnelle qui

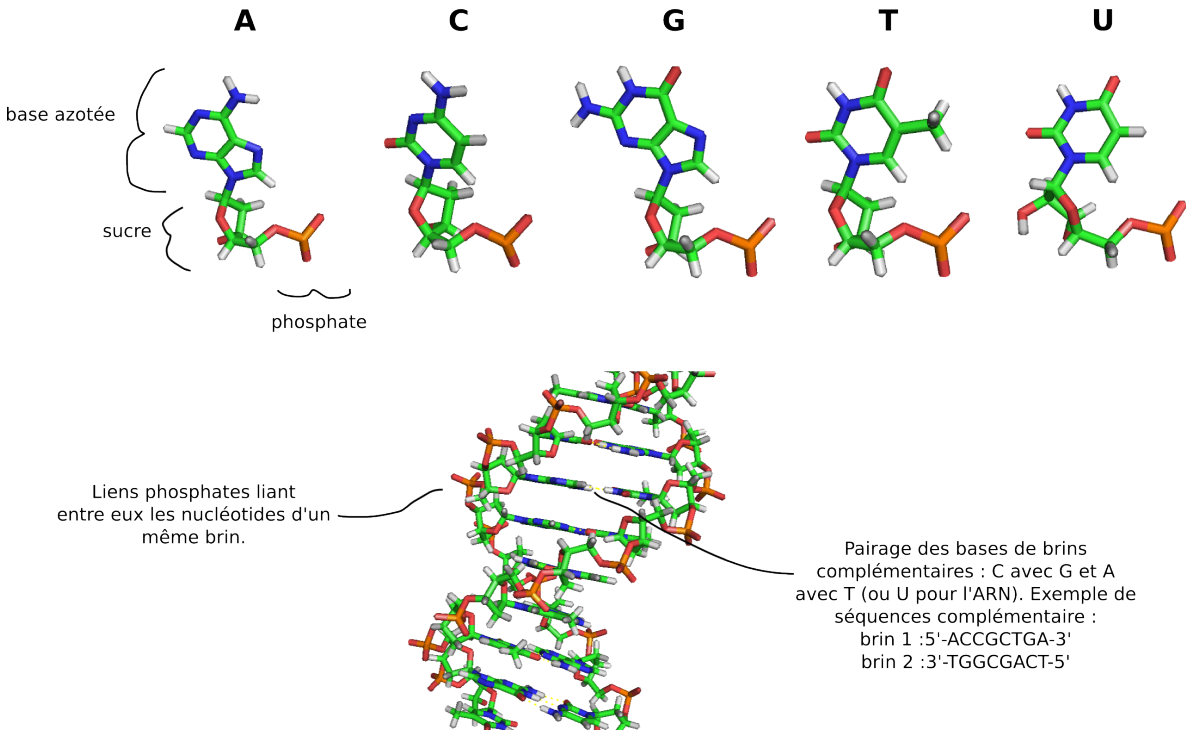


Figure 1: Nucléotides et ADN sous sa forme double brin.

dépend essentiellement de la séquence d'acides aminés qui les composent. Différentes séquences donneront différentes structures, et par conséquent différentes propriétés chimiques et possibilités de fonctions. C'est là que se trouve probablement l'explication de l'ubiquité des protéines : la combinatoire des acides aminés a permis une évolution vers l'improbable et stupéfiante machinerie cellulaire contemporaine. Par exemple, les protéines d'*actine* s'assemblent en filaments qui sont utilisés, entre autres, pour donner une forme à la cellule (cytosquelette) et participer à la contraction des muscles. Les protéines de la classe *hémoglobine*, qui forment les globules rouges, sont capables de lier l'oxygène grâce à un ion de fer et ainsi transporter l'oxygène dans le sang. Un dernier exemple pourrait être celui de l'*insuline* et de son *récepteur de l'insuline*. Sans entrer dans les détails, le récepteur de l'insuline est une protéine *transmembranaire* (à cheval entre l'intérieur et l'extérieur de la cellule) que l'on retrouve abondamment à la surface des cellules du foie, des muscles et tissus adipeux. L'insuline quant à elle est sécrétée par les cellules du pancréas et qui peut lier et activer les récepteurs correspondants. Cette activation provoque une cascade de

réactions à l'intérieur de la cellule qui ultimement résultent en l'absorption de glucose et son stockage sous une autre forme. Dans ce cas précis, les protéines servent à transmettre un message d'une cellule à une autre par le biais de l'insuline, une *hormone*.

Les quelques protéines mentionnées ci-haut ne sont que des exemples parmi les dizaines de milliers de protéines humaines (différentes séquences d'acides aminés), et les milliards de protéines différentes qui existent ou qui ont pu exister au cours de l'évolution de la vie sur terre. Ce qu'il faut retenir, c'est essentiellement l'importance de la séquence d'une protéine pour sa fonction, ce qui ramène à l'ADN, lieu de stockage de l'information de ces séquences protéiques, molécule de l'hérédité.

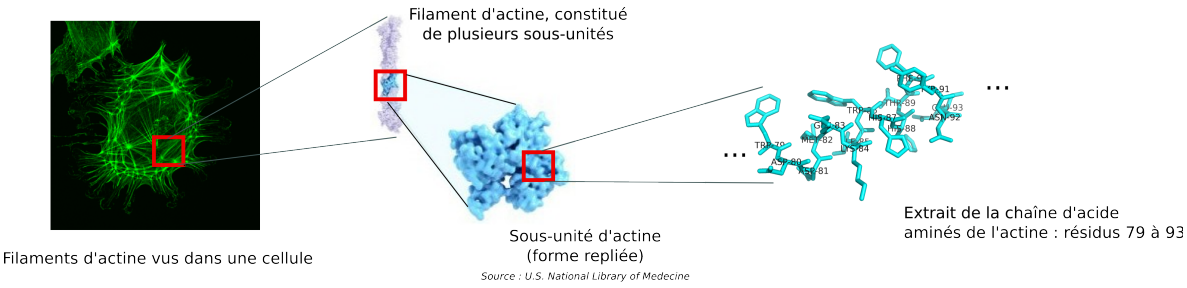


Figure 2: L'actine, un exemple de protéine.

Introduction au gène

Un *gène* est une section d'une molécule d'ADN codant (généralement) pour une protéine. L'ensemble des molécules d'ADN d'un organisme donné est appelé *génome.* Ce dernier encode évidemment plusieurs gènes et leurs protéines correspondantes. Il est question ici de *code,* car l'alphabet de l'ADN (A, C, T, G) n'est pas le même que celui des protéines (les 20 acides aminés). Plus spécifiquement, chaque acide aminé est encodé par trois bases (codon) selon un code génétique relativement uniforme d'une espèce à une autre.

Ainsi, la molécule d'ADN contient les plans pour la construction des différentes protéines. Le passage de l'ADN aux protéines se produit en deux étapes, soit la *transcription* et la *traduction*. Ces deux étapes constituent ce qui est communément appelé

le *dogme central de la biologie moléculair*e : l'information génétique se propage de l'ADN vers les protéines.

Lors de la transcription, la séquence d'un gène dans l'ADN double-hélice est transcrite en plusieurs copies. Ce sont ces copies qui peuvent ensuite transiter ailleurs dans la cellule et qui seront ultimement utilisées pour fabriquer la protéine correspondant au gène. Un peu à la manière dont serait traité un manuscrit ancien de grande valeur, il est judicieux de transcrire l'original en plusieurs copies, qui quant à lui reste bien à l'abri au musée. Plus particulièrement, l'ADN est lu par un enzyme (étant lui-même un assemblage de plusieurs protéines...) nommé *polymérase*. Cet enzyme s'attache à un des brins de l'ADN et glisse le long de celui-ci dans une direction particulière, dite de 5' en 3'. Notons que la polymérase ne s'attache et ne se détache pas n'importe où le long du génome. La transcription commence généralement aux alentours de séquences dites « promoteurs » et se termine aux « terminateurs ». Lors de son glissement sur l'ADN, la polymérase « attrape » des nucléotides libres déjà présents dans le milieu et les concatène en une chaîne complémentaire au brin d'ADN « lu ». Cette nouvelle chaîne n'est toutefois pas composée d'ADN, mais d'*ARN* (*acide ribonucléique)*. L'ARN est un polymère de nucléotides, tout comme l'ADN, mais à quelques différences près : le sucre est un ribose plutôt qu'un désoxyribose et l'uracile (U) remplace la thymine (T) parmi les quatre bases azotées possibles. L'ARN est aussi généralement simple brin, quoiqu'il se replie fréquemment en duplex double-brin avec lui-même. Soulignons finalement que l'ARN n'occupe pas seulement la fonction de messager entre l'ADN du génome et la protéine encodée. On appelle *ARN messagers* (ARNm) les ARN contenant la séquence d'une protéine, mais l'ARN en général peut occuper bien d'autres fonctions.

La *traduction* est le processus par lequel l'ARN messager d'un gène est lu et traduit en une chaîne d'acides aminés qui formera la protéine encodée par le gène (Figure 3). Notons que la transcription décrite précédemment exploite la complémentarité 1:1 des bases de l'ARN avec l'ADN pour produire la molécule d'ARN messager. De façon similaire, la traduction exploite la complémentarité des bases azotées, mais cette fois par

l'entremise d'un *ARN de transfert* (ARNt). Un ARN de transfert est essentiellement une molécule d'ARN avec deux caractéristiques essentielles. La première est son *anticodon*, une série de trois nucléotides situés dans une boucle à la base de l'ARNt. L'anticodon permet à l'ARNt de reconnaître une série de trois nucléotides complémentaires sur un ARNm. La seconde est le type d'acide aminé qu'il peut porter et qui dépend du type particulier d'ARNt. Sans trop entrer dans les détails, seules certaines combinaisons anticodon/acide aminé sont possibles chez un organisme donné. La raison est que les ARNt sont eux-mêmes transcrits à partir du génome où seules certaines combinaisons y sont présentes. Il est maintenant possible de comprendre comment les ARNt servent de *molécule adaptatrice* entre les acides aminés et les codons de l'ARNm, suivant un code génétique bien précis.

Pour terminer, la traduction ne se réalise cependant pas spontanément. Elle nécessite aussi l'intervention du ribosome, gigantesque complexe d'ARN (ARNr) et de protéines. Le rôle du ribosome est essentiellement de lire l'ARNm et de le traduire en attirant un ARN de transfert complémentaire au codon courant, d'y détacher l'acide aminé porté pour joindre cet acide aminé au peptide naissant à mesure que le ribosome progresse.

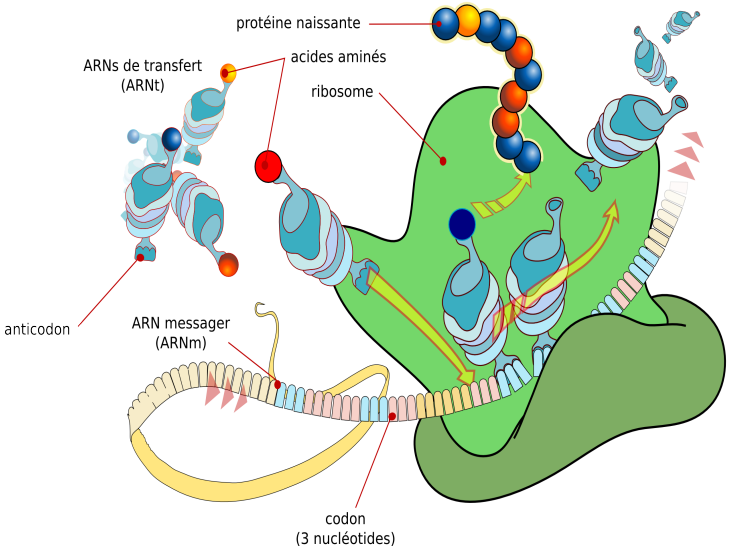


Figure 3: Schématisation du processus de traduction d'un gène. Source: Wikipedia.org.

Régulation de l'expression

Le génome d'un organisme particulier contient des milliers de gènes différents, et ce génome est généralement le même d'une cellule à une autre. Cependant, il est évident que d'un tissu à un autre, à travers le temps ou dans différentes conditions, le programme d'expression des gènes n'est pas le même. Autrement dit, différentes cellules dans différentes conditions n'expriment pas nécessairement les gènes dans les mêmes quantités. Par exemple, les cellules des tissus adipeux (*adipocytes*), spécialisées dans l'entreposage d'énergie sous forme de gras, n'ont certainement pas le même programme d'expression qu'une cellule rétinienne, qui elle doit exprimer des protéines telles que l'opsine et la rhodopsine pour accomplir son travail de photoréception. Un autre exemple pour illustrer la chose est celui de la réponse au stress : la stimulation des cellules du foie par l'épinéphrine (adrénaline) provoque une cascade de signalisation qui mène à la transcription d'une panoplie de gènes d'enzymes impliqués dans la production de glucose (Figure 4). Dans ce dernier cas, une condition différente (exposition à l'épinéphrine) a modifié l'expression des enzymes de la gluconéogenèse dans la cellule.

De façon générale, on peut dire que l'expression des gènes du génome est un processus hautement régulé, car l'expression d'un gène donné peut être sous le contrôle d'un ou plusieurs autres gènes. Par exemple, certains gènes encodent des protéines appelées *facteur de transcription*. Ce type de protéines jouent un rôle de régulation : un facteur de transcription est une protéine capable de lier la région promotrice d'un ou plusieurs gènes afin d'augmenter ou de diminuer le taux de recrutement de la polymérase, modifiant de ce fait le taux de transcription (copies/seconde) du gène cible (les facteurs de transcription peuvent eux-mêmes être sous contrôle d'un autre facteur de transcription, et ainsi de suite). Les gènes de *microARN* encodent de petits ARN complémentaires à un ou d'autres ARNm et peuvent ainsi s'y hybrider, bloquant leur traduction ou encourageant leur dégradation. Il existe bien d'autres mécanismes par lesquels l'expression des gènes est régulée, et les détails varient d'une espèce à une autre. Ces mécanismes peuvent intervenir à différentes étapes de l'expression d'un gène, que ce soit avant la transcription (p.e. au niveau de la

chromatine, ADN compacté), durant la transcription, après la transcription (p.e. *splicing*, *editing*) ou après la traduction (en fait, les protéines subissent toute une panoplie de modifications suite à leur traduction en protéine, par exemple la *glycosylation*). On peut trouver une discussion complète sur ce sujet dans plusieurs manuels comme Latchman[2].

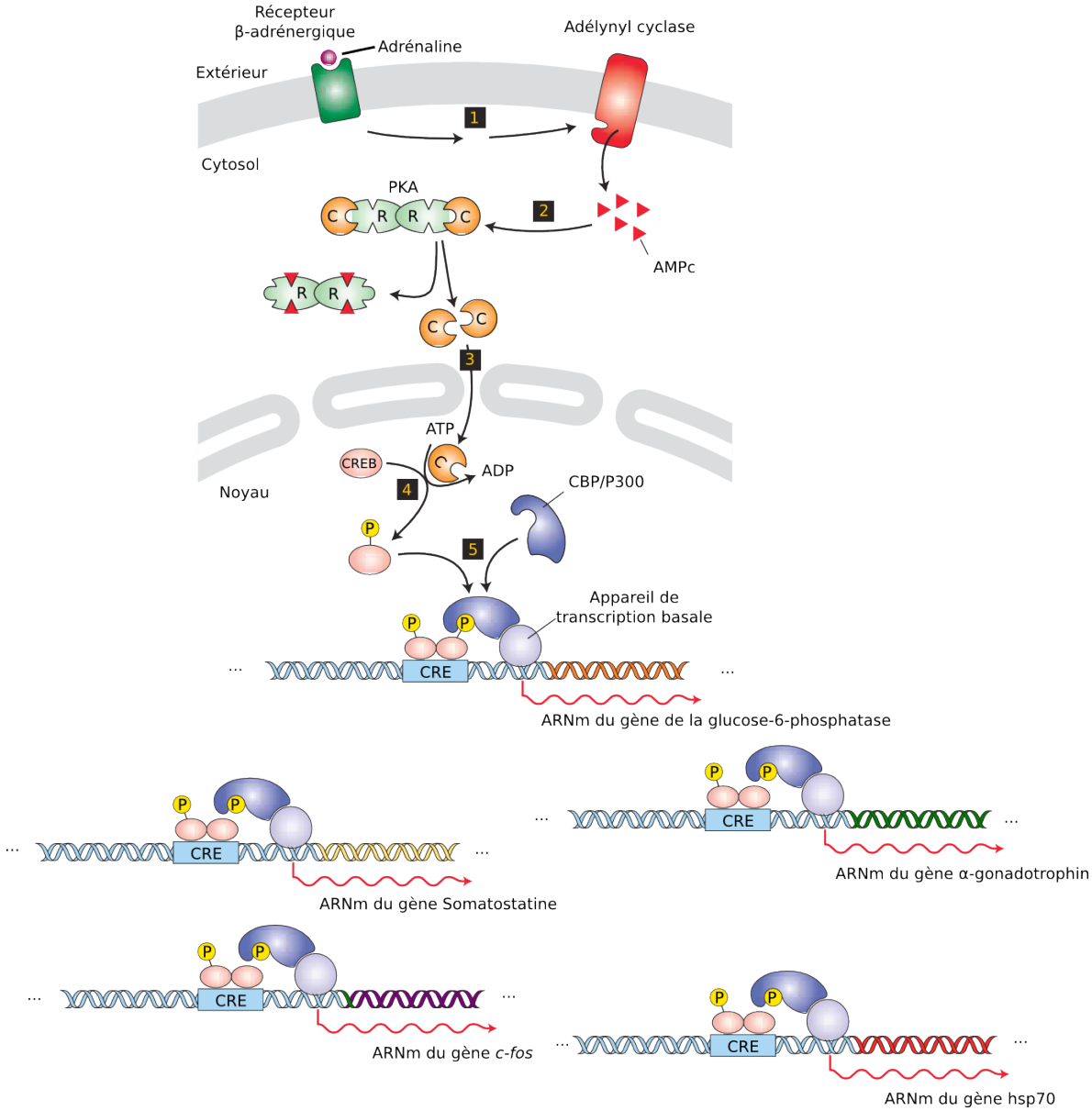


Figure 4: Exemple de mécanisme de (co-)régulation de gènes en réponse à l'adrénaline (épinéphrine). Cette hormone est sécrétée par les glandes surrénales en réponse au stress. L'hormone transite par le sang et active un récepteur β-adrénergique, une protéine encastrée dans la membrane d'une cellule du foie.

L'activation du récepteur provoque une cascade de réactions intermédiaires menant à l'activation (ajout d'un groupe phosphate qui modifie la conformation de la protéine) du facteur de transcription CREB. Ce facteur de transcription ainsi activé lie ensuite de façon spécifique l'élément CRE, une courte séquence d'ADN retrouvée en amont de plusieurs gènes dans le génome. La présence du facteur de transcription a pour effet d'augmenter le taux de recrutement de la machinerie de transcription et ainsi d'augmenter le nombre de copies de gènes liés (entre autres) à la gluconéogenèse (non montré). Le résultat final est une augmentation du niveau de glucose sanguin nécessaire à la réponse au stress. Image modifiée de Lodish, et al., *Molecular Cell Biology*, cinquième édition.

Mesure de l'expression

Cela peut apparaître trivial, mais l'identité et les interactions des molécules impliquées dans un mécanisme donné ne peuvent être élucidées simplement qu'en « regardant à l'intérieur de la cellule ». La raison évidente est que le monde à l'échelle moléculaire est relativement impénétrable en temps réel, encore moins lorsque des milliers de types différents de molécules sont simultanément impliqués. Notre compréhension des mécanismes moléculaires est plutôt le fait d'observations qui, la plupart du temps, sont très incomplètes, indirectes ou sinon à très faible résolution spatiale.

Par exemple, l'interaction entre le facteur de transcription CREB et les éléments génomiques CRE peut être confirmée par une expérience de type ChIP[1]. La cellule est d'abord traitée au formaldéhyde, ce qui a pour effet de fixer (*cross-link*) les protéines à l'ADN. Les cellules sont ensuite lysées (détruites), pour ensuite repêcher spécifiquement, à l'aide d'un anticorps (immunoprécipiter), la protéine d'intérêt en complexe avec l'ADN. On obtient donc finalement seulement la région d'ADN fixée in vivo à la protéine. Pour s'informer sur la réponse à l'épinéphrine, on pourrait aussi créer des souris dîtes *knock-out*, pour lesquelles certains gènes dont on suspecte a priori l'implication (le récepteur par exemple) dans la réponse ont été désactivés. On pourrait ensuite comparer la réaction à l'adrénaline des cellules de ces souris à celle de cellules de souris normales. Ce genre d'étude est toutefois lente, indirecte et très coûteuse.

À défaut de pouvoir suivre dans le temps et l'espace chaque espèce moléculaire de la cellule, il est toutefois dorénavant possible de *quantifier les niveaux d'expressions* des ARNm d'à peu près tous les gènes dans un échantillon biologique donné. À partir d'ici, *niveau d'expression* fera référence à l'expression des ARNm, et non des protéines. Cette équivalence n'est pas nécessairement exacte, mais elle se révèle utile en pratique[3,4]. Cette prouesse est rendue possible par des technologies automatisées de séquençage de l'ADN et d'autres avancées technologiques. En étudiant quels gènes sont exprimés ou non, en quelles quantités, sous différentes conditions expérimentales ou environnementales, on peut en venir à une plus grande compréhension des différents mécanismes cellulaires. Le cas de la réponse à l'épinéphrine peut être donné en exemple. On pourrait par exemple comparer les niveaux d'expression d'un échantillon de cellules exposées à l'épinéphrine à celles d'un échantillon contrôle. Une telle expérience constaterait, sans aucune connaissance préalable sur le mécanisme, l'implication des gènes contrôlés par CREB. Un processus semblable pourrait être répété pour différents types de tissus, à différents stades de développement, sous différents traitements pharmacologiques, etc.

2. Microarrays

Il existe plusieurs techniques de mesure de l'expression des gènes, chacune avec leur niche d'utilisation, avantages et inconvénients. La technologie sur laquelle porte ce mémoire, celle des microarrays, se distingue de ses prédécesseurs par sa capacité de mesurer simultanément l'expression de milliers de gènes dans un échantillon biologique, et ce à un coût relativement bas en peu de temps. Avant de poursuivre, il est pertinent de mentionner que les microarrays ne servent pas uniquement à mesure l'expression des gènes. Ils sont également employés par exemple pour le génotypage (SNPs), pour identifier les régions génomiques liant une protéine (ChIP-chip), la détection de micros ARNs, ou pour l'hybridation génomique comparative[5]. Cependant, ce mémoire ne considère les microarrays que dans un contexte de mesure de l'expression, ce qui n'implique pas que les résultats ne puissent être pertinents dans un autre contexte.

Dans sa formulation la plus générale, un microarray est un substrat sur lequel ont été fixés des ADN simple brin complémentaires aux molécules d'ARN (ou d'ADN) que l'on cherche à quantifier. Le terme microarray (« microgrille ») vient du fait que ces ADN ont été positionnés de façon systématique, c'est-à-dire que l'identité de l'ADN fixé à une position donnée est prédéterminée. Chacune de ces positions dans la grille est appelée *sonde* (*probe*). C'est sur le phénomène d'hybridation des acides nucléiques que reposent le fonctionnement des microarrays : comme les séquences complémentaires ont tendance à s'hybrider l'une à l'autre, les molécules d'acide nucléique de séquence S_c dans l'échantillon (la cible) devraient s'hybrider aux molécules d'une sonde dont la séquence S_s est complémentaire (idéalement, S_s est une sous-séquence complémentaire à S_c). Pour chacune des sondes, le nombre d'hybrides sonde/cible devrait augmenter avec la quantité de molécules de la cible présente initialement dans l'échantillon. Une quantification du nombre d'hybrides peut être accomplie si les molécules de la cible ont été préalablement marquées, par exemple en utilisant un fluorochrome; Suite à une étape de lavage, l'intensité lumineuse de chaque sonde, en réaction à un éclairage laser quantifie alors le nombre d'hybrides sondes/cible. Cette intensité est liée au nombre de molécules de la cible présentes initialement dans l'échantillon.

Quoiqu'elles soient toutes basées sur le principe d'hybridation, on peut répertorier deux types de microarrays fortement utilisés en pratique :

- *Les puces à ADNc.* Les sondes de ce type de microarray sont préparées séparément à partir de clones d'ADNc, soit de l'ARNm mature (sans les *introns*) rétrotranscrit (*reverse-transcribed*) en ADN, puis cloné (chez *E. coli* par exemple), puis amplifié par PCR. Un robot pipetteur est ensuite utilisé pour déposer une goutte de chaque sonde préparée sur le substrat, typiquement une lame de verre. Cette technologie, plus ancienne, comporte des avantages et des inconvénients discutés en détail par Kohane[4]. Les puces à ADNc sont typiquement associées (mais pas exclusivement) à l'approche *deux-couleurs* pour laquelle deux échantillons portant deux marqueurs différents y sont hybridés. L'approche deux-couleurs est statistiquement

intéressante, car elle implémente directement la comparaison de deux échantillons, ce qui confond certaines sources de variabilité.

- «Les microarrays à *oligonucléotides* emploient des sondes entièrement synthétisées, dont la séquence est prédéterminée. Les étapes de clonage, de PCR, source de bruit et d'erreurs expérimentales, sont évacuées du processus, en plus d'offrir une flexibilité totale quant aux questions posées par la puce (p.e. SNPs, isoformes). Les oligonucléotides peuvent être synthétisés directement sur le substrat (in situ) par photolithographie (p.e. Affymetrix, voir section suivante), par jet « d'encre » (*inkjet* p.e. Agilent). Une autre technologie, celle des *beadArrays* de la compagnie Illumina, attache les oligonucléotides à des billes de verres qui sont ensuite dispersées aléatoirement dans des puits hémisphériques sur un substrat de verre. À chaque oligonucléotide est aussi ajouté une séquence signature, grâce à laquelle un astucieux algorithme de décodage[6] par hybridations successives peut retrouver l'identité de chaque oligonucléotide de chaque bille pour chacune des positions sur la puce.

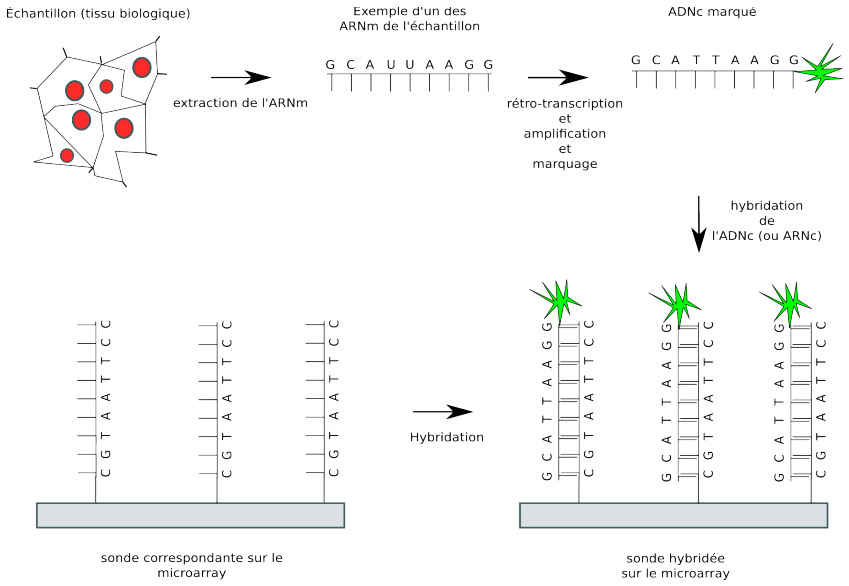


Figure 5 : Principe général des microarrays dans le contexte de mesure de l'expression. L'ARNm est extrait de l'échantillon biologique puis rétrotranscrit en ADNc (ADN complémentaire) pour être ensuite amplifié et marqué avec un fluorochrome. Cet ADNc peut ensuite être hybridé sur le microarray. La

quantification du nombre d'hybrides cible/sonde peut procéder en éclairant le microarray au laser (non illustré).

La technologie GeneChip® d'Affymetrix

Les puces à oligonucléotides à haute densité d'Affymetrix (Santa Clara, CA, É.-U.) ont longtemps fait figure de référence pour la mesure de l'expression par microarrays, comme en témoigne le nombre d'expériences soumises à ce jour au Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). C'est d'ailleurs pour cette raison qu'il a été choisi d'effectuer l'analyse de comparaison proposée dans ce mémoire sur les GeneChip®. Cette section introduit donc quelques aspects importants de cette technologie. De plus amples détails peuvent être trouvés entre autres dans Lipschutz[7] ou dans la documentation technique disponible sur le site web d'Affymetrix.

Les sondes de la technologie GeneChip sont des oligonucléotides de longueur 25 dont la séquence est complémentaire aux gènes sondés. Les oligonucléotides sont synthétisés *in situ* et en parallèle sur une gaufrette (qui donnera plusieurs microarrays) de quartz par un procédé photolithographique décrit par Pease et al.[8]. Le procédé repose sur l'emploi d'un groupement chimique *de liaison* capable de s'attacher à un nucléotide et d'empêcher la liaison d'un autre nucléotide, mais qui est toutefois photosensible (détruit par la lumière UV). En employant une série de masques photolithographiques ne laissant passer les UV qu'aux positions désirées de la gaufrette, il est possible de « photodéprotéger » uniquement les positions prédéterminées, pour ensuite exposer la gaufrette à une solution d'un des quatre nucléotides voulus. Le résultat est un microarray à très haute densité pour lequel l'espace occupé par une sonde est de l'ordre du micron. Lors de son introduction, cette technologie rendit possible la quantification de l'ensemble des transcrits de génomes de grande taille comme celui de l'humain, de la souris, drosophile, etc.

Afin de compenser pour la courte longueur des sondes, plusieurs sondes différentes (typiquement 11 à 20) interrogent un même gène. L'ensemble des sondes d'un même gène (plus rigoureusement, une *séquence référence*) est alors appelé *probeset* (Figure 6), que l'on

pourrait traduire par *ensemble-sonde*. Le nombre de probesets présents sur une puce est de l'ordre de plusieurs dizaines de milliers, et parfois un même gène peut être représenté par plus d'un probeset à la fois. Inversement, une sonde ou un probeset en entier peut représenter plus d'un gène, ces derniers étant le plus souvent membre d'une même famille de gènes. Les sondes d'un même probeset sont disposées aléatoirement sur les puces récentes, afin d'être plus robustes à d'éventuels artefacts spatiaux. Notons aussi que ce type de puces est appelé 3'-IVT, les sondes étant choisies surtout en région 3', pour des raisons techniques, car c'est à cette extrémité (la queue poly-A) que débute la rétrotranscription. La sélection des sondes, c'est-à-dire le choix de quelles régions des gènes seront sondés par la puce, est en soi une question assez complexe, discutée par Mei et al.[9] dans le cas des puces Affymetrix. Pour résumer la chose, il s'agit de choisir les sondes de façon à ce qu'elles soient spécifiques au gène, à limiter l'hybridation croisée (des séquences quasi complémentaires peuvent parfois s'hybrider) et à maximiser la liaison entre la sonde et la cible. Le problème en est un de thermodynamique de l'hybridation des acides nucléiques qui n'est pas nécessairement trivial.

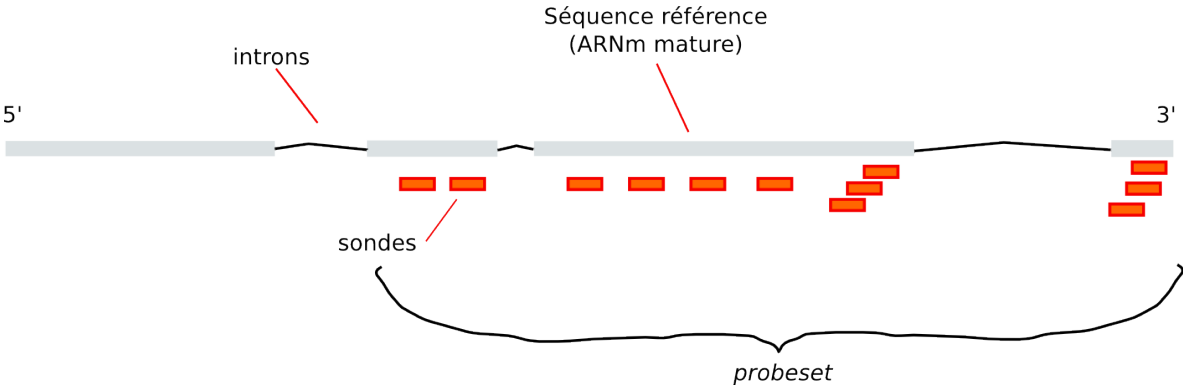


Figure 6: Illustration d'un probeset sondant un gène. En pratique, le choix initial de la position des sondes fut basé sur une version donnée d'un génome qui peut avoir changé depuis. Certaines sondes peuvent correspondre à plusieurs transcrits et vice-versa.

Une caractéristique particulière des puces GeneChip est que chaque sonde se révèle en fait être une *paire de sonde* (*probe pair,* côte à côte sur la puce) : la première, *PM* (*perfect match*) correspond exactement à la séquence sondée alors que la seconde *MM* (*mismatch*) contient une erreur pour la base en position centrale de l'oligonucléotide. L'idée

initiale des sondes MM fut d'estimer les niveaux d'hybridation non spécifiques. Quoique certains algorithmes d'analyse des données exploitent les valeurs MM, leur utilité réelle reste incertaine (elles occupent, après tout, la moitié de l'espace sur la puce).

Quoique ce mémoire ne s'intéresse qu'à l'analyse des intensités brutes, il peut être utile de mentionner la procédure technique qui mène ultimement à ces dernières lors d'une expérience[4]. Pour un organisme eucaryote (dont les ARNm sont dotés d'une queue poly(A)), l'ARNm est extrait puis rétrotranscrit en ADNc double brin à l'aide d'amorces oligo-dT (thymine) aussi porteuses d'un site de liaison pour la polymérase T7. De l'ARNc marqué à la biotine est ensuite transcrit in vitro à partir de l'ADNc, qui se voit donc amplifié. L'ARNc résultant est fragmenté en des longueurs typiques de 25 à 200 paires de bases, puis injecté dans la cartouche du microarray qui est mis au four à hybridation pendant quelques heures. Le microarray subit un lavage pour ensuite être traité au marqueur fluorescent streptavidine-phycoercine (SAPE), qui lie la biotine portée par l'ARN. L'étape de lecture s'en suit, alors qu'un microscope confocal capte la fluorescence émise par le microarray exposé à un balayage laser. L'image résultante contient plusieurs pixels sur une échelle d'intensités de 16 bits. Une grille est superposée à l'image, la séparant en cellules chacune correspondant à une espèce d'oligonucléotide particulière. L'intensité de chaque cellule (sonde) est rapportée comme le 75e centile de l'intensité des pixels (échelle de 16 bits) correspondants, en excluant les pixels des bordures des cellules. Le tout est écrit dans un fichier au format .CEL, point de départ de l'analyse des données traitées par ce mémoire.

3. Les étapes de l'analyse de données de microarrays

Les intensités brutes provenant d'une expérience de microarrays ne peuvent pas être utilisées directement. La raison est que cette technologie est affligée de nombreuses sources de variation, telles que des variations à l'étape de préparation de l'ARNm, du marquage fluorescent, des différentes affinités des sondes, de l'analyse d'image, etc[5]. Les données de microarrays sont généralement considérées comme étant très *bruitées*, voir parfois même peu reproductibles[10,11], un problème exacerbé par sa haute dimensionalité et le faible niveau

de réplication technique dans un contexte pratique. Heureusement, il est possible d'estimer *a posteriori* certains effets non spécifiques et de les retirer en employant divers algorithmes de *pré-traitement.* Plusieurs méthodes statistiques complexes ont aussi été avancées pour l'analyse statistique des niveaux d'expression. Cette section présente les différentes étapes du paradigme actuel (*stepwise*[1]) pour l'analyse des données provenant de microarrays GeneChip d'Affymetrix, particulièrement pour l'identification des *gènes différentiellement exprimés* (DEGs). Ultimement, la comparaison de différentes méthodes proposées à cet effet est le but du travail rapporté par ce mémoire.

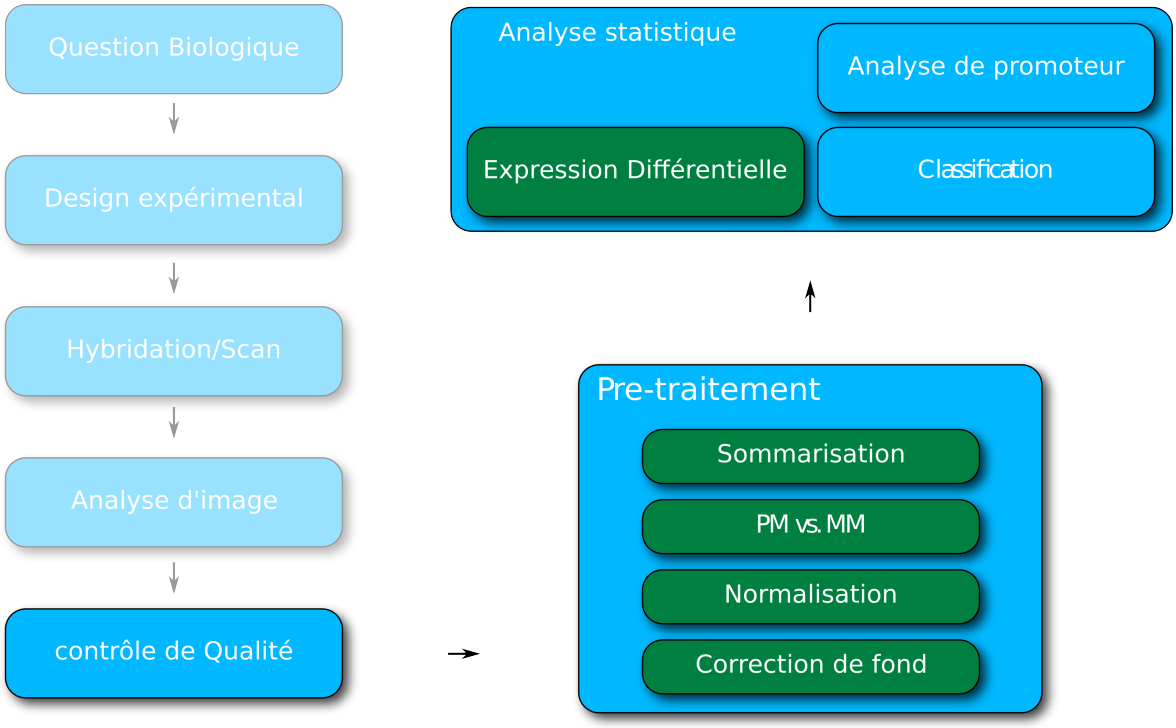


Figure 7: Déroulement typique d'une expérience de microarrays sur la plateforme GeneChip d'Affymetrix.

Pré-traitement

La correction de fond est une étape qui vise à ajuster les intensités brutes afin d'en retirer le signal non spécifique et ainsi augmenter la sensibilité (aptitude à détecter

1 L'approche stepwise consiste à séparer le problèmes en étapes distinctes qui peuvent chacune être effectuées par un algorithme différent, plutôt que par un seul modèle statistique global[12].

l'expression, si expression il y a réellement) du microarray[13]. Cette étape est justifiable par le fait qu'une partie du signal peut provenir d'hybridation d'ARNc non spécifique (non complémentaire), de bruit dans l'instrumentation, ou encore de tout artefact spatial dont peut être affligée une puce.

La normalisation est l'opération qui vise à rendre les puces d'une même expérience *comparables* (ou les différents *canaux* dans le cas des puces à deux couleurs). C'est le processus par lequel on tente de corriger pour des variations techniques confondantes, telle que la quantité initiale d'ARN hybridé sur les différentes puces, des efficacités de marquage fluorescent différentes ou des réglages différents du scanneur, pour n'en nommer que quelques unes[14,15]. La nécessité de normaliser se constate assez facilement en observant la distribution des intensités brutes de n'importe quelle expérience de microarrays (Figure 8).

La correction PM est une étape propre à la technologie GeneChip d'Affymetrix, dont la moitié des sondes est de type MM (mismatch). Tel que discuté précédemment, ces sondes sont présentes sur les puces GeneChip dans le but d'estimer la contribution du signal non spécifique au signal réel. Le but d'une méthode de correction PM est donc de combiner les intensités des sondes PM et MM en une valeur proportionnelle au nombre réel de transcrits cible hybridés à la sonde PM.

La sommarisation une étape propre à toute plateforme pour laquelle un même transcrit est sondé par plusieurs sondes que l'on doit résumé en une seule *valeur d'expression*. Dans le cas particulier des GeneChip d'Affymetrix, les sondes d'un même probeset ne sondent pas la même partie de la séquence cible (Figure 6), ouvrant la porte à des méthodes de sommarisation plus sophistiquées (plus qu'une simple moyenne, par exemple).

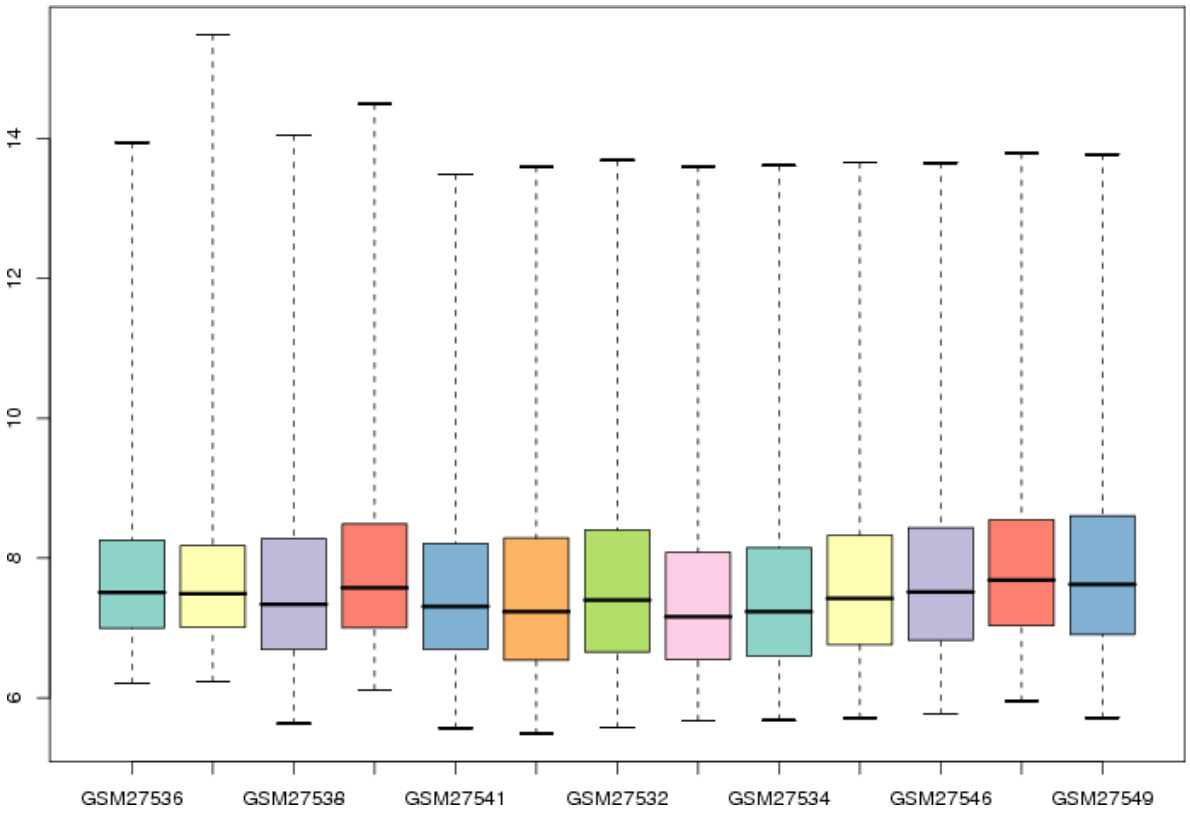


Figure 8: Boxplot des intensités brutes (log2) d'une expérience de microarrays (Wood et al. 2005). Chaque boîte illustre la distribution des intensités de chaque puce. La ligne intérieure d'une boîte indique la valeur médiane, les extrémités les 1er et 3e quartiles. Les différences de médianes sont de bons indices de différences globales qui devraient être nivelées avant d'entreprendre des comparaisons entre les puces.

Expression différentielle, statistiques d'ordonnancement

L'analyse de l'expression différentielle consiste à identifier quels gènes voient leur niveau d'expression varier *entre différentes conditions biologiques* (Figure 9).

Il est important de noter que la comparaison des niveaux d'expression *absolus* ne s'effectue généralement pas *directement entre* les gènes, mais se limite plutôt entre les différents échantillons hy*bridés*. La raison généralement invoquée est que d'un gène à un autre, les unités de la mesure de l'expression ne sont pas les mêmes[12]. Par conséquent, un microarray n'est généralement pas utilisé pour répondre à la question « le gène A est-il plus exprimé que le gène B », mais pour répondre à des questions telles que « le gène A répond-

il à la stimulation X? », « le gène A répond-il plus que le gène B à la stimulation X? » ou encore « l'expression du gène A est-elle corrélée à celle du gène B? ». Ces dernières portant sur les niveaux d'expression relatifs plutôt qu'absolus.

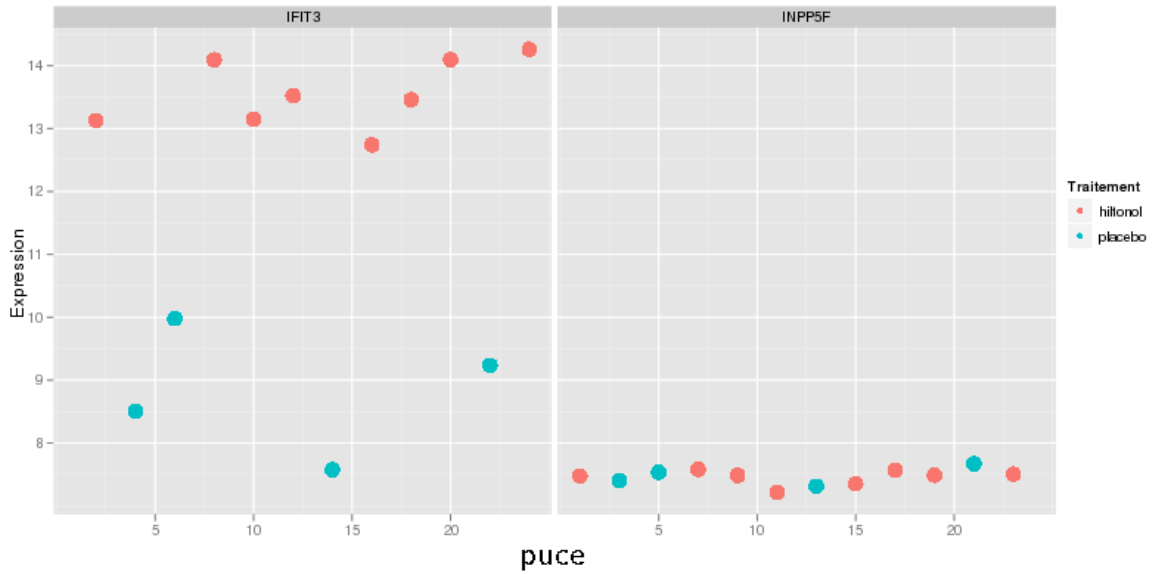


Figure 9: Illustration de l'expression différentielle pour deux gènes. Pour cet exemple, le gène IFIT3 (*interferon-induced protein with tetratricopeptide repeats 3*) est clairement différentiellement exprimé entre les cellules mononucléaires sanguines (PBMC) de patients échantillons traités au hiltonol (poly-ICLC) et les patients du groupe placebo.

Avant de poursuivre, il est aussi important de mentionner qu'en pratique, l'analyse de l'expression différentielle suit le paradigme *gène par gène*. En effet, bien que l'expression génique soit un phénomène coordonné, et que par conséquent les niveaux d'expression des gènes soient fortement dépendants, notre connaissance de cette dépendance (donc de la distribution jointe...) est à ce jour plutôt limitée. Chaque « ligne » de la matrice d'expression est donc habituellement analysée individuellement.

L'analyse de l'expression différentielle suit habituellement la procédure suivante :

Formuler un test d'hypothèse statistique. On définit ici quantitativement ce que l'on entend par expression différentielle en formulant une hypothèse nulle et une hypothèse alternative. Par exemple, on testera l'hypothèse nulle que la moyenne entre deux groupes

(ou le coefficient de régression d'un modèle linéaire) est égale à zéro. Cette hypothèse nulle est la plus courante, mais d'autres sont possibles, comme par exemple de tester si la différence de moyennes dépasse un certain seuil arbitraire ou encore un test sur la différence des médianes.

Calculer une statistique pour l'expression différentielle pour chaque gène. Cette statistique peut être par exemple la simple différence des moyennes (le *fold-change*), le *t* de Student, la différence des médianes, une statistique non paramétrique, etc. .

Si applicable, comparer la valeur de la statistique obtenue précédemment avec la distribution de cette dernière sous l'hypothèse nulle. Par exemple, le *t* de Student suit une distribution bien connue du même nom. Cette étape fournit habituellement une valeur-p, qui correspond à la probabilité d'erreur de type I sous l'hypothèse nulle.

Ordonner les gènes selon la significativité de leur test d'hypothèse. Noter que dans le cas des microarrays Affymetrix, l'ordonnancement par significativité est souvent le même que celui de la valeur absolue de la statistique puisque chaque test comporte le même nombre de degrés de liberté. Pour ce mémoire, on appelle aussi la statistique de l'expression différentielle statistique d'ordonnancement, même si à proprement parler, l'ordonnancement s'effectue la plupart du temps sur une valeur de significativité comme la valeur-p.

Recalculer la significativité pour tenir compte que plusieurs hypothèses ont été testées simultanément (problème des tests multiples). L'analyste sera fréquemment intéressé à obtenir une sous-liste parmi les gènes les plus significatifs. Même si la valeur-p d'un test pris individuellement donne la probabilité d'erreur de type I (hypothèse nulle rejetée faussement, ou faux positif), évaluer la significativité d'une liste de plusieurs tests demande un calcul supplémentaire. L'idée se comprend assez facilement : le bruit expérimental fait inévitablement en sorte que certains gènes, en réalité non différentiellement exprimés, résultent en des valeur-p significatives. La significativité d'une liste sélectionnée par un simple seuil sur la valeur-p n'est donc manifestement pas quantifiée par ce seuil. Une nouvelle définition de la significativité, ainsi qu'une procédure pour l'évaluer sont donc

nécessaires. Par exemple, les procédures pour tests multiples de type FWER (*Family-Wise Error Rate*) tentent d'évaluer probabilité que la liste sélectionnées contiennent au moins un faux positif, alors que les procédures de type FDR (False Discovery Rate) évaluent la *proportion* de faux positifs d'une sous-liste donnée. En fait, la recherche de telles procédures constitue un pan de la recherche en statistiques en soi[16] : la situation se complexifie rapidement si l'on considère par exemple la dépendance des tests d'hypothèse ou si l'on considère les différentes façons de calculer la correction (p.e. ré-échantillonage). Ce mémoire ne s'y attardera pas plus en détail et n'inclura pas l'étape de correction pour tests multiples dans la méthode de comparaison. La justification est que la méthode de comparaison proposée se base sur *l'enrichissement de listes de gènes* (section suivante), mais aussi principalement pour garder la complexité de la comparaison à un niveau acceptable. Aussi semble-t-il opportun de mentionner à ce point que la correction pour tests multiples n'affecte pas l'ordonnancement des gènes. La méthode de comparaison proposée effectue cette dernière sur les ordonnancements complets, sans fixer de seuil spécifique. Un retour sur ce point sera effectué à la fin de la présente introduction.

Analyse d'enrichissement (GSEA)

L'analyse de l'expression différentielle résulte essentiellement en une liste de gènes ordonnés selon une mesure de leur association avec les variables expérimentales. Souvent, l'analyste souhaitera aussi savoir si certains groupes de gènes prédéfinis sont aussi associés aux variables expérimentales, ce qui facilite l'interprétation. Ces groupes de gènes peuvent être, par exemple, des voies moléculaires (*pathways*), des gènes associés à des fonctions biologiques, des gènes cibles d'un même facteur de transcription, etc. Cette étape de l'analyse de microarrays est souvent appelée « analyse d'enrichissement » (Gene Set Enrichment Analysis). Cette étape est discutée ici puisque dans la méthode de comparaison proposée, l'enrichissement de groupes est employé comme métrique pour comparer les méthodes de pré-traitement et statistiques d'ordonnancement.

Plusieurs méthodes statistiques ont été proposées pour effectuer une analyse d'enrichissement. Strimmer et Ackermann[17] proposent à cet égard une taxonomie exceptionnellement utile, en plus d'effectuer une revue assez complète des méthodes disponibles. On y distingue notamment deux types d'approches :

- Les approches *globales*[18,19] qui calculent l'enrichissement d'un groupe de gènes en ajustant un modèle statistique (régression) est directement aux valeurs d'expression.

- Les approches de *niveau-gènes* qui prennent la statistique d'expression différentielle comme point de départ pour l'analyse. Cette seconde catégorie est la plus utilisée en pratique, et les différentes implémentations diffèrent par les critères suivants :

    - la statistique d'expression différentielle employée : celle-ci peut être par exemple une statistique *t*, une statistique *t* régularisée, le coefficient de régression, le coefficient de corrélation, etc.

    - La transformation appliquée à la statistique : rang, binaire (~test de Fisher sur la distribution hypergéométrique), valeur-p, carré, etc.

    - La statistique pour le groupe de gène : moyenne des rangs (Mann-Whithney U, celle qui sera utilisée pour ce projet), *maxmean*[20], moyenne, médiane, etc.

    - Le type d'hypothèse nulle[21-23] recherche-t-on 1) les groupes de gènes dont l'association avec les variables expérimentales est différente de celle du reste des gènes, cette hypothèse est testable en permutant les gènes; ou 2) les groupes de gènes dont l'association avec les variables expérimentales est significative. Ceci est testable en permutant les gènes.

Quoiqu'il s'agisse d'une sujet très intéressant, il serait difficile et peu pertinent d'expliciter l'ensemble des méthodes d'analyse d'enrichissement proposées dans la littérature à ce jour. Un développement récent mérite toutefois mention. Plusieurs auteurs ont en effet souligné que des méthodes simples à calculer, comme par exemple, la moyenne

des statstiques-t, ou leur carré, offrent des performances largement supérieures à GSEA (statistique Kolmogorov-Smirnov pondérée par la statistique d'expression différentielle), une méthode pourtant largement plus populaire en pratique[21,23-25].

4. Contexte : la nécessité de comparer

Tel que mentionné précédemment, le nombre de méthodes de pré-traitement ou de statistiques de l'expression différentielle proposées dans la littérature est ahurissant. La situation est telle qu'en 2009, le journal *Bioinformatics* a cru bon de rappeler en éditorial que tout nouvel algorithme d'analyse de microarrays publié doit offrir un avantage important par rapport aux meilleures méthodes préexistantes, et que cet avantage doit être démontré sur un éventail de données biologiques réelles, et non seulement sur des données simulées[26].

L'« enthousiasme » de la communauté bio-informatique et statistique pour le problème de l'analyse de données de microarrays s'explique probablement de la façon suivante. Premièrement, il y a certainement place à amélioration : la technologie des microarrays est parfois critiquée comme livrant des résultats peu reproductibles, et la méthode d'analyse employée est reconnue comme étant un facteur déterminant des résultats obtenus[27].

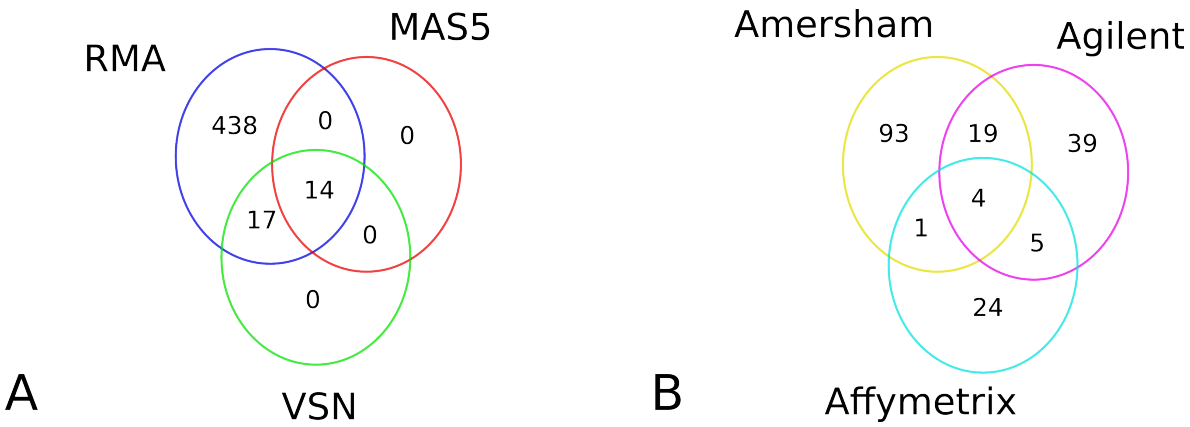


Figure 10 : Illustration du problème de la reproductibilité. (A) Trois méthodes différentes de pré-traitement ont été appliquées aux données de Wood et al.[28]. Le seuil pour l'expression différentielle est ici fixé

à une valeur-q de la statistique-t de 0.05. (B) Diagramme de Venn tirée de Tan et al.[29] comparant les gènes différentiellement exprimés d'une même expérience pour trois plateformes commerciales différentes. Shi et al. ont démontré que le problème de reproductibilité est en grande partie dû à l'emploi des valeurs-p seules sans considération du fold-change[30].

Deuxièmement, le défi intellectuel que représente le problème d'inférences multiples à faible puissance, couplé au prestige scientifique de la génomique, attire de nombreux chercheurs vers le domaine. Parmi ces derniers, bon nombre d'entre eux, provenant de la recherche statistique fondamentale tentent d'appliquer des techniques hautement sophistiquées. Malheureusement, le gain en performance n'est pas nécessairement toujours rendez-vous.

Un troisième facteur contribuant à l'inflation du nombre de méthodes proposées est l'absence d'un critère objectif de comparaison satisfaisant (banc d'essai, angl. *benchmark*), empêchant la matérialisation d'un consensus solide. La comparaison de différentes approches méthodologiques s'effectue généralement en utilisant un ou plusieurs jeux de données de contrôle pour lesquels la « vérité » est connue. Toutefois, un tel jeu de données n'existe pas encore pour les microarrays. On recense par contre deux types de banc de bancs d'essais qui peuvent s'en rapprocher[31]

- Les « spike-ins ». Il s'agit d'expériences artificielles pour lesquelles différents ARN ont été ajoutés en concentrations connues. Par exemple, les *Latin Square* effectués sur deux plateformes humaines Affymetrix différentes[32], sont caractérisés par un petit nombre de gènes différentiellement exprimés (0.2%) avec des fold-change allant de moyen à très élevés (2 à 512). Un mélange complexe d'ARN humain est aussi ajouté à la solution à titre de gènes non différentiellement exprimés. Un autre exemple de spike-in est le « golden-spike » de Choe et al[33] pour lequel environ 10% des gènes sont différentiellement exprimés. Même si la distribution des intensités moyennes est supposée semblable à celle d'une expérience réelle, les auteurs remarquent eux-mêmes que l'expression différentielle de ce jeu de données demeure

artificielle, étant non seulement unidirectionnelle, mais aussi limitée aux hautes intensités.

- Les s*imulations*. Il s'agit de générer des données d'expression *in silico,* la plupart du temps en estimant différents paramètres à partir de données réelles. Les simulations sont fréquemment employées pour démontrer la performance d'une nouvelle méthode[34].

Quoi qu'il en soit, un jeu de données pour lequel la « vérité » est connue n'existe tout simplement pas[31], et les alternatives souffrent de lacunes sérieuses. D'un côté, les *spike-ins* demeurent artificiels et la variabilité contre laquelle on teste une méthode est « technique » plutôt que biologique. Par exemple, le nombre de gènes différentiellement exprimés, leur intensité moyenne, la grandeur de cette expression différentielle, leur direction, écart-type, etc. ne sont pas représentatifs d'une expérience réelle. Du côté des simulations, elles reposent sur des modèles possiblement trop simples, ou pire, biaisés, en ce sens qu'ils reposent sur les mêmes hypothèses que la méthode qui doit être testée.

5. Présentation de l'outil de comparaison

Ce mémoire propose une approche nouvelle et originale à la comparaison des méthodes d'analyse : à défaut d'avoir en main plusieurs expériences pour lesquelles le véritable profil d'expression différentielle est connu, nous disposons, d'un côté, d'un *très grand nombre d'expériences réelles, et d'un autre, d'un très grand nombre de courtes listes de gènes pour lesquelles il est raisonnable d'attendre une certaine cohérence de leur niveau d'expression différentielle*. La cohérence telle que définie ici constitue un critère qui ratisse plutôt large, au sens où il importe que le niveau d'expression différentielle commun soit plus grand que pour une liste de gènes tirés au hasard. L'ensemble des listes de gènes cohérents englobe donc plusieurs catégories, dont les gènes directement co-régulés, impliqués dans une même voie moléculaire, annotés d'un même terme d'ontologie, etc. De telles listes seront aussi appelées *signatures moléculaires*. Par exemple, il est fort probable

que les transcrits des gènes membres de la voie moléculaire « Signalisation de l'interféron » (Figure 12) se voient différentiellement exprimés ensemble pour une expérience qui déclencherait une réponse immunitaire.

L'idée fondamentale est donc la suivante : les ordonnancements de gènes résultants de différentes méthodes d'analyse peuvent être vus comme des versions *bruitées* d'un ordonnancement *réel* que l'on tente d'estimer. En conséquence, la meilleure méthode d'analyse est celle qui, pour plusieurs expériences, parvient à fournir les ordonnancements les moins bruités par rapport aux ordonnancements réels. Une façon de mesurer le niveau de bruit d'un ordonnancement peut être dérivée de l'observation que l'effet attendu (au sens statistique) de l'introduction de bruit dans un ordonnancement est la *diminution* du score d'enrichissement de signatures moléculaires pertinentes à une expérience.

Pour illustrer la chose, supposons qu'une expérience consiste en la stimulation d'échantillons de sang avec l'interféron-α. Supposons aussi que l'on dispose d'une technologie de mesure de l'expression de précision et d'exactitude infinies. Assurément, la voie de signalisation de l'interféron-α se verrait fortement enrichie, ses gènes membres occupants une place élevée dans l'ordonnancement réel résultant. Comment se comparerait le score d'enrichissement d'un ordonnancement intentionnellement bruité (par exemple en permutant aléatoirement quelques gènes)? De toute évidence, le score pourrait augmenter simplement par chance, mais l'effet moyen observé, suite à plusieurs répétitions l'expérience, serait une nette diminution. Autrement formulé, donnerait-on raison à un biologiste qui choisirait une méthode d'analyse A plutôt que B, si les gènes de sa voie moléculaire préférée sont systématiquement mieux représentés? La réponse à cette question dépend du nombre d'expériences : un nombre élevé d'expériences augmente la confiance envers le biais observé. Sur un grand nombre d'expériences, une meilleure méthode reflètera systématiquement mieux la biologie connue, exprimée sous la forme de signatures moléculaires.

Ce mémoire explore l'idée d'utiliser les scores d'enrichissements de plusieurs signatures moléculaires pour plusieurs expériences comme métrique de comparaison (Figure 12). Les points suivants récapitulent la logique derrière la méthode de comparaison proposée :

1. Pour une expérience donnée, il existe un ordonnancement *réel* des gènes selon leur niveau d'expression différentielle. Cet ordonnancement existe et pourrait être retrouvé avec des instruments parfaits.

2. Le scores d'enrichissement de différentes signatures moléculaires pertinentes à une expérience donnée sont plus élevés que pour une liste de gènes aléatoire, puisqu'ils sont susceptibles d'être simultanément différentiellement exprimés (co -régulation).

3. Les erreurs de mesure, dues par exemple à la technologie elle-même, à l'amplification de l'ARN, et particulièrement la méthode d'analyse introduisent un bruit qui perturbe l'ordre *réel* pour finalement produire l'ordre observé. L'effet net du bruit sur les scores d'enrichissements des signatures moléculaire est une diminution vers la valeur attendue pour des listes aléatoires.

4. Par conséquent, différentes méthodes d'analyse peuvent être comparées sur la base de scores d'enrichissements qu'elles produisent pour différentes signatures moléculaires. Il semble raisonnable de supposer que les scores résultants de signatures non pertinentes à une expérience donnée n'offrent aucun avantage comparatif et ne font que diminuer le pouvoir de discrimination de la procédure de comparaison.

Pour terminer, il semble important de souligner que la méthode de comparaison proposée juge les méthodes sur la base de l'ordonnancement de tous les gènes de la puce plutôt que sur le contenu d'une sous-liste produite en fixant un seuil dans la liste complète des gènes. Cette approche risque de confondre certains lecteurs qui en pratique, s'attendent d'une expérience de microarrays de lui produire une *liste de gènes différentiellement*

exprimés pouvant être utilisées pour une étape subséquente de validation. Il est important de comprendre que la pratique de fixer un seuil n'est en fait qu'un cas particulier de l'utilisation complète de l'ordonnancement, où un poids unitaire est donné aux gènes classés au dessus du seuil, et un poids nul donné aux gènes sous le seuil. Dans ce contexte, il semble raisonnable de supposer qu'une méthode qui produit un meilleur ordonnancement produira de meilleures sous-listes tronquées.

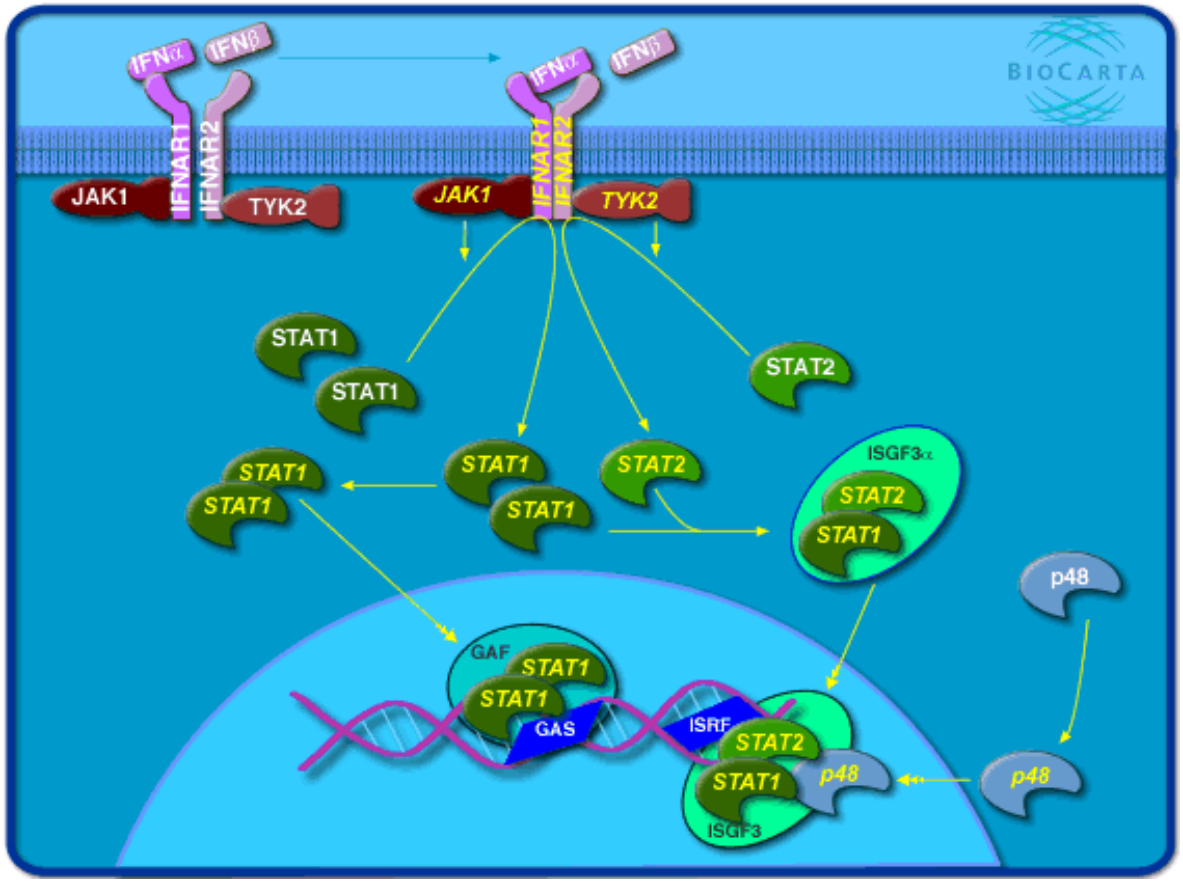


Figure 11 : Exemple de signature moléculaire : Signalisation de IFNα. Source de l'image : Biocarta[35].

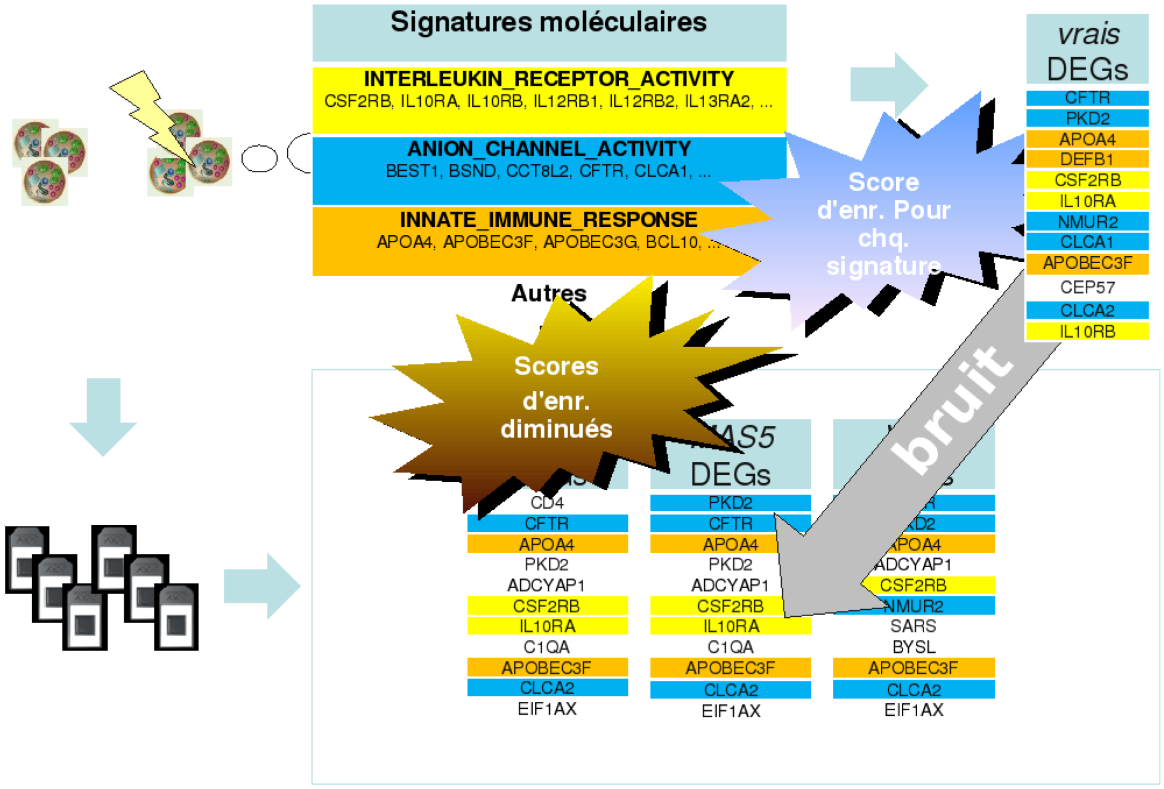


Figure 12 : Illustration du principe derrière la méthodologie de comparaison proposée. Pour une expérience donnée, les scores d'enrichissement de différentes signatures moléculaires peuvent servir de métrique de ressemblance d'un ordonnancement observé (issu d'une méthode d'analyse donnée) avec l'ordonnancement réel.

Méthodes

1. Pipeline d'analyse

Cette section passe en revue les différents algorithmes applicables aux GeneChip d'Affymetrix qui ont été soumis à la procédure de comparaison. Cette liste est loin d'être exhaustive étant donné la pléiade de méthodes proposées dans le passé et les ressources de calcul limitées, face à une combinatoire augmentant rapidement avec le nombre d'algorithmes inclus. En général, le choix d'inclure ou non telle ou telle méthode s'est justifié par : 1) La disponibilité d'une implantation dans Bioconductor, ou du moins en R, gage de popularité, 2) une vitesse d'exécution raisonnable et 3) la « valeur méthodologique » ajoutée. Un retour sera fait sur ce choix lors de la présentation des résultats.

Correction de fond

rma2, le modèle de convolution RMA

Cet algorithme a été proposé par Irizarry et al.[36] dans le cadre de la solution complète de pré-traitement *RMA*. En s'appuyant sur l'observation de la forme de la distribution des intensités brutes d'un microarray typique, ce modèle suppose que pour une puce donnée, le signal observé des sondes, S, est la somme du signal spécifique X, suivant une distribution exponentielle $exp(\alpha)$, et d'un bruit de fond Y, suivant une normale $N(\mu, \sigma)$ strictement positive. Le modèle suppose que la valeur du bruit de fond est indépendante du signal. La théorie statistique élémentaire stipule que la densité de probabilité conjointe de $S = X + Y$ est alors la *convolution* des deux distributions, ce qui donne :

$$f_{X,Y}(x, y) = \alpha \exp(-\alpha x) \frac{1}{\sigma} \phi(\frac{y-\mu}{\sigma}) \text{ avec } y > 0, x > 0$$

où ϕ est la distribution normale et Φ est la densité normale. Il est possible démontrer que la valeur attendue de *X* étant donné une valeur observée de S est:

$$E(X|S=s) \approx a + b\frac{\phi(\frac{a}{b})}{\Phi(\frac{a}{b})}$$

Où $a = s - \mu - \sigma^2\alpha$ et $b = \sigma$. Ici, μ, σ et α sont estimés directement des données par une procédure ad-hoc.

Ce modèle peut aussi se comprendre intuitivement. On suppose que le signal non spécifique suit une loi normale centrée quelque part aux basses intensités, alors que le signal spécifique suit une distribution exponentielle décroissante s'étendant vers les hautes intensités. En observant la distribution des milliers d'intensités du microarray, il est possible d'estimer assez fidèlement les paramètres de ces deux distributions. Si le signal observé est élevé, la contribution attendue du bruit est faible étant donné la faible probabilité d'obtenir des valeurs élevées de bruit, moindre que la probabilité d'obtenir un signal spécifique élevé. Aux basses intensités, la contribution du bruit devient graduellement plus importante. Sur une courbe intensité/intensité corrigée, le résultat final peut être qualitativement décrit comme un « aplatissement » lisse aux basses intensités suivi d'une droite identité. La force de cette méthode est de pouvoir déterminer quantitativement où se termine cet aplatissement.

Notons que cette méthode ne fait aucun usage de l'information spatiale des sondes, p.e. détection d'artefacts évidents sur la puce. Cette tâche est probablement reléguée au modèle niveau-sonde robuste de la méthode RMA dont elle fait partie. Notons aussi que des améliorations de la procédure ont été proposées dans la littérature[37,38].

Avant de poursuivre, il semble intéressant de faire le lien entre la correction de fond et une fausse conception entourant les intensités *négatives*. En effet, certaines méthodes de correction de fond mènent à des intensités négatives : par exemple une soustraction de l'intensité locale, ou encore un soustraction de la moyenne des sondes non spécifiques (MM ou contrôles négatifs). Les intensités négatives sont considérées problématiques par

plusieurs analystes car il est impossible de leur appliquer une transformation logarithmique. Cela ne constitue cependant pas une explication valable. La véritable raison est plutôt que les intensités négatives *n'existent tout simplement pas,* et sont plutôt le résultat d'un estimé du bruit de fond irréaliste, puisque supérieur au signal lui-même. Par exemple, si l'intensité observée pour une sonde est de 90, 100 n'est alors certainement pas la valeur réelle du bruit de fond pour cette observation. Qu'en est-il de 91? Et de 90 ? Un estimé du bruit de fond devient-il soudainement approprié si la valeur observée l'excède? En fait, la réponse appropriée à l'obtention d'intensités négatives semble être de calculer de meilleurs estimés du bruit (par exemple, un modèle statistique), et non quelque contorsion ad hoc d'arithmétique discontinue visant à permettre la transformation logarithmique.

mas, local specific correction d'Affymetrix

Cet algorithme fait référence à l'implantation R de la librairie Bioconductor *affy* de l'algorithme de correction de fond du logiciel MAS 5.0, tel que décrit par Affymetrix[39]. La puce est divisée en k régions (16 par défaut) pour chacune desquelles le bruit de fond b_k est défini comme la moyenne du 2% des intensités les plus basses (cette méthode corrige également les intensités MM). L'écart type n_k correspondant est aussi calculé. Un poids dépendant de la distance entre une position donnée de la puce et le centroïde chaque région k est défini comme :

$$w_k(x,y) = \frac{1}{d_k^2(x,y) + \text{smooth}} \ (\text{smooth}{=}100)$$

Pour chaque position (x,y) de la puce, $B(x,y)$ (ou $N(x,y)$) est défini comme la somme des b_k (n_k) de chaque région, pondérée par w_k. Finalement, si $P(x,y)$ est l'intensité observée de la sonde à la position (x,y), sa valeur corrigée est calculée comme:

$$S(x,y) = max(P_{x,y} - B(x,y), N_f * N(x,y)) \ (N_f = 0.5)$$

La correction apportée par cette méthode est un raffinement (plutôt ad hoc faut-il avouer) d'une méthode plus naïve qui ne ferait que soustraire pour chaque sonde une valeur

de bruit de fond estimé à partir des intensités les plus basses de toute la puce. Plutôt, le bruit de fond est évalué pour une région par un lissage des bruits des autres régions. Cet algorithme tente en quelque sorte de soustraire un bruit de fond dont l'intensité peut varier spatialement sur la puce. L'étape du max en fin d'algorithme permet d'éviter les valeurs négatives en les remplaçant par ce qui semble être un estimé de la variabilité du bruit de fond.

Normalisation

constant, la régression linéaire

Cette méthode de normalisation naïve suppose que l'intensité d'une sonde (ou d'un probeset) ne diffère que d'un facteur multiplicatif d'une puce à une autre, et que ce facteur multiplicatif est le même pour les différentes sondes. Par conséquent, il ne suffit que d'effectuer une régression linéaire (sans ordonnée à l'origine) entre chaque puce et une puce référence afin de trouver les facteurs multiplicatifs. Essentiellement, cela revient, pour chaque puce i, à diviser par sa moyenne le vecteur des intensités des sondes puis de le multiplier par la moyenne du vecteur des intensités de la puce référence.

La supposition de linéarité est cependant quelque peu naïve et des alternatives non linéaires comme celle de Schadt et al.[40] permettent en quelque sorte un facteur multiplicatif variable avec l'intensité. Une autre façon d'améliorer les méthodes basées sur la régression est de ne considérer que certaines sondes lors de cette dernière, par exemple des sondes dont l'intensité est stable (au sens du rang) d'une puce à une autre. À cet égard, la méthode dChip de Li et Wong[41] est assez populaire.

loess, la régression non-linéaire cyclique

La normalisation loess (*locally weighted scatterplot smoothing*) cyclique, décrite par Bolstad et al.[14] est semblable aux méthodes de régression, mais diffère sur trois aspects. Premièrement, la régression s'effectue sur les valeurs appelées *MA (MeanAverage)*, qui sont essentiellement le résultat d'une transformation logarithmique suivie d'une rotation de 45

degrés. Pour la sonde k et pour une régression entre les puces i, $M_k = log_2(x_{ki}/x_{kj})$ j et $A_k = log_2(x_{ki}x_{kj})$. Dans cet espace, $M_k = 0$ implique donc des valeurs égales pour la sonde k sur les deux puces. Deuxièmement, la régression loess est locale selon la procédure loess[42] ce qui signifie que la courbe de régression n'est pas contrainte de suivre une forme fonctionnelle particulière. La valeur de la courbe de régression peut ensuite être utilisée pour corriger les intensités dans l'espace original. Si M'_k est la valeur de la régression à l'intensité de la sonde k, $M'_k = M_k - \hat{M}_k$ est la correction locale à apporter (rappelons que $M = 0$ implique l'égalité et que dans l'espace log, une soustraction devient une division). Un peu d'algèbre indiquerait que les valeurs originales ainsi corrigées sont $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ et $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$. Finalement, cette méthode de normalisation est cyclique car plutôt que d'utiliser une des puces comme référence, l'ajustement total provient de la contribution de chacune des autres puces (probablement la moyenne des valeurs ajustées). L'opération est répétée jusqu'à convergence.

Pour résumer, loess calcule un facteur multiplicatif local à l'intensité, et ce pour toutes les paires de puces. La correction moyenne est appliquée et l'opération est itérée.

quantiles

Cette méthode normalise les intensités des sondes des puces en forçant l'égalité de tous les quantiles des intensités des sondes entre les puces. Ainsi, la plus petite intensité aura la même valeur sur toutes les puces, la seconde plus petite de même, la troisième, et ainsi de suite. La procédure d'égalisation des quantiles employée est ici la moyenne : pour chaque puce séparément, on trie les sondes en ordre croissant d'intensité, pour ensuite remplacer l'intensité de la sonde au rang n par la moyenne des sondes de rang n à travers toutes les puces. D'une certaine façon, il s'agit d'une version extrême d'un centrage de la moyenne, qui ne forcerait que l'égalité du 50-ème percentile, ou encore d'un compromis à une méthode entièrement non paramétrique qui remplacerait les intensités par leur rang. Une discussion plus élaborée de cette normalisation se retrouve dans Bolstad et al. 2003[14].

Correction PM

mas

Les versions initiales de l'algorithme de correction PM soustrayaient simplement la valeur de la sonde MM à celle de la sonde PM, ce qui mène pour certaines sondes à des valeurs négatives. Son successeur, la procédure *Ideal Mismatch,* cherche à corriger le problème des valeurs négatives en calculant une valeur ajustée de MM, nommée IM. Soit $PM_{i,j}$ et $MM_{i.j}$ les intensités PM et MM pour la sonde j du probeset i. Si $MM_{i,j} \leq PM_{i,j}$, alors MM est considéré comme un bon estimé du signal non spécifique et $IM_{i,j} = MM_{i,j}$. Dans le cas contraire, Affymetrix propose de se baser sur les rapports *PM/MM des autres sondes du même probeset* pour estimer une valeur raisonnable du signal non spécifique IM : $IM_{i,j} = \frac{PM_{i,j}}{2^{SB_i}}$ où $SB_i = T_{bi}(log_2(PM_{i,j}) - log_2(MM_{i,j}))$, soit une moyenne robuste des rapports (écarts en échelle log). Affymetrix ajoute une condition supplémentaire aux valeurs de SB_i qui ne sera pas discutée ici. Le lecteur intéressé à toute cette procédure plutôt ad hoc peut se référer à Affymetrix[39].

pmonly

Comme le nom l'indique, cette méthode correspond à n'effectuer aucune correction. Les valeurs brutes des sondes PM sont employées directement.

Sommarisation

avgdiff

Cette méthode de sommarisation ne consiste qu'à calculer la moyenne des sondes PM pour chaque probeset de chaque puce. Noter que la version employée ici n'effectue pas de transformation logarithmique préalable.

medianpolish

On dit de de la méthode mediapolish qu'elle est « multipuces », car la sommarisation des intensités des sondes d'une puce donnée est informée de l'intensité des mêmes sondes sur les autres puces du même jeu de données. Spécifiquement, un modèle linéaire additif robuste est ajusté aux intensités log2 d'un même probeset. Pour la sonde i sur la puce j,

$$\mathrm{P}M_{ij} = \alpha_i + \beta_j + \epsilon_{ij}$$

où α_i est l'effet de la sonde$i$ et β_j la valeur d'expression de la puce j, quantité que l'on cherche à estimer[36]. Noter qu'il s'agit ici plutôt d'un modèle multiplicatif qui équivaut à un modèle additif en échelle logarithmique.

Une première force de ce modèle vient du fait qu'il reconnait que d'une puce à une autre, un effet de sonde existe (la cinétique de l'hybridation varie évidemment avec la séquence), effet que l'on peut raisonnablement supposer identique d'une puce à une autre puisque la séquence sondée est la même. La seconde force vient de l'estimation robuste des coefficients par la méthode d'où elle tire son nom, *medianpolish*[43], qui la rend moins sensible aux artefacts spatiaux, qui la plupart du temps n'affectent pas simultanément plusieurs sondes d'un même probeset[12].

FARMS – FARMS I/NI

La sommarisation FARMS (*Factor Analysis for Robust Microarray Summarization*) est basée sur le modèle suivant pour les intensités PM :

$$log(PM_{ij}) = z_i(\sigma + \tau_j) + \mu + \gamma_j + \epsilon_{ij}$$

où z_i représente le score-z de la quantité réelle d'ARN dans l'échantillon hybridé sur la puce i alors que cette quantité réelle est même de moyenne μ et variance σ. S'additionnent l'effet de la sonde j γ_j sur la moyenne et τ_j sur la variance. Une procédure complexe d'analyse factorielle optimise ensuite les paramètres du modèle en usant d'une

méthode a posteriori de maximum Bayésien, sous l'hypothèse d'erreur de mesure gaussienne[44]:

$$x = \lambda z + \epsilon$$

Ce modèle suppose qu'un facteur caché $z$est sous-jacent aux intensités des sondes x dont la matrice λ décrit la structure de corrélation.

L'algorithme FARMS peut être augmentée de la procédure de filtrage I/NI (*informative*/*non-informative*)[45] basée sur la quantité $var(z|x)$, mesure de la variance des sondes x expliquée par le facteur z. En effet, les intensités des sondes d'un probeset non informatif (qui n'est pas porteur de signal) ne devraient pas être corrélées, ce qui peut être détecté par l'analyse factorielle. En résumé, la procédure I/NI est un filtre non spécifique (c.-à-d non informé du design expérimental) qui exploite l'architecture en probeset des GeneChip d'Affymetrix. Les auteurs ont par ailleurs démontré que cette méthode fonctionne généralement bien si le nombre total de puces de l'expérience analysée est de six ou plus.

Solutions complètes de pré-traitement

Par solution complète, on entend un assemblage d'algorithmes proposés dans la littérature.

MAS5, l'algorithme d'Affymetrix

Il s'agit d'une implémentation R qui produit des valeurs d'expression telles que générées par le logiciel d'analyse d'Affymetrix (*Microarray Analysis Suite 5.0*). La correction de fond locale spécifique *mas* est d'abord appliquée, pour ensuite sommariser les intensités des sondes (corrigées en soustrayant les valeurs IM) par une moyenne robuste (Tukey biweight) en échelle logarithmique. Affymetrix propose de normaliser chaque probeset en employant une version robuste de la régression linéaire :

$$ReportedValue(i) = 500\frac{2^{SLV_i}}{TrimMean(2^{SLV_i}}), 0.02, 0.98)$$

où SLV_i est l'intensité sommarisée du probeset i sur une puce donnée. Il s'agit donc essentiellement de diviser chaque probeset d'une puce par la moyenne tronquée des 2 centiles extrêmes de tous les probesets de la même puce. Noter que l'algorithme MAS5 ne retourne pas des valeurs en échelle logarithmique, et qu'Affymetrix recommande depuis une méthode plus avancée, PLIER[46].

RMA

La très populaire méthode *Robust Multiarray Average*[36] combine la correction de fond *rma*, la normalisation *quantiles* et la sommarisation *medianpolish*.

GCRMA

GCRMA se distingue de RMA par l'emploi d'une correction de fond tenant compte du fait que l'affinité entre la sonde et la cible, autant que l'affinité entre la sonde et des séquences non spécifiques peuvent varier selon la séquence[47]. En effet, il a déjà été mentionné comment Affymetrix a initialement introduit les sondes MM dans le but d'estimer le signal non spécifique. Les auteurs de GCRMA soulignent que d'abandonner les sondes MM comme le fait RMA (*avgdiff*) revient à renoncer à un certain gain en exactitude (la justesse de la valeur attendue pour l'estimé du signal spécifique) au profit d'un grand gain de précision (la variance de l'estimé). GCRMA tente de pallier à ce problème grâce à un estimateur du signal dérivé d'un modèle statistique additif dans lequel le terme pour le signal spécifique est dépendant de la séquence de la sonde.

VSNRMA, *vsn* comme alternative au $\log_2$

L'une des principales raisons d'effectuer une transformation logarithmique est qu'il s'agit d'une transformation *stabilisatrice de variance*. Une telle transformation vise à fournir des valeurs qui respectent le mieux possible les exigences des tests statistiques classiques (par exemple le test-t, ANOVA, modèle linéaire, etc.) qui suivent le pré-traitement : la normalité et l'homogénéité de la variance (hétéroscédasticité)[48]. Plus spécifiquement, une transformation stabilisatrice de variance cherche à rendre la variance

indépendante de l'intensité moyenne. Il existe plusieurs transformations candidates, telles que la transformation logarithmiques[49] ou la transformation Box-Cox[50].

Durbin et al.[51] ont toutefois remarqué que la transformation logarithmique n'apporte pas nécessairement l'effet stabilisateur recherché sur les données de microarrays, particulièrement aux basses intensités où cette dernière exagère la variance et mène à des fold-change dont la significativité n'est plus comparable à ceux des hautes inentsités[48]. En fait la transformation logarithmique est stabilisatrice seulement si la variance augmente linéairement avec la moyenne, ce qui pour les microarrays n'est pas nécessairement le cas aux basses intensités[52].

L'expression d'un gène quelconque sur une puce peut être modélisée de la façon suivante, tel que proposé par Rocke et Durbin[53] :

$$Y = \alpha + \beta e^{\eta} + \nu$$

Ce modèle implique que pour une valeur d'expression réelle β d'un gène donné, la valeur observée sur un microarray est une variable aléatoire égale à la somme d'un niveau de base $\alpha + \nu$ avec $\nu \sim N(0, s_{\nu}^2)$ et β multiplié par e^{η} où $\eta \sim N(0, s_{\eta}^2)$. Or il se trouve que ce modèle d'erreur implique une relation approximativement quadratique, et non linéaire, entre la valeur attendue de l'expression Y et sa variance. Cette déviation de la linéarité aux basses intensités est également observée en pratique.

Pour remédier à cette situation, les auteurs de *vsn* dérivent du modèle de Rocke et Durbin la transformation stabilisatrice suivante :

$$h_{ij} = glog_2 \frac{I_{ij} - b_j}{k_j} + \epsilon_{ij}$$

où I_{ij} est l'intensité observée, b_j un estimé du bruit de fond, k_j un facteur de normalisation et $glog_2$ fait référence au logarithme généralisé soit $glog_2(x) = \frac{x + \sqrt{x^2+1}}{2}$. La procédure d'estimation des paramètres ne sera pas discutée ici, mais notons qu'elle suppose que la plupart des gènes ne sont pas différentiellement exprimés, ce qui en soi peut

être une faiblesse. Quoi qu'il en soit, le modèle de *vsn* sous-entend donc qu'il existe un bruit de fond additif ainsi qu'un facteur multiplicatif qui affecte l'intensité de chaque sonde. Sous ce modèle, les auteurs démontrent que la transformation logarithmique généralisée rend la variance approximativement indépendante de l'intensité. L'inclusion du facteur multiplicatif *k* et du bruit additif *b* dans la transformation rend implicite la correction de fond et la normalisation à la procédure de stabilisation de la variance.

La méthode VSNRMA qui a été employée pour le présent projet combine la normalisation *vsn* (*variance stabilization normalization*[52]) suivie de la sommarisation de RMA.

Statistiques d'ordonnancement

FC, le *fold-change*

Dans le cadre du présent projet, FC fait référence à la différence entre les moyennes respectives pour deux groupes comparés. Soit y_{ij}, l'expression du gène j pour la puce i , pour les deux groupes d'échantillons biologiques G_1, G_2 comparés,

$$FC_j = \bar{y}_{i \in G_1, j} - \bar{y}_{i \in G_2, j}$$

Si les valeurs d'expression sont en échelle logarithmique, la différence des moyennes arithmétiques est équivalente au ratio des moyennes *géométriques* des valeurs d'expression non logarithmées, d'où l'appellation *fold-change* en anglais.

Modèles linéaires

De toute évidence, les expériences de microarrays ne sont pas toujours aussi simples qu'une comparaison de deux groupes. En langage statistique, on dirait que les designs expérimentaux n'impliquent pas nécessairement qu'un seul covarié d'intérêt, et que ces covariés ne sont pas nécessairement de type « catégoriques » (covarié binaire indiquant l'appartenance à une classe). Plus généralement, le fold-change est donc défini comme la

valeur de l'estimé d'un coefficient d'un modèle linéaire multivarié[54]. Cette section présente les faits saillants de la théorie des modèles linéaires, qui se formule élégamment et en toute généralité à partir d'opérations simples d'algèbre linéaire.

Comme le même modèle linéaire multivarié est ajusté à chaque gène, on peut utiliser la formulation matricielle suivante[55] :

$$E(\mathbf{y_j}) = \mathbf{X}\alpha_\mathbf{j}$$

$$\mathbf{var}(\mathbf{y_j}) = \sigma_\mathbf{j}^\mathbf{2}$$

Les termes de ces équations s'expliquent aisément. α est le vecteur des coefficients du modèle linéaire. Il y a autant de covariés que de coefficients, ces derniers spécifiant la relation linéaire avec la variable dépendante y. Chacune des colonnes de X, la *matrice de design*, spécifie les valeurs d'un covarié pour les différentes observations (puces) sur la variable dépendante (expression du gène j). Un modèle linéaire formulé correctement est caractérisé par une matrice de design dont les colonnes sont linéairement indépendantes. Par exemple, supposons qu'une expérience comporte trois groupes biologiques quelconques. Un modèle pourrait être ajusté tel que les coefficients $\alpha_\mathbf{j}$ représentent la moyenne de chaque groupe pour le gène j. La matrice de design appropriée en serait une ou chaque ligne i « encode » l'appartenance au groupe biologique de la i-ème observation par un «1». La seconde équation définit la variance totale (réelle) σ_j^2 des valeurs d'expression du gène j.

L'*ajustement* d'un modèle linéaire consiste à calculer un estimé des coefficients et de leurs variances. Le plus souvent, la méthode des moindres carrés (grâce au pseudo-inverse de Moore-Penrose) est utilisée pour estimer les coefficients du modèle:

$$\hat{\alpha}_\mathbf{j} = [(\mathbf{X^T X})^{-1}\mathbf{X^T}]\mathbf{y_j}$$

On suppose que variance (réelle) du m-ème coefficient suit l'expression :

$$\mathbf{var}((\hat{\alpha}_\mathbf{j})_\mathbf{m}) = \sigma_\mathbf{j}^\mathbf{2}((\mathbf{X^T X})^{-\mathbf{1}})_\mathbf{mm}$$

En remplaçant σ_j^2 par un estimé il est ensuite possible d'effectuer ensuite différents tests statistiques sur ces coefficients.

Souvent, il sera intéressant de travailler sur des *combinaisons linéaires* des coefficients, appelées *contrastes*. Ces combinaisons linéaires sont spécifiées par les lignes de la matrice des contrastes $\mathbf{C}$, tel que :

$$\beta_\mathbf{j} = \mathbf{C}\alpha_\mathbf{j}$$

la variance des contrastes est donnée par l'expression :

$$\mathbf{var}((\hat{\beta}_\mathbf{j})_\mathbf{m}) = \sigma_\mathbf{j}^\mathbf{2}(\mathbf{C^T}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{C})_\mathbf{mm}$$

(mm réfère au m-ème élément diagonal de la matrice). La formulation des contrastes est souvent nécessaire, car les coefficients α du modèle initial ne répondent pas directement à la question biologique d'intérêt. Cela reflète le fait que plusieurs formulations équivalentes du même modèle linéaire sont possibles, pour autant que les colonnes de la matrice de design soient linéairement indépendantes.

lm, *nolm* et le test-t de Student

Souvent, une expérience comportera plusieurs groupes biologiques pour lesquels certaines différences sont d'intérêt. Par exemple, une expérience pourrait tester l'effet de deux nouveaux traitements, notés T1,T2, par rapport à un traitement contrôle, noté T0 (Figure 13).

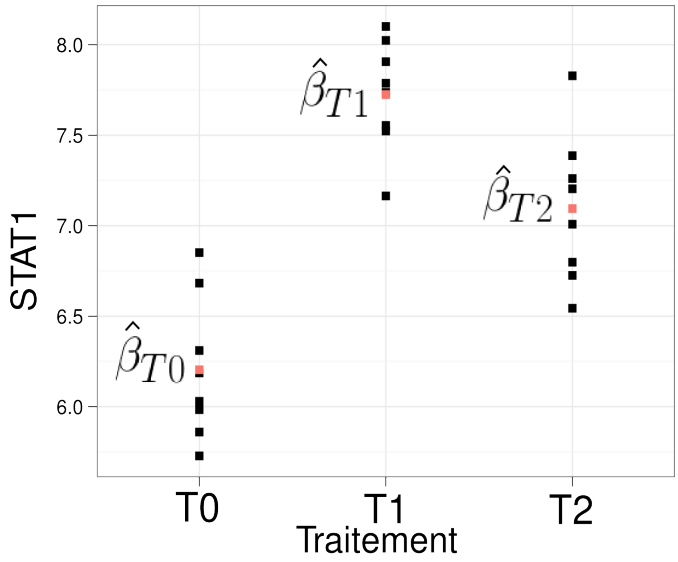


Figure 13 : Modèle linéaire et comparaisons groupe à groupe. Les données ont été simulées ici telles que $\beta_{T0} = 6.5, \beta_{T1} = 7.5, \beta_{T2} = 7$ et un écart-type des résiduels $\sigma = 1$. Supposons que l'on veuille tester l'hypothèse nulle habituelle pour la différence (contraste) $T1 - T0$. Le test-t à deux groupes indique une valeur $p = 5.55 \cdot 10^{-7}$. En comparaison, on obtient $p = 6.29 \cdot 10^{-8}$ pour un test-t sur le contraste $\beta_{T1} - \beta_{T0}$ du modèle linéaire $STAT1 \sim \beta_{T0}x_{T0} + \beta_{T1}x_{T1} + \beta_{T2}x_{T2}$. Sur plusieurs simulations, la tendance est d'obtenir des valeurs-p inférieures pour l'approche par modèle linéaire (non illustré) : l'estimé $\hat{\sigma}$ est différent entre les deux tests puisque amélioré par l'inclusion des observations de T2.

Une première approche possible (*nolm*), consiste à ne considérer que les puces impliquées dans la comparaison d'intérêt pour le test statistique. Cette approche est fréquemment utilisée en pratique car elle ramène typiquement le problème à des tests statistiques simples, comme par exemple le test-t entre les deux groupes, bien compris par la plupart des expérimentateurs :

$$t_{nolm} = \frac{FC_j}{\sqrt{s_j^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

où $s_j^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, s_1, s_2 étant les variances échantillonnages individuelles des groupes comparés. Cette statistique peut être utilisée pour effectuer le test d'hypothèse habituel :

$$H_0 : FC = 0, H_1 : FC \neq 0$$

La probabilité d'observer une valeur donnée de la statistique t plus grande que celle observée sous l'hypothèse nulle H_0 (valeur-p) peut alors être calculée à partir de la distribution-t de Student à $n_1 + n_2 - 1$ degrés de liberté.

La seconde approche (*lm*), consiste à ajuster un modèle linéaire à toutes les puces d'une expérience, pour ensuite faire les tests sur les comparaisons de groupes voulues. L'approche *lm* devrait être préférée à l'approche naïve *nolm,* non seulement car elle permet de tester directement les «effets» voulus sans se limiter bêtement aux comparaisons de deux groupes, mais aussi car la variance échantillonnale est alors estimée en incluant toutes les puces de l'expérience (~nombre de degrés de liberté des résidus plus grand, donc tests statistiques plus puissants). Dans le contexte des modèles linéaire, la statistique-t pour le m-ème contraste du modèle se formule en toute généralité :

$$t_{lm} = \frac{(\hat{\beta}_j)_m}{\sqrt{\mathbf{var}((\hat{\beta}_\mathbf{j})_\mathbf{m})}} = \frac{(\hat{\beta})_m}{s_j \sqrt{(C^T(X^TX)^{-1}C)_{mm}}}$$

où s_j, l'écart-type des résidus, estime σ_j. Sous l'hypothèse de normalité qui ne sera pas discutée ici, t_{lm} suit une distribution-t de Student à d_g degrés de liberté, soit le nombre de coefficient du modèle moins 1.

Dans le cadre du présent projet, les versions *lm* et *nolm* sont comparées, si applicable. De plus, le modèle linéaire ajusté aux données sera toujours tel que les coefficients α correspondent aux moyennes des groupes biologiques et que les contrastes β correspondent à toutes les différences (fold-change) groupe à groupe possible.

Cyber-T, une statistique-t *régularisée*

L'équation pour t_{lm} à la section précédente montre que la statistique-t ordinaire repose sur un estimé de la variance, et que cet estimé est la variance échantillonnale du gène même. De nombreux auteurs ont reconnu que l'utilisation de la variance échantillonnale résulte en des statistiques instables dans le contexte des microarrays, où le nombre de tests effectués est immense et le niveau de réplication très faible. On affirme par

exemple que des gènes au fold-change négligeables peuvent résulter en des statistiques-t exagérées si leur variance échantillonnale est petite simplement par hasard[56].

Pour pallier ce problème, de nombreux auteurs ont proposé leur version d'une statistique-t régularisée, statistique pour laquelle l'estimé de la variance d'un gène donné est amélioré en « empruntant » de l'information aux autres gènes. La première statistique de ce genre qui est comparée pour ce projet est Cyber-T[57] :

$$CyberT = \frac{FC\sqrt{n}}{\zeta_p}$$

où ζ_p est un estimé bayésien de l'écart-type :

$$\zeta_p^2 = \frac{\nu_0\zeta_0^2+(n-1)\zeta}{\nu_0+n-2}$$

Ici, ζ est l'écart type des résidus habituel, $\zeta_0$un *prior* sur l'écart-type calculé comme la moyenne des écarts-types pour les gènes de niveau d'expression semblables, et $\nu_0$un facteur de pondération quantifiant la confiance attribuée au prior. L'effet du prior sera de comprimer les variances vers l'estimé local, le but de la régularisation.

L'implantation R fournie par l'auteur a été employée pour calculer cette statistique. Noter que cette statistique n'est implantée qu'en version *nolm*, mais qu'elle est possible si les tailles de chaque groupe sont différentes, contrairement à la formulation ci-haut.

SAM-R

La statistique-t SAM (Significance Analysis of Microarrays[58]) est définie comme :

$$SAMr = \frac{FC}{s_0+s}$$

où la valeur de s_0 , facteur « fudge » sur la variance, est la valeur de s^* qui minimise le coefficient de variation de la statistique elle-même, $\frac{FC}{s^*+s}$, sur tous les gènes. Autrement dit, on recherche la valeur de s_0 parmi toutes les valeurs possibles s^* qui sur tous les gènes

résulte en une statistique-t la moins variable possible. La librairie R *samr* a été employée pour calculer cette statistique, qui n'est possible qu'en version *nolm*.

ebayes

ebayes est une autre statistique de l'expression différentielle fortement populaire en pratique, en grande partie car elle est calculée par défaut par la librairie R *LIMMA*[55], spécialement conçue pour l'ajustement de modèles linéaires aux données de microarrays. Sans entrer dans les détails, ebayes est une méthode dite *bayes empirique*, ce qui signifie que non seulement la valeur de la variance échantillonnale est comparée à un prior (une distribution sous-jacente), mais que ce prior est évalué à partir des données mêmes. Qualitativement, l'idée des méthodes de Bayes empiriques s'articule ainsi

1. On estime d'abord la distribution des variances réelles à partir des données mêmes. Cette distribution est le *prior*. Le prior est souvent estimé en supposant que seule une fraction des gènes sont véritablement différentiellement exprimés.

2. On dérive un *posterior* de la précédente distribution, c'est à dire la densité de probabilité de la variance réelle *étant donnée* la variance échantillonnale observée. Il est alors possible de remplacer la valeur observée de la variance échantillonnale par la valeur attendue du posterior. Par exemple, on peut comprendre intuitivement qu'il est peu être plus probable qu'une valeur élevée de la variance échantillonnale provienne d'une variance réelle moyennement élevée que d'une variance réelle elle-même élevée, cette dernière étant elle-même peut probable selon le prior.

Cette façon de raisonner peut sembler étrangement circulaire, mais il semblerait qu'elle fonctionne bien en pratique, c'est-à-dire que les estimés de la variance sont en moyenne plus rapprochés de la variance réelle. La modération s'effectue en quelque sorte en « empruntant » de l'information à l'ensemble de gènes pour aider dans l'inférence sur un gène particulier. Dans le cas précis de *ebayes*, la statistique-t modérée est définie comme :

$$t_{ebayes} = \frac{\hat{\beta}_j}{\tilde{s}_j \sqrt{v_j}}$$

où $v_j = (C^T(X^TX)^{-1}C)_{mm}$ tel que vu précédemment et $\tilde{s}$ est l'estimé posterior de la variance. On suppose en fait que la véritable variance σ_j du gène j, s'il n'est pas différentiellement exprimé, est tirée de la distribution :

$$\frac{1}{\sigma_j^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

où s_0 est par conséquent la valeur attendue des variances réelles. De cette dernière équation, il est possible de démontrer que

$$\tilde{s}_j^2 = \frac{d_0 s_0^2 + d_j s_j^2}{d_0 + d_j}$$

Les quantités s_0, d_0 sont estimées par une procédure complexe qui ne sera pas discutée ici. La statistique résultante suit une distribution-t au nombre de degrés de liberté résiduels plus grand que la statistique-t ordinaire. Cette statistique peut ensuite être employée pour un test d'hypothèse habituel, typiquement celui de l'expression différentielle nulle $H_0 : \beta_j = 0$.

TREAT

TREAT[59] est une variante intéressante de la statistique modérée *ebayes*. Sa particularité est l'application d'un seuil sur la valeur du coefficient de régression testé. Plutôt que de définir l'hypothèse nulle comme $H_0 : \beta = 0$ avec $H_1 : \beta \neq 0$, la valeur du contraste est plutôt testée par rapport à un seuil : $H_0 : \beta > |\tau|, H_1 : \beta \leq \tau$. La motivation derrière ce test est de ne retenir que les gènes dont le fold-change est biologiquement « significatif ». Ce test d'hypothèse composite est implanté dans la librairie Bioconductor LIMMA.

Note sur le paradigme « stepwise »

Gentleman et al.[12] ont expliqué comment il est d'usage de diviser l'analyse des données de microarrays étapes distinctes, ce qu'ils appellent l'approche *stepwise*. L'ordre usuel dans lequel est abordée chaque étape est celui dans lequel elles ont été présentées précédemment. De plus, toutes les puces sont habituellement pré-traitées ensemble (on pourrait bien décider, par exemple, de pré-traiter individuellement les comparaisons groupe à groupe). Ce mémoire s'en tiendra à ce paradigme (à l'exception des solutions complètes VSNRMA et MAS5). L'approche « s*tepwise »* pourrait très bien cependant suivre un ordre différent. Bolstad[60] s'intéresse par exemple à deux autres versions de la normalisation quantiles : une première ajuste *ensemble* les sondes d'un même probeset et une seconde qui normalise *après* sommarisation. Rien non plus n'oblige aussi à suivre l'approche *stepwise*, diviser le problème en étapes successives ne semble être qu'un compromis technique. Par exemple, Bolstad[60] et Lemieux[61] ont proposé des modèles linéaires qui estiment à la fois les effets de sondes et de traitement, ce qui rend implicite au modèle statistique l'étape de sommarisation.

2. Métrique d'enrichissement

Il a déjà été décrit que la méthode de comparaison cherche à quantifier laquelle des méthodes d'analyses de l'expression différentielle reflète le mieux l'information de co-régulation contenue dans les signatures moléculaires (groupes de gènes). Pour cela une métrique d'enrichissement doit être calculée pour chaque paire signature/ordonnancement de gènes. Une panoplie de statistiques d'enrichissement sont à disposition à cet effet, toutes ne sont cependant pas adéquates.

Tout d'abord, étant donné le grand nombre de scores d'enrichissement à calculer, le temps de calcul est un critère important. D'emblée, cela exclut des métriques d'enrichissement basées sur la permutation des échantillons (ou des gènes), qui demandent de recalculer la statistique d'expression différentielle pour chaque permutation (normalement de 1000).

Ensuite, la métrique d'enrichissement doit être applicable à plusieurs statistiques d'expression différentielle. Par exemple, une métrique qui effectue une transformation binaire (fixer un seuil sur un niveau de significativité) comme le test exact de Fisher sur la distribution hypergéométrique n'est pas applicable au fold-change. Certaines métriques sont cependant effectivement applicables à tout un éventail de statistiques. PAGE[62] par exemple, calcule un score z sur les statistiques de chaque gène d'un groupe de gène, pour ensuite se baser sur le théorème central limite afin d'en calculer la significativité. Toutefois, une métrique faisant le moins d'assomptions possibles a semblé préférable, quitte à utiliser une métrique de moindre exactitude : en effet, puisque l'on cherche à comparer les scores provenant de différentes méthodes d'analyse de l'expression différentielle, l'important est qu'un « meilleur enrichissement » produise un meilleur score. Une statistique d'enrichissement d'exactitude moindre diminuera au pire le pouvoir de discrimination de la méthode de comparaison.

Le choix s'est arrêté sur une statistique simple, non paramétrique, donc qui ne travaille que sur les rangs des gènes, l'*AUC* (*Area under the ROC curve*). Soit O l'ordonnancement des gènes selon leur statistique d'expression différentielle (valeur absolue, pas de distinction sur/sous-expression), S le groupe de gènes pour lequel on recherche l'enrichissement de O et R la somme des rangs de de S dans O. L'AUC se calcule comme

$$AUC = 1 - \frac{U}{|S|(|O|-|S|)}$$

où U est la statistique Wilcoxon-Mann-Whitney-U définie comme

$$U = R - \frac{|S|(|S|+1)}{2}$$

La valeur attendue de l'AUC a une valeur attendue de 0.5 si les membres de S de sont parfaitement dispersés à travers O, 1 au haut et 0 au bas de l'ordonnancement.

S = {HOXB9, HOXA4, HOXB13}

O1	
Rang	Gène
1	CBX8
2	HOXB13
3	HOXB9
4	GTF3C4
5	HOXA4
6	AEBP2
7	FBS1
8	BCL2L2
9	FEZ2
10	CRSP7CS

U = 17

AUC = 0,81

O2	
Rang	Gène
1	CBX8
2	FBS1
3	BCL2L2
4	GTF3C4
5	HOXA4
6	AEBP2
7	HOXB13
8	FEZ2
9	HOXB9
10	CRSP7CS

U = 6

AUC = 0,29

Figure 14: Example fictif illustrant l'AUC.

Ainsi, l'AUC est une version normalisée de la statistique Wilcoxon-Mann-Whithney-U, soit une comparaison de la somme des rangs des membres de S avec son complément dans O. Plus le rang moyen des éléments de S dans O est élevé, plus l'AUC se rapproche de 1.

Une interprétation équivalente de l'AUC est celle de l'*aire sous une courbe ROC* si l'on considérait les éléments de S comme les « instances vraies » (gènes véritablement différentiellement exprimés) d'un classificateur binaire à seuil variable qui teste les éléments de O. C'est d'ailleurs de cette interprétation qu'est née l'idée de la méthode de comparaison. La valeur numérique de l'AUC est alors égale à la probabilité que le classificateur ordonne une instance positive choisie aléatoirement mieux qu'une instance négative choisie aussi aléatoirement.

3. Signatures moléculaires

Il a été expliqué en introduction que l'outil de comparaison proposé repose sur un grand nombre de listes de gènes susceptibles d'être simultanément différentiellement exprimés. La base de données MSigDB[63] de Subramanian et al. (version 2.5) fut identifiée comme candidat idéal à cet effet, étant donné sa taille, la fiabilité des données et sa popularité. Cette base de données, construite initialement pour effectuer des analyses d'enrichissement, est divisée en différentes *collections* (Tableau 1) selon la provenance des

listes de gènes, qui seront aussi appelées *signatures moléculaires,* en référence à la nomenclature de MsigDB.

Tableau 1: Description sommaire des collections composant MSigDB.

Collection MSigDB	Description
c1	Listes de gènes positionnelles pour chaque chromosomes humains et chaque bande cytogénétique.
c2	Listes de gènes compilées, provenant de BD de voies moléculaires, de publications dans PubMed et connaissances d'experts.
c3	Listes de gènes motifs, basées sur la conservation d'un motifs cis-régulatoire dans une analyse comparative des génomes humain, du rat et chien.
c4	Listes de gènes computationelles, extraites de l'analyse de jeux de données de microarrays liés au cancer.
c5	Listes de gènes annotés par un même terme GO.

c1 – *positional gene sets*

Cette collection rassemble 386 listes de gènes correspondants à chaque chromosome et chaque bande cytogénétique contenant au moins un gène. Les bandes cytogénétiques sont des régions chromosomiques visibles au microscope après marquage et servent de marqueurs de position le long des chromosomes. L'expression des gènes d'une même bande peuvent donc être affectés par des délétions ou amplifications chromosomiques, du « silencing » épigénétique, de la compensation de dose et autres effets « régionaux ». À titre d'exemple, il est typique pour les cellules tumorales de voir certaines régions chromosomiques contenant un oncogène être dupliquées[64]. Pour résumer, la collection c1 est susceptible de capturer la co-régulation due à la position des gènes le long du génome.

c2 – *curated gene sets*

Cette collection contient 1892 listes de gènes compilées par des experts. Elle est divisée en deux sous-collections. La première, c2:cp (*canonical pathways*, taille 639), est tirée (représentation canonique) de différentes base de données de voies moléculaires (*pathways*) métaboliques ou de signalisation telles que BioCarta[65], KEGG[66] et Signaling Gateway[67], la liste complète étant disponible sur le site web de MsigDB[68]. La seconde, c2:cgp (*chemical and genetic perturbations*, taille 1186) est le résultat d'une revue de littérature d'expériences de microarrays, dans le but d'identifier des signatures (groupes de gènes) particulières à différentes perturbations chimiques et génétiques (p.e. gènes dont l'expression diminue après un traitement de cellules GH3 avec le LIF[69]). Le problème de feedback que pourrait causer cette dernière sous-collection fut une préoccupation importante du projet et sera traité en résultats.

c3 – *motif gene sets*

Cette collection réuni 837 listes de gènes partageant un motif cis-régulatoire en région promotrice ou en 3'-UTR. 500 listes de la sous-collection c3:tft contiennent des gènes qui partagent un site de liaison tel que défini dans la base de données de facteurs de transcription TRANSFAC[70]. La sous-collection c3:mir contient des gènes dont la région 3'-UTR contient un motif de liaison à un des 222 miRNA déjà connu[71]. Le restant de c3 (105 listes), inclue dans c3:tft, consiste en des listes de gènes dont les régions promotrices partagent un des 105 motifs découverts par Xie et al.[72] comme étant fortement conservés entre les génomes humains, de la souris, du rat et du chien.

Il sera discuté plus loin que, comme dans le cas de c1, c3 fourni des listes de gènes pour lesquels la co-régulation est fortement probable en plus d'être inférée à partir de données de séquence seulement. Par conséquent, ces collections sont absolument libres de tout biais en faveur d'une méthode d'analyse ou d'une autre.

c4 – *computational gene sets*

Cette collection est composée de 883 listes de gènes provenant de l'analyse de grandes expériences de microarrays sur la thématique du cancer. Elle se divise en deux sous-collections qu'il est nécessaire de décrire plus en détail, afin de disperser certains doutes qui deviendront plus clairs ultérieurement.

La sous-collection c4:cm (*cancer modules*, taille 456) a été construite par Segal et al.[73]. Les modules en question sont essentiellement des groupes de gènes fortement enrichis dans un compendium d'expériences de microarrays reliées au cancer. Sommairement, ils ont été générés de la façon suivante :

1. Collecter 1 975 microarrays d'expériences sur une grande variété de cancer. Les microarrays sont tirés de la *Stanford Microarray Database*[74] et du *Center for Genomic Research at the Whitehead Institute* (voir pour plus de détails). Identifier pour chacune les gènes différentiellement exprimés (normalisation par centrage de chaque gène, suivi de $FC_{ratio}>2$).

2. Collecter 2 849 listes de gènes initiales, provenant de la GeneOntology[75], KEGG[66], GenMAPP[76] et des listes de gènes fortement exprimés spécifiquement dans certains tissus[77]. En plus, d'autres listes sont générées à partir des gènes dont l'expression est fortement corrélée dans les expériences de l'étape 1.

3. Trouver pour quelles expériences de microarrays les listes de gènes de l'étape 2 sont significativement enrichies (test hypergéométrique, FDR 5%).

4. Effectuer un clustering hiérarchique[78] (corrélation de Pearson comme métrique de distance) des listes de gènes dans l'espace de la significativité de l'enrichissement pour chaque expérience (0 ou 1). Définir une partition si la distance est supérieure à 0,05. S'en suit un test statistique afin d'épurer les groupes de gènes (union des listes initiales composant un noeud) qui ne sera pas décrit ici. Finalement, toute cette

procédure a pour effet de fusionner *en modules* les listes de gènes initiales selon leur profil d'enrichissement à travers les expériences de microarrays.

La sous-collection c4:cgn (*cancer gene neighborhoods*, taille 427) prend une approche différente afin d'identifier des signatures moléculaires de cancers. Partant d'une liste initiale de 380 gènes reliés au cancer compilée par Brentani et al.[79], le « voisinage » de chacun de ces gènes est défini comme l'ensemble des gènes dont les profils d'expression dans quatre grands jeux de donnée sont fortement corrélés (Pearson > 0.85, taille minimale 25 gènes) avec le gène initial. Ces quatre jeux de données sont : le *Novartis normal tissue compendium*[80], le *Novartis carcinoma compendium*[81], le *Global cancer map*[82] et finalement une série d'expériences « maison » du Broad Institute.

c5 – Gene Ontology (GO)

Par définition, une ontologie spécifie un vocabulaire contrôlé et les relations entre les termes de ce vocabulaire. Dans le contexte de la biologie moléculaire, elles fournissent une représentation systématique des connaissances à propos des gènes[83,5].

Tout comme il est important de spécifier et de contrôler les termes du vocabulaire et leurs relations dans le système de classification d'une bibliothèque, il y va de même pour les centaines de milliers de gènes des systèmes biologiques. Le consortium GO[75] (*Gene Ontology*) offre le standard international à cet effet, et facilite donc grandement l'interprétation de grandes listes de gènes produites par les méthodes à haut débit, dont les microarrays font évidemment partie.

GO offre trois ontologies pour les gènes et leurs produits, soit *molecular functions*, *biological processes*, et *cellular locations*. La figure 15 montre un exemple des termes des trois différentes ontologies annotant le gène FOXP2 et de leur structure. On y voit que les termes sont des nœuds d'un graphe acyclique orienté (DAG) et que si un gène est annoté par un certain terme, il sera aussi annoté par tout les termes parents dans le graphe. Cet aspect est souligné ici, car cela implique que lorsque l'on tente de faire passer les annotations GO au paradigme de listes (de créer une liste par terme pour ensuite remplir ces

listes des gènes annotés par ce terme), les listes résultantes ne sont évidemment pas indépendantes et seront redondantes étant donnée la structure en arborescence des termes GO. Par exemple, les listes résultant des termes *putamen development* et *anatomical structure* hériteront toutes deux de FOXP2 comme membre. Cette possible « dépendance » statistique entre les listes de gènes de c5 (tout comme les dépendances possibles des autres collections et les dépendances inter-collection comme entre c2 et c5) a toutefois été ignorée dans l'implantation l'outil de comparaison proposée. Les conséquences possibles et des pistes de solution seront traitées en discussion.

Les sous-collections c5:cc, c5:bp et c5:mf (tailles 233, 825, 396) sont construites en regroupant ensemble les gènes annotés d'un même terme GO, respectivement des ontologies c*ellular component*, *biological process* et *molecular function*. Toutefois, seules les annotations portant les codes d'évidences[a] suivants sont inclues : IDA, IPI, IMP IGI, IEP ISS et TAS ont été considérées. De plus, les listes résultantes dont la taille finale est inférieure à 10 ou celles qui sont entièrement redondantes ont été omises.

a Code qui indique le type de preuve scientifique qui justifie l'annotation, voir pour la liste complète voir Genome Group of the Gene Ontology Consortium[75].

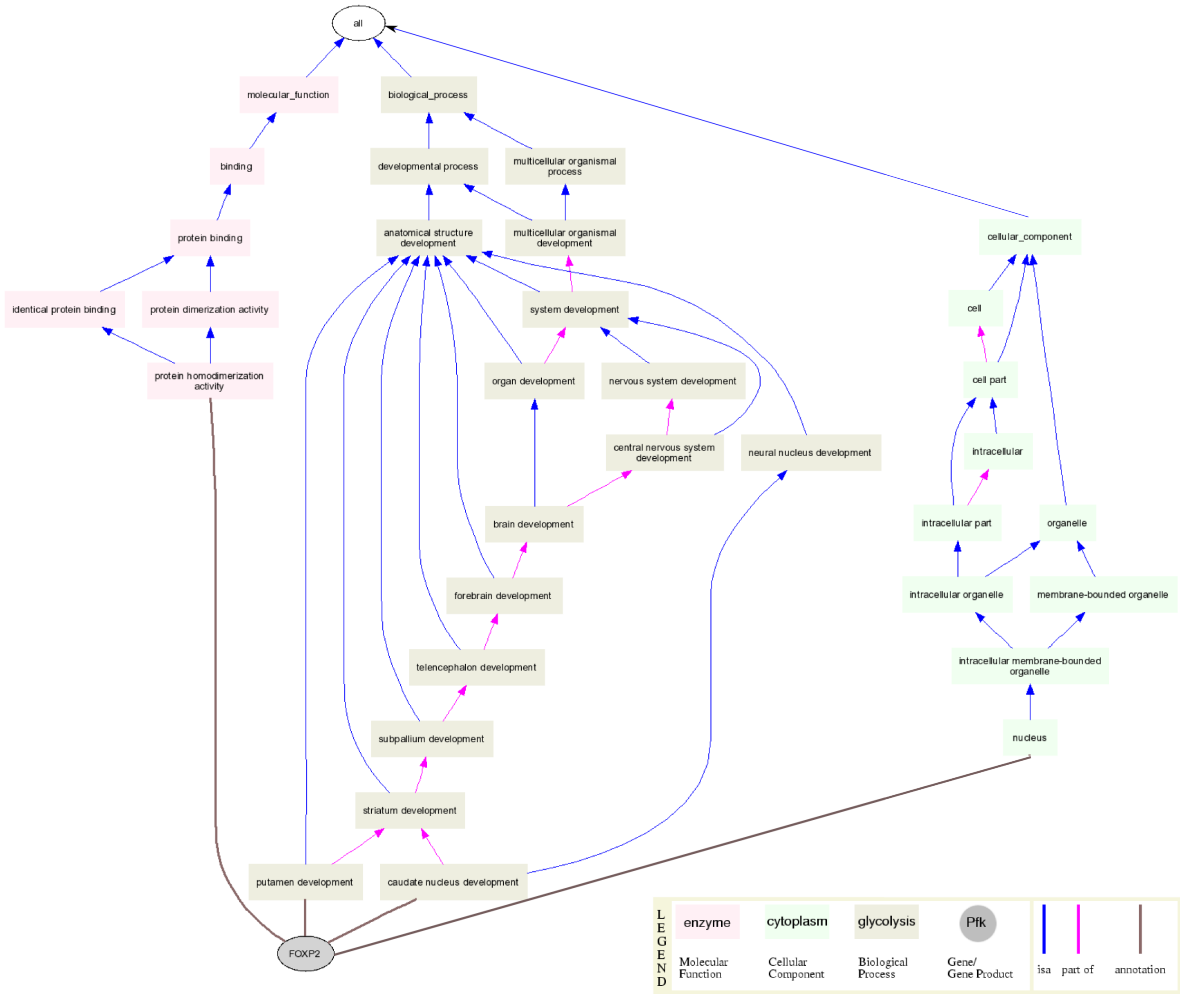


Figure 15: Exemple des termes des trois ontologies de GO pour le gène FOXP2. Cet exemple illustre comment les connaissances des fonctions, des processus biologiques, et de la localisation cellulaire d'un gène sont formalisées dans GO. Figure éditée de *GenNav*, http://mor.nlm.nih.gov/perl/gennav.pl

Résultats

Tel qu'il fut discuté en introduction, la méthodologie de comparaison proposée repose sur l'usage d'un grand nombre d'expériences de microarrays et de signatures moléculaires. Je commence donc ici par une description des résultats de collecte et de curation de ces données.

1. Expériences de microarrays

Collecte et curation

Dans un premier temps, il fut entrepris de retrouver le plus grand nombre d'expériences de microarrays couvrants un grand nombre de thématiques biologiques, de façon à exploiter le plus possible l'information biologique représentée par les signatures moléculaires. Il est présumé que plus d'expériences sont incluses, plus la capacité de résoudre les différences de performances entre différentes méthodes d'analyse sera grande. Il fut préalablement établi que les expériences recueillies doivent remplir les critères suivants :

1. *Les données (intensités) brutes doivent être disponibles.* Sans ces dernières, il est évidemment impossible de réanalyser les expériences avec différentes méthodes aux fins de comparaisons.

2. Si des expériences provenant de plateformes différentes sont incluses, ces p*lateformes doivent préférablement avoir des propriétés statistiques semblables*. Il est difficile de définir précisément ce dernier concept, mais il est toutefois certain que de nombreux facteurs tels que le nombre de sondes, leur longueur, les protocoles d'hybridation aient une influence.

3. *Les propriétés statistiques et les thématiques des expériences doivent être variées.* La raison pour exiger la diversité des thématiques biologiques a déjà été

mentionnée. Quant aux « propriétés statistiques », même si la comparaison des méthodes d'analyse se fera sur une base par expérience, l'ensemble des expériences ne doit pas être biaisé envers un régime statistique particulier afin que les conclusions sur la meilleure méthode soient générales. Par exemple, une méthode A pourrait mieux performer sur une expérience « facile » où l'expression différentielle est très évidente (par exemple, neurone vs tissu musculaire), alors qu'une méthode B pourrait mieux performer sur une expérience où seulement quelques gènes sont différentiellement exprimés. Encore ici, ce critère apparaît difficile à quantifier et il apparaît raisonnable de se contenter d'expériences dont les thématiques biologiques semblent variées.

4. Pour une même plateforme, le nombre d'expériences disponibles doit être le plus grand possible, idéalement plus d'une vingtaine. Cela se justifie par le fait que les méthodes d'analyse seront comparées sur une base par expérience.

5. De plus, les plateformes se doivent d'être méthodologiquement « compatibles » au sens où tous les algorithmes à l'étude doivent être applicables à toutes les plateformes. Par exemple, les plateformes Affymetrix ne sont pas compatibles avec la technologie deux-couleurs, ces dernières ne reposant pas sur l'architecture en probesets qui requiert alors l'étape de sommarisation.

6. Les contrastes offerts par une expérience doivent avoir une signification biologique. La raison est que l'expression différentielle doit résulter d'une perturbation quelconque du réseau de régulation sous-jacent, car, tel qu'exposé en introduction, on souhaite comparer les méthodes en se basant sur leur capacité à faire « ressortir la biologie » telle qu'exprimée par les signatures moléculaires.

7. L'expérience doit être minimalement répliquée. Cela signifie qu'à chaque condition biologique doivent correspondre au moins deux microarrays. Sans réplication, l'estimation de la variance d'un gène individuel est impossible, rendant inapplicable

toute méthode basée sur l'ajustement d'un modèle statistique, par exemple, le test-t de Student.

8. La taille du jeu de donnée doit être raisonnable pour que les données brutes puissent être chargées en mémoire et traitées par toutes les méthodes d'analyse. Nous disposions initialement de 4 Go de mémoire vive.

Une recherche par accès programmatique dans de la base de données *Data Sets* du *Gene Expression Omnibus* a révélé que les plateformes Affymetrix *HG-U133A* (GPL96) et *HG-U133 Plus 2* (GPL570) offrent le plus d'expériences pour lesquelles les données brutes sont disponibles. Comme ces deux plateformes furent les seules trouvées répondant aussi aux critères 2 à 5, il fut décidé de ne se concentrer que sur ces dernières.

La librairie Bioconductor *GEOquery*[84] fut employée pour retrouver les fichiers CEL et le design expérimental des expériences du site ftp de GEO. Malgré les efforts de curation de l'équipe de GEO, il fut nécessaire d'inspecter chaque expérience afin de non seulement s'assurer du respect des critères établis, mais aussi d'apporter des corrections aux designs des expériences retenues. Ces corrections, listées en Annexe 2, consistent typiquement à :

- Ig*norer le facteur indiquant l'individu source de l'échantillon biologique.* Pour un bon nombre d'expériences, le design expérimental inclut un facteur tel que l'i*ndividu*, le p*atient*, le s*ujet, etc.,* alors que le nombre d'individus est clairement insuffisant pour faire une quelconque estimation des variabilités inter-patient, résiduelles, d'intérêt biologique (et d'intérêt pour l'outil de comparaison). Pour ces expériences, la seule possibilité de les exploiter fut donc de confondre la variabilité inter-patient en retirant du design les facteurs du type *individu*, *patient*, *sujet*, etc. Les conséquences possibles de ce choix seront discutées plus loin dans ce mémoire.

- *Retirer du design des microarrays dont les données brutes ne sont pas disponibles ou dont le fichier est endommagé.*

- *Retirer du design des microarrays associés à des combinaisons de niveaux factoriels non répliqués.*

- *Retirer des microarrays qui sont identiques à un autre dans l'expérience.*

- *De fusionner des expériences qui manifestement ne devraient en former qu'une seule.*

Ce processus de collecte et d'inspection laissa finalement un total de 87 et 32 expériences pour les plateformes GPL96 et GPL570 respectivement, dont l'Annexe 1 présente un sommaire accompagné de quelques statistiques sur le design de chacune.

L'ensemble des 119 expériences totalise 2372 microarrays, 423 groupes biologiques (assimilables à des échantillons biologiques) fournissant 809 comparaisons de groupes intra-expérience. La réplication médiane (médiane du nombre médian de réplicat par groupe par expérience) est de 4, valeur typiquement peu élevée. Les thématiques biologiques investiguées semblent suffisamment variées pour satisfaire au critère 3) énoncé plus haut.

On notera que l'expérience GDS2204[85] est une expérience de dilution pour laquelle différentes concentrations (1, 2, 5, 10 µg) d'un mélange de quantités égales de 10 différentes lignées cellulaires ont été hybridées sur des microarrays. Il fut choisi d'inclure cette expérience même s'il n'est pas clair si l'expression différentielle résultante peut vraiment être utilisée pour discriminer entre les performances de différentes méthodes d'analyse. En effet, les gènes les plus susceptibles d'être différentiellement exprimés dans une telle expérience sont ceux qui sont les plus exprimés, qui à leur tour sont possiblement susceptibles d'être liés fonctionnellement.

Contrôle de qualité

Il a été mentionné que la base de données *Data Sets* de *GEO*[86] est plus qu'un simple entrepôt où les chercheurs déposent hâtivement leurs données, mais bien une base de

données revue par des experts pour laquelle on s'attend à un bon niveau pour la qualité des données et des annotations. Cependant, il a été jugé nécessaire d'effectuer un contrôle de qualité sur les expériences recueillies, car tel qu'il fut constaté précédemment, les experts de *GEO* n'ont de toute évidence pas apporté le même soin aux données brutes qu'aux valeurs d'expression soumises par les auteurs (on a vu que certains fichiers sont endommagés, manquants, identiques, etc.). En effet, une correspondance avec des responsables a permis de découvrir que l'équipe de *GEO* a effectué, aussi tard qu'en 2009 seulement (soit vers la fin de ce projet) une série de tests de « centrage médian »[2] pour s'assurer de la normalisation adéquate des données (format SOFT) soumises par l'auteur. On apprit du même coup que trois de nos expériences, soit GDS1574, GDS558, et GDS749 ont été retirés de *Data Sets après* avoir échoué le test en question. Essentiellement, ce test détectera si les microarrays d'une expérience offrent des valeurs d'expression plus ou moins sur la même échelle (donc adéquatement normalisés par l'auteur).

On ne peut se fier uniquement au test effectué par GEO pour évaluer le potentiel d'inclusion d'une expérience dans le jeu de donnée de l'outil de comparaison. Dans un premier temps, puisque ce test n'a été effectué que sur les données normalisées par l'auteur, il est possible que l'expérience soit parfaitement utilisable, si normalisée d'une autre façon plus adéquate. Dans un deuxième temps, il a été démontré que des contrôles de qualités plus sophistiqués comme les méthodes *PLM*[87] peuvent arriver à détecter des microarrays fautifs là où des techniques comme le centrage médian ou les boxplots sont moins sensibles. En résumé, le contrôle de qualité ne semble pas avoir été une préoccupation importante pour *GEO* jusqu'à cette date, à plus forte raison en ce qui concerne les données brutes.

L'ensemble des contrôles de qualité présentés par Gentleman et al.[12] été appliqué aux expériences collectées afin de détecter d'éventuels microarrays problématiques. Le langage R[88] et les libraires *affy*, *affyPLM* du projet Bioconductor[89] ont été utilisé à cette fin. Ce procédé a généré des milliers d'images et de graphiques (impossible de tous les présenter ici) qui ont été inspectés sommairement.

2 Échange courriel avec Stephen Wilhite, PhD, geo@ncbi.nlm.nih.gov

Tel qu'attendu, plusieurs expériences contiennent quelques microarrays dont la qualité est suspecte. Plus particulièrement, GDS1286, GDS1288, GDS2090, GDS2241, GDS2287 et GDS2744 ont été identifiées comme particulièrement inquiétantes compte tenu de leur taille, du nombre de microarrays déviants et par conséquent, de l'impact potentiel sur la validité des ordonnancements de gènes en découlant. Par exemple, il est évident qu'une expérience de 4 microarrays comparant 2 groupes sera sérieusement compromise (au sens d'être peu ou pas informative pour discriminer entre les méthodes) par un microarray fautif alors qu'une autre de 20 le sera moins.

Les figures 16 et 17 montrent quelques résultats pour l'expérience GDS2287[90]. Alors qu'on ne saurait trop que faire (rejeter ou espérer une correction par les algorithmes de pré-traitement) des contrôles de qualité classiques (figure 16, intensités brutes et figure 17a, b), les contrôles basés sur les résiduels d'un modèle PLM (figure 17c,d), révèlent un désaccord sérieux des intensités du microarray suspect avec le reste, et ce même après la correction de fond et une normalisation agressive (quantiles). Il est remarquable que cette expérience, comme bien d'autres, ait passé à travers les mailles du filet de GEO.

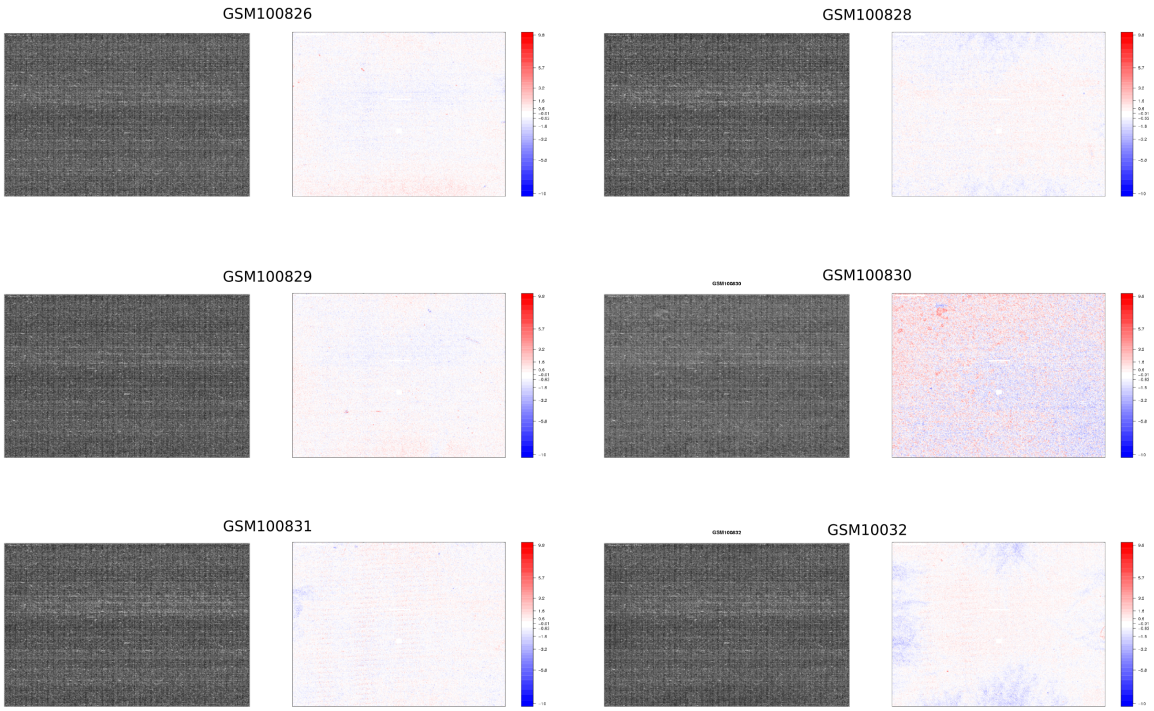


Figure 16: Images des intensités brutes logarithmées et images des résiduels de l'ajustement d'un modèle PLM pour l'expérience GDS2287. Le modèle PLM est le même que pour la méthode de sommarisation robuste RMA discutée en Méthodes, où l'intensité d'une sonde (fond corrigé et normalisation quantiles) est expliquée par la somme d'un terme représentant l'expression du gène sondé et d'un terme représentant l'effet spécifique dû à la séquence de la sonde (qui est présumé être le même d'un microarray à un autre). Les images des résiduels résultantes sont donc libres des valeurs d'expression qui autrement, obscurcissent des artefacts spatiaux sur la puce (par exemple sur GSM10032). On peut remarquer à quel point les résiduels chez GSM100830 sont élevés par rapport aux autres puces.

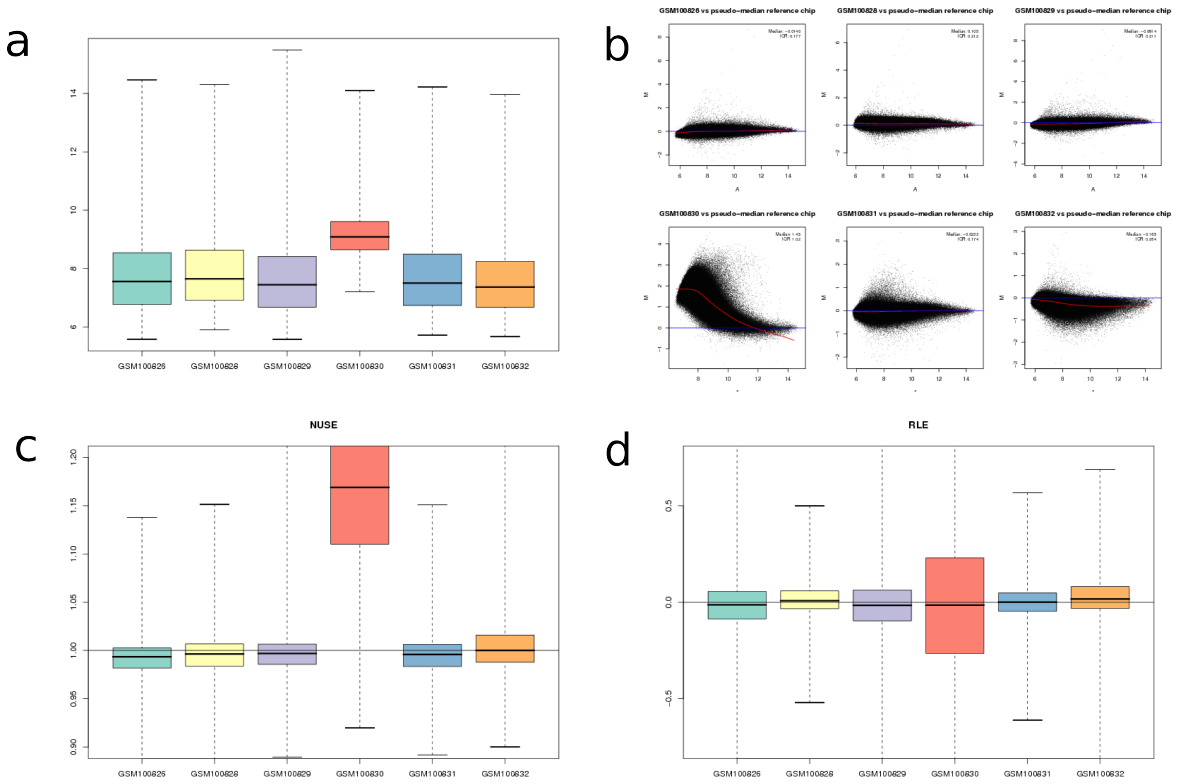


Figure 17: Contrôles de qualité pour l'expérience GDS2287. (a) Boxplot des intensités, échelle logarithmique. Un boxplot est une représentation simplifiée de plusieurs distributions où les extrémités de la boîte indiquent les premiers et troisièmes quartiles et la ligne centrale, la valeur médiane. **(b)** MA plot utilisant une pseudo-puce médiane comme référence, échelle logarithmique. Chaque point d'un MA plot de ce type est une rotation de 45 degrés d'un graphe de dispersion comparant les intensités d'une puce avec l'intensité médiane à travers toutes les puces de l'expérience. Une majorité de points loin de l'abscisse indique que les sondes d'une puce ont tendance à diverger des sondes homologues sur les autres puces. **(c)** NUSE (*Normalized unscaled Standard Error*). Cette visualisation consiste à observer la distribution des erreurs types du coefficient $\hat{\theta}_{gi}$ (standardisés) dans un modèle PLM. Si les intensités des sondes d'une puce ont tendance à ne pas suivre le modèle PLM, la distribution de ces erreurs est non seulement plus variable, mais se centre au-dessus de 1. **(d)** RLE (*Relative Log Expression*). Cette visualisation exploite le fait que l'expression de la plupart des transcrits ne devrait pas varier d'une puce à l'autre et que par conséquent, les différences entre $\hat{\theta}_{gi}$ (d'une puce donnée) et la valeur médiane des $\hat{\theta}_{gi}$ devraient, pour une puce défectueuse apparaître beaucoup plus variables que pour les autre puces.

Les résultats de l'étape du contrôle de qualité soulèvent naturellement la question difficile du traitement des microarrays suspects. Trois options s'offrent alors :

1. *L'exclusion des microarrays suspects*. Cette option, plutôt intuitive, n'est pas nécessairement la meilleure. Premièrement, le choix d'exclure ou non un microarray reste à ce jour hautement subjectif. Deuxièmement, il ne faut pas oublier que le but du projet est de *comparer* les méthodes. Aussi étrange que cela puisse paraître, l'inclusion de microarrays de moins bonne qualité pourrait être une bonne chose en permettant ainsi de comparer la *robustesse* ou la capacité de *« sauvetage »* des méthodes d'analyses. Troisièmement, les données soumises à GEO sont celles qui furent utilisées dans les publications originales. Ne pas réanalyser les expériences dans leur intégralité reviendrait à employer une méthodologie différente de celle des chercheurs en pratique.

2. *L'inclusion des microarrays suspects.* Il est évident que pour un chercheur ayant effectué une expérience peu répliquée, la présence d'un ou deux microarrays défectueux compromet sérieusement le potentiel scientifique de toute l'expérience. Encore une fois, il ne faut pas perdre de vue que l'objectif du projet est ici de comparer les méthodes d'analyse, et ce en se basant sur un très grand nombre d'expériences. Dans le pire des cas, une partie des expériences seront rendues non-informatives par la présence de microarrays de mauvaise qualité. À ce titre, cela présuppose qu'une méthode performant mieux (AUC plus élevés) sur des données de bonne qualité performera aussi bien, sinon mieux sur des données de mauvaise qualité. En d'autres termes, une augmentation du *bruit* entraîne dans le pire des cas une diminution du pouvoir de discrimination entre les méthodes d'analyse, et non un changement dans le classement de leur performance.

3. *L'inclusion ET l'exclusion des microarrays suspects.* Cette option consisterait à intégrer le contrôle de qualité dans le pipeline d'analyse, permettant ainsi de mesurer l'effet d'avoir exclu les microarrays suspect sur les scores AUC.

C'est la seconde voie, soit d'analyser intégralement les expériences soumises à *GEO*, qui fut choisie pour la suite du projet. En effet, il a été jugé que l'outil d'analyse, basé sur plusieurs expériences, est suffisamment robuste pour permettre l'inclusion de quelques expériences de mauvaise qualité. Quant aux raisons justifiant de ne pas avoir choisi la troisième option, soit un dédoublement des temps de calcul et d'espace, elles deviendront évidentes dans les sections qui suivent.

2. Calcul de l'expression différentielle

Une fois les expériences collectées, l'étape suivante de l'implémentation de l'outil de comparaison est d'ordonner les gènes (les s*ondes*, à proprement parler) selon leur niveau d'expression différentielle. Cette opération est répétée pour chacune des comparaisons de groupes offertes par chacune des expériences, en suivant chacune des combinaisons possibles des algorithmes d'analyse présentés en Méthodes (Figure 18).

Tous les calculs ont été effectués en R à l'aide des librairies suivantes[89]: *affy* (ver. 1.16.0), *farms* (ver. 1.3.1), *vsn* (ver. 3.2.1), *gcrma* (ver. 2.10.0), *limma* (ver. 2.18.0), *SAMr* (ver. 1.25), *bayesreg* (1.0beta). Les valeurs des paramètres sont celles par défaut.

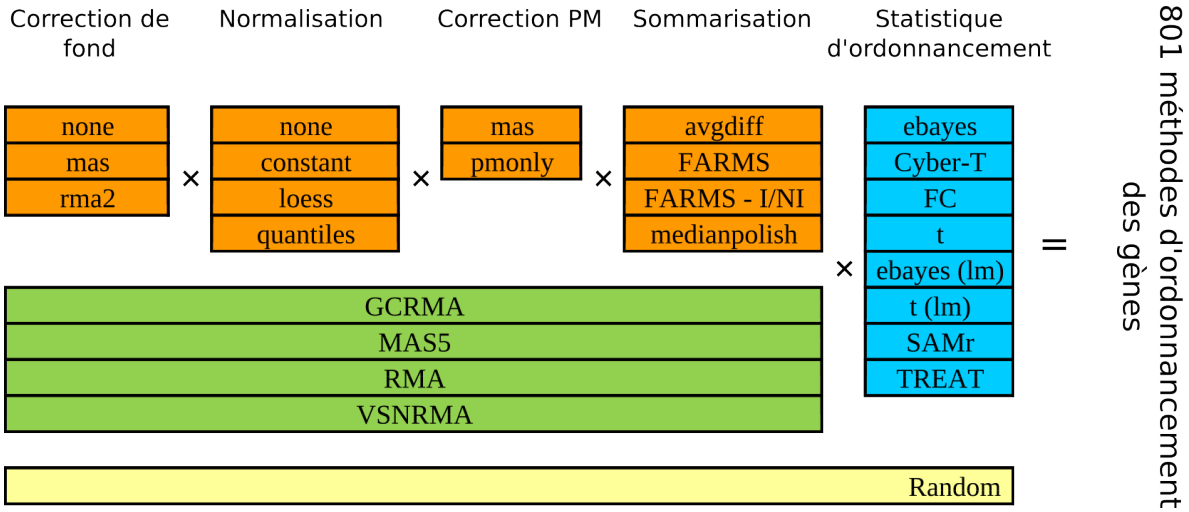


Figure 18: Combinatoire des algorithmes formant le pipeline d'analyse de l'expression différentielle. « none » aux étapes de correction de fond et de normalisation correspond à n'effectuer aucun traitement.

Pré-traitement

Les différentes méthodes de pré-traitement ont donc été appliquées aux microarrays de chaque expérience. Rappelons que le *pré-traitement* (corr. De fond, normalisation, correction PM, sommarisation) consiste à obtenir des valeurs d'expression sommarisées et comparables, certains effets systématiques ayant été estimés et soustraits. La statistique d'ordonnancement fait référence à la statistique employée pour trier les gènes selon le niveau de preuve pour l'expression différentielle.

Pour des raisons évidentes de limites en espace mémoire et en temps de calcul, il fut impossible d'inclure tous les algorithmes de pré-traitement disponibles via le projet Bioconductor, et encore moins toutes les méthodes proposées dans la littérature ou disponibles dans les logiciels commerciaux. Il en est de même pour les statistiques d'ordonnancement qui sont elles aussi légion. Le choix de retenir les algorithmes de la figure 18 se base essentiellement sur la supposition que la disponibilité dans Bioconductor est indicative de popularité, sur des temps d'exécution raisonnables et dans le cas de FARMS, car cette sommarisation sort gagnante la compétition A*ffycomp II*[91]. Les solutions complètes de pré-traitement MAS5 et VSNRMA sont incluses parce qu'elles ne se décomposent pas selon les étapes successives habituelles (voir *Méthodes*). GCRMA et RMA se distinguent par leur algorithme de correction de fond (*rma*, *gcrma*).

Statistiques d'ordonnancement

Pour chaque expérience, toutes les comparaisons de groupes ont été définies (illustré à la Figure 19), pour ensuite ordonner les gènes selon une des statistiques d'ordonnancement appliquées aux valeurs d'expression obtenues à la suite du pré-traitement. Soulignons que les gènes sont toujours tirés selon la valeur absolue de la statistique, aucune distinction n'étant faite entre les gènes surexprimés ou sous exprimés.

Tel que décrit en Méthodes, TREAT, ebayes(lm) et t(lm) ont été calculées à partir d'un modèle linéaire ajusté sur les valeurs d'expression d'un probeset pour tous les microarrays d'une même expérience, plutôt que seulement pour les puces impliquées dans

la comparaison de groupes. Noter par contre que le pré-traitement est malgré tout effectué en incluant toutes les puces d'une expérience. Le seuil de FC pour TREAT a été fixé arbitrairement à $log_2 1.5 \approx 0.74$. Il a aussi été remarqué que la transformation en logarithme ne se fait qu'au niveau de la sommarisation, et que la sommarisation avgdiff n'effectue pas une telle transformation logarithme.

L'algorithme nommé ici FARMS-I/NI joint la sommarisation FARMS et les appels I/NI (*informative/non-informative calls*). Comme la métrique d'enrichissement choisie demande l'ordonnancement complet des sondes, la liste triée des probeset non-informatifs a été jointe à la fin de la liste triée des probeset informatifs.

La statistique *random* correspond à un ordonnancement aléatoire. Cet ordonnancement permettra d'apprécier comment se comportent (variance et moyenne) les scores d'AUC lorsqu'aucun « signal biologique » n'est présent dans une expérience, où de façon équivalente, dans les signatures moléculaires.

Au total, l'étape du calcul de l'expression différentielle aura donc généré $809 \times 801 = 648009$ ordonnancements.

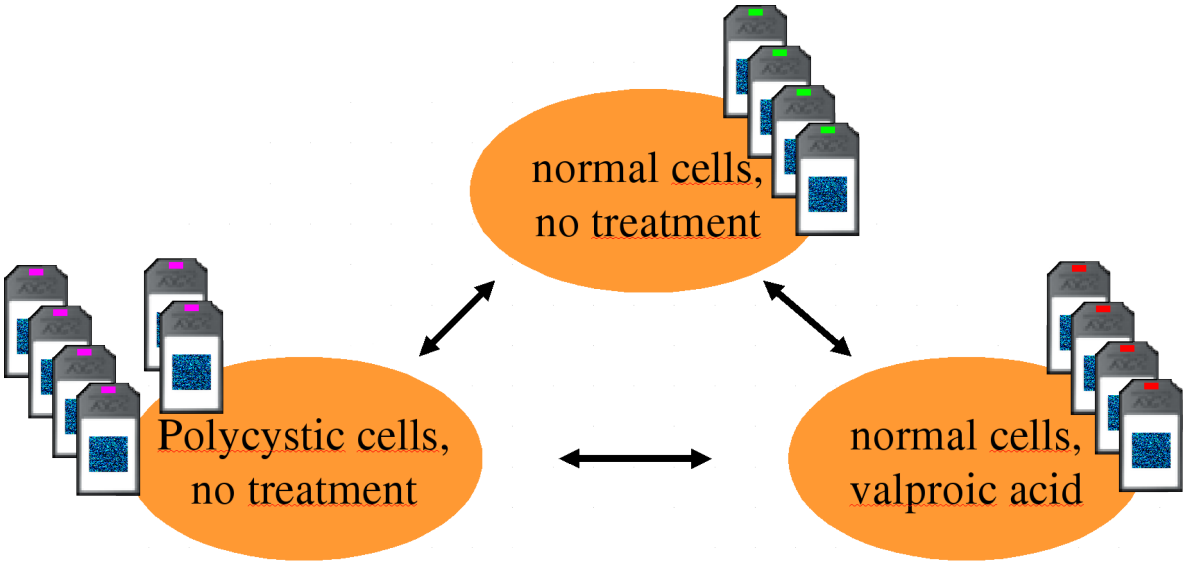


Figure 19: Exemple de la façon par laquelle les contrastes ont été définis. Le design de l'expérience GDS1050 (Wood et al.) comporte deux facteurs. Le premier, *disease state* avec les niveaux *normal cells* ou

polycystic cells, le second, *agent,* avec les niveaux *untreated* ou *valproic acid*. Cette expérience fournira donc trois ordonnancements qui sont ensuite soumis aux 801 méthodes d'analyses de l'expression différentielle.

3. Signatures moléculaires

Tel que mentionné en Méthodologie, MSigDB est la source des listes de gènes utilisée lors du projet. Cette section présente les résultats de collecte, de conversion, de nettoyage et d'inspection des signatures moléculaires pour lesquelles les scores d'AUC seront ultimement calculés.

Conversion en listes de probesets

L'entièreté des signatures moléculaires décrites en Méthodologie a été téléchargée du site web de MSigDB (format XML version 2.5), puis lue en mémoire avec la librairie Bioconductor *GSEABase*[89].

Les fichiers d'annotation, c'est à dire donnant la correspondance entre les *symboles*[a] et les identifiants probesets Affymetrix des puces HG-U133A et HG-U133 Plus 2 ont été récupérés sur le site ftp de MSigDB (fichiers *.chip*). Selon ces annotations, respectivement 13 828 et 21 229 symboles uniques sont sondés par les 22 283 et 54 675 probesets des puces HG-U133A et HG-U133 Plus 2. Les annotations de MSigDB ont été préférées à d'autres, plus à jour, pour des raisons de synchronisation de version avec les signatures moléculaires elles-mêmes. Un retour sur la qualité de ces annotations sera effectué en discussion.

Un bref survol des annotations a révélé quelques erreurs dans le fichier d'annotation qui a dû être corrigé, typiquement des problèmes de tabulation, de séparateur ou d'espaces en trop. Cette formalité est mentionnée, car il fut découvert plus tard que l'équipe de MSigDB offre déjà un programme de conversion des symboles en probesets (*Chip2Chip*), et qu'il est fort probable que la conversion utilisée ici soit de meilleure qualité.

---

a Identifiant unique à chaque gène; *gene symbol* tel que défini par le HUGO *Gene Nomenclature Committee*.

Ainsi donc, l'étape suivante fut effectivement de faire passer les signatures moléculaires de la forme « listes de symboles » vers « listes de probesets », pour plus tard calculer leurs enrichissements (scores AUC) pour un ordonnancement de probesets, lui-même, rappelons-le, généré par l'application d'une méthode d'analyse donnée sur une expérience donnée. Le tableau 2 présente les tailles des collections MSigDB avant et après conversion, ainsi que quelques statistiques de MSigDB après conversion. Noter que ces tailles sont différentes après conversion, puisque certains gènes ne sont tout simplement pas sondés par les microarrays, laissant ainsi certaines signatures moléculaires avec une taille nulle (seules les listes de deux probesets et plus ont été retenues).

Tableau 2: Tailles des collections, ainsi que quelques statistiques de MSigDB avant et après conversion en listes de probesets.

Collection	avant	après	
		HG-U133A	HG-U133 Plus 2
c1	*386*	*345*	*356*
c2:cp	*639*	*637*	*637*
c2:cgp	1 186	1 186	1 186
c3:mir	222	222	222
c3:tft	615	615	615
c4:cm	456	456	456
c4:cgn	427	427	427
c5:cc	*233*	*232*	*232*
c5:bp	825	825	825
c5:mf	396	396	396
TOTAL	5 452	5 408	5 419
Nbr. de symboles uniques représentés dans MSigDB	38 509	13 731	20 737
Nbr. de probesets uniques représentés dans MSigDB		21 058	45 519
Nbr. médian de symboles par signatures moléculaire	47	42	47
Nbr. médian de probesets par signatures moléculaire		72	118

Inspection des signatures : À la recherche du *feedback*

Il fut décrit en introduction comment les scores d'enrichissement servent de proxy à la précision, alors qu'un score plus élevé pointe vers une plus grande ressemblance avec l'ordonnancement réel. Cependant, un score plus élevé peut être causé par autre chose qu'une plus grande ressemblance avec l'ordonnancement réel, et cette « autre chose » peut systématiquement jouer en faveur d'une méthode d'analyse ou d'une autre. C'est que l'on qualifie alors de *biais*, et tout outil de comparaison doit autant que possible tenter de les éliminer.

À cet effet, une attention particulière a été apportée à la façon avec laquelle les signatures moléculaires de MSigDB sont construites. La raison est bien simple : ces signatures moléculaires ne doivent pas être dérivées d'une expérience elle-même utilisée pour l'outil de comparaison, sans quoi il y a risque de créer un effet de rétroaction (*feedback)* envers la méthode initialement employée pour analyser l'expérience. Quant aux signatures moléculaires dérivées d'expérience de microarrays qui ne font pas partie du jeu de donnée utilisé pour la comparaison, elles peuvent être informatives, mais doivent être approchées avec prudence : ces dernières fournissent assurément de « l'information biologique », mais rien n'assure qu'elles ne sont pas porteuses d'un biais quelconque propre à cette technologie.

Ainsi, les collections ont été inspectées manuellement à la recherche d'un effet de feedback. Les collections c1 et c3, étant strictement inférées de séquences génomiques, elles ne posent évidemment aucun problème, les scores d'enrichissement qui en sont tirés sont donc particulièrement fiables. L'origine de chacune des listes composant la collection c2:cp (p.e. KEGG, BioCarta) est difficilement traçable, mais il semble raisonnable de supposer que les voies moléculaires dont elles sont tirées résultent d'expériences « classiques » méticuleusement compilées et non de microarrays.

Le cas de c2:cgp est problématique, considérant que la plupart de ces listes de gènes sont dérivées d'expériences de microrarrays. La liste des publications originales de chacune

des expériences collectées pour l'outil a donc été croisée avec la liste des publications de chaque signature moléculaires de cette collection. Un effet de feedback a été trouvé pour deux expériences, GDS785 GDS810 (toutes deux sur la plateforme HG-U133A) avec respectivement 10 et 4 signatures moléculaires. Les 10 signatures LEE_TCELL provenant de GDS785 (Lee et al.[92]) ont été construites par pré-traitement RMA suivi d'une approche ad hoc test-t / FC / clustering par SOM. Les 4 signatures ALZHEIMERS_DISEASE -_INCIPIENT provenant de GDS810 (Blalock et al.[93]) ont quant à elles été construites par pré-traitement MAS5, suivi d'un test de corrélation avec certaines variables psychométriques, pour ensuite identifier des gènes différentiellement exprimés par un ANOVA (test-t).

Le cas de c4 est semblable à c2:cgp, leurs signatures étant inférées d'expériences de microarrays. Toutefois, l'inspection de ces dernières a révélé qu'aucune d'entre elles n'est utilisée pour l'outil de comparaison.

En ce qui concerne les signatures de c5, on a tiré avantage des codes d'évidence afin de retracer l'origine des signatures qui composent cette collection. En effet, seules les signatures de type IEP (*Inferred from Expression Pattern*), code d'ailleurs réservé à l'ontologie biological processes (c5:bp), sont susceptibles d'être inférées de données de microarrays. Une inspection du fichier gene2go[a] a révélé 10 associations gène/terme GO portant le code d'évidence IEP et inférées d'expériences de microarrays, dont aucune membre du jeu de donnée de l'outil de comparaison. Considérant la taille de c5:bp (825 signatures) et qu'au plus 10 gènes sont potentiellement affectés, il semble raisonnable de supposer la collection c5 comme libre de tout biais envers l'une ou l'autre des méthodes d'analyse de microarrays.

La conclusion de cette section est la suivante : les collections c1, c2:cp, c3 et c5 fournissent des signatures assurément non biaisées, tandis que les scores d'enrichissements

a (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz) Fichier listant les annotations par terme GO des gènes. Ce fichier ne liste toutefois que les annotations minimales, c'est-à-dire que les annotations implicites des parents d'un terme sont omises.

calculés sur les collections c2:cgp et c4 doivent être interprétés avec prudence, car issus d'expériences de microarrays. Pour la suite des résultats, il fut décidé de n'exclure aucune des signatures de MSigDB, quitte à analyser les scores de différentes sous-collections et différentes expériences séparément. Il sera vu, plus loin, comment les scores peuvent être observés, soit par expérience, soit par sous-collection.

4. Calcul et Analyse des AUC

Classement général des méthodes

L'AUC a donc été calculé pour toutes les combinaisons de méthodes d'analyse, d'expériences et de signature moléculaires décrites précédemment, les quelques milliards de scores résultants étant sauvegardés dans une base de données MySQL. Noter que dans le cas qui nous occupe, la moyenne arithmétique a été choisie comme statistique sommaire des AUC. Les différentes méthodes d'analyses sont donc comparées sur la base de l'AUC moyen, où l'AUC moyen peut être défini de différentes façons tout dépendant de la question à laquelle on cherche à répondre.

Une première approche aux scores d'enrichissement est de simplement classer les différents pipelines pris individuellement, c'est à dire par combinaisons d'algorithmes. Rappelons que pour chaque pipeline, on dispose alors d'une valeur d'AUC par combinaison de signature moléculaire et de contraste. La première étape vers un critère de classement général des pipelines est donc de faire la moyenne sur l'ensemble des signatures moléculaire. Deux objections peuvent être opposées à cette façon de faire : 1) tel qu'exposé à la section précédente, les signatures moléculaires sont divisées en sous-collections dont la fiabilité de certaines est douteuse et 2) les signatures moléculaires ne sont pas toutes « pertinentes » à une expérience donnée. Le premier point sera traité à la section suivante alors que les sous-collections seront considérées individuellement. En réponse à la seconde objection, il fut déjà discuté en introduction que l'effet attendu de l'inclusion de signatures moléculaires non pertinentes n'est qu'une diminution du pouvoir de discrimination (sensibilité) entre les méthodes comparées. Autrement dit, les différences observées entre

les méthodes sont valables (*spécificité* de la comparaison), mais certaines différences risquent d'être masquées (*sensibilité* de la comparaison).

L'étape suivante est de résumer la performance d'un pipeline donné à travers les expériences. Une approche naïve pourrait être de simplement calculer la moyenne globale des AUC pour un pipeline donné. Toutefois, le poids d'une expérience particulière dans une telle moyenne globale est alors proportionnel au nombre de contrastes qu'elle comporte. Considérant les nombres inégaux de contrastes par expérience, et comme il semble important d'exploiter l'indépendance de ces dernières, la seconde étape vers un critère de classement est de faire la moyenne des AUC par pipeline et expérience. Les scores d'AUC moyens par pipeline et par expérience sont ici approchés de deux façons différentes. La première est de calculer l'AUC moyen pour toutes les expériences. La seconde, que l'on pourrait qualifier de non paramétrique, calcule plutôt le rang moyen d'une méthode à travers les expériences, l'objectif étant d'accorder un poids égal à chacune, peu importe la grandeur du « signal » qu'elles offrent. Un sommaire du classement final est rapporté aux tableaux 3 et 4.

On peut conclure des deux classements présentés que :

- Les scores d'enrichissement sont supérieurs à la valeur attendue de 0.5 pour la plupart des pipelines d'analyse. Cela démontre la tendance des gènes qui composent les signatures moléculaires à se retrouver dans le haut des listes de gènes ordonnés selon leur niveau d'expression différentielle. La section suivante va montrer que cette tendance est présente pour la plupart des expériences prises individuellement, et non le fait de quelques expériences seulement.

- TREAT et FC, les statistiques d'ordonnancement basées sur le fold-change, dominent largement la tête du classement. Quelle que soit la combinaison d'algorithmes de pré-traitement, FC et TREAT affichent la meilleure performance. L'étape qui a le plus d'influence sur la performance est de toute évidence la statistique d'ordonnancement.

- Les solutions complètes de pré-traitement se classent relativement bien, mais de meilleures combinaisons d'algorithmes existent. VSNRMA termine première parmi celles-ci. Noter que la correction de fond et la normalisation de VSNRMA n'ont pas été testées séparément au cours de cette étude. Peut-être qu'une combinaison avec la sommarisation FARMS plutôt que medianpolish améliorerait significativement sa performance.

Tableau 3 : Sommaire du classement des pipelines pris individuellement pour la plateforme HG-U133A. Sont ici montrés les 10 premiers pipelines, les deux derniers, la première occurrence d'un algorithme particulier (en soulignement) et la première occurrence d'une solution complète de pré-traitement.

Rang final	Correction de fond	Normalisation	Correction PM	Sommarisation	Stat. d'ord.	AUC moyen	Rang moyen
1	none	quantiles	pmonly	FARMS	TREAT	0.58014	5.17
2	none	quantiles	pmonly	FARMS-I/NI	TREAT	0.58013	5.25
3	mas	quantiles	pmonly	FARMS-I/NI	TREAT	0.57971	5.95
4	mas	quantiles	pmonly	FARMS	TREAT	0.57971	6.22
5	rma2	quantiles	pmonly	FARMS-I/NI	TREAT	0.57939	7.98
6	rma2	quantiles	pmonly	FARMS	TREAT	0.57939	8.24
7	none	loess	pmonly	FARMS	TREAT	0.57947	8.45
8	none	loess	pmonly	FARMS-I/NI	TREAT	0.57944	8.75
9	mas	loess	pmonly	FARMS-I/NI	TREAT	0.57902	9.02
10	mas	loess	pmonly	FARMS	TREAT	0.57902	9.37
13	mas	constant	pmonly	FARMS-I/NI	TREAT	0.57545	25.56
15	rma2	quantiles	mas	FARMS-I/NI	TREAT	0.56732	27.63
19	none	loess	pmonly	medianpolish	TREAT	0.54683	45.83
24	VSNRMA				TREAT	0.54552	50.54
25	none	loess	pmonly	FARMS-I/NI	FC	0.54156	54.38
50	rma2	quantiles	mas	avgdiff	FC	0.53950	77.74
59	rma2	none	mas	avgdiff	FC	0.53878	84.17
61	none	loess	pmonly	FARMS-I/NI	Cyber-T	0.53771	90.03
72	MAS5				FC	0.53649	105.72
94	RMA				TREAT	0.53522	131.63
103	GCRMA				TREAT	0.53425	148.87
108	none	loess	pmonly	FARMS-I/NI	ebayes	0.53207	154.33
110	none	loess	pmonly	FARMS-I/NI	ebayes (lm)	0.53197	156.30
119	none	loess	pmonly	FARMS-I/NI	SAMr	0.53140	161.77
125	none	loess	pmonly	FARMS-I/NI	t	0.53103	171.10
126	none	loess	pmonly	FARMS-I/NI	t (lm)	0.53100	171.39
800	rma2	none	mas	avgdiff	t (lm)	0.50349	666.71
801	random					0.50022	722.14

Tableau 4 : Sommaire du classement des pipelines pris individuellement pour la plateforme HG-U133 Plus 2. Sont ici montrés les 10 premiers pipelines, les deux derniers, la première occurrence d'un algorithme particulier (en soulignement) et la première occurrence d'une solution complète de pré-traitement.

Rang final	Correction de fond	Normalisation	Correction PM	Sommarisation	Stat. d'ord.	AUC moyen	Rang moyen
1	none	loess	pmonly	FARMS-I/NI	TREAT	0.566743	13.50
2	none	loess	pmonly	FARMS	TREAT	0.566735	14.00
3	none	quantiles	pmonly	FARMS-I/NI	TREAT	0.565887	16.22
4	none	quantiles	pmonly	FARMS	TREAT	0.565880	16.84
5	rma2	loess	mas	avgdiff	FC	0.564732	20.19
6	rma2	quantiles	mas	avgdiff	FC	0.564343	20.97
7	rma2	constant	mas	avgdiff	FC	0.563921	21.66
8	mas	loess	pmonly	FARMS-I/NI	TREAT	0.565137	23.06
9	rma2	none	mas	avgdiff	FC	0.564103	23.09
10	mas	loess	pmonly	FARMS	TREAT	0.565110	24.44
38		VSNRMA			TREAT	0.556358	59.03
40	none	loess	pmonly	medianpolish	TREAT	0.557506	60.38
57		MAS5			FC	0.556132	76.09
75		GCRMA			TREAT	0.550496	117.47
78	none	loess	pmonly	FARMS-I/NI	Cyber-T	0.549068	122.00
86	none	loess	pmonly	FARMS-I/NI	ebayes (lm)	0.547224	141.45
88	none	loess	pmonly	FARMS-I/NI	ebayes	0.547226	142.39
96	none	loess	pmonly	FARMS-I/NI	SAMr	0.546546	149.09
104	none	loess	pmonly	FARMS-I/NI	t (lm)	0.546140	156.05
110	none	loess	pmonly	FARMS-I/NI	t	0.546088	157.70
214		RMA			TREAT	0.540132	253.59
791			random			0.499725	732.19
800	mas	quantiles	mas	medianpolish	FC	0.489455	775.03
801	mas	quantiles	mas	medianpolish	TREAT	0.485449	778.00

Départage des algorithmes par étape d'analyse

Cette section s'intéresse aux performances des méthodes d'analyse prises individuellement pour chacune des étapes de l'analyse.

Tel que précédemment, l'AUC moyen a d'abord été calculé pour chaque combinaison de signature moléculaire, pipeline d'analyse et expérience afin de donner un poids égal à chaque expérience. L'objectif étant de comparer les algorithmes d'une même étape d'analyse entre eux, ce calcul a cette fois été suivi du calcul d'une moyenne pour chaque algorithme d'une étape d'analyse (et expérience, et signature moléculaire), peu importe la méthode employée pour le reste du pipeline. Les AUC moyens résultants pour chaque algorithme ont ensuite été combinés de deux façons différentes :

- Par expérience, afin d'observer la contribution de chaque expérience.
- Par sous-collection MSigDB, afin d'observer la contribution de chaque sous-collection aux AUC moyens calculés à la section précédente, et surtout de s'assurer

que les tendances observées ne sont pas le fruit d'un phénomène de rétroaction tel que discuté précédemment.

Les « séries » d'AUC moyen résultantes sont tracées pour chacune des étapes d'analyse aux figures 20, 21, 22, 23 et 24.

Correction de fond

Les AUC moyens montrés à la figure 20 suggèrent que le choix, parmi les trois méthodes de correction de fond testées, soit *rma2*, *mas* et aucune correction (*none*), a peu d'impact sur l'analyse de l'expression différentielle sur la plateforme HG-U133A. Malgré un léger avantage en faveur de *rma2* par rapport à l'absence de correction ($P_{Wilcoxon} = 0.01$), les rangs moyens des trois méthodes restent forts semblables. La situation est différente pour la plateforme HG-U133 Plus 2.0, alors que *rma2* se classe systématiquement mieux (rang moyen $1,28$) que *mas* ou *none* ($P_{Wilcoxon} = 2,62 \times 10^{-6}$). Mentionnons que les AUC moyens ont le même comportement au niveau des sous-collections MSigDB. De plus, il est aussi intéressant de noter que le meilleur pipeline sur les deux plateformes n'effectue pas de correction de fond, et que les trois méthodes sont présentes parmi les 10 meilleurs pipelines (Tableaux 3 et 4).

Rappelons que le premier objectif de la correction de fond est l'amélioration de la sensibilité. Il est en effet facile de démontrer que la présence de bruit biaise l'estimation du fold-change vers l'unité aux basses intensités[94]. Le second objectif de la correction de fond est l'amélioration de la spécificité lorsque le nombre de réplicats est faible, accomplit en éliminant la variance (non informative) aux basses intensités afin d'éviter le « surpeuplement » du haut du classement avec des gènes dont la statistique-t ou le fold-change n'est significatif que par chance, dû au bruit. La présence de ce genre de faux positifs est aussi appelée *fanning*[95]. Dans ce contexte, il est possible que la correction de fond gagne en importance à mesure que le nombre de sondes de faibles intensités augmente. La différence de l'importance relative de la correction de fond observée entre les deux platformes pourrait donc être une conséquence du nombre plus élevé de sondes pour

HG-U133 Plus 2, ainsi que la proportion beaucoup plus élevée de ces dernières qui ne sondent pas un gène connu, soit 25 %, comparativement à 9 % pour HG-U133A.

Quoi qu'il en soit, les résultats obtenus n'ont pas démontré le caractère absolument indispensable de la correction de fond, du moins pour les deux méthodes testées, *mas* et *rma2*. Cette situation est surprenante, mais ne remet pas en question l'acuité de la méthodologie de comparaison proposée. D'une part, peut-être que le pouvoir de discrimination est insuffisant pour départager les algorithmes de correction de fond, et que le départage serait possible avec un plus grand nombre d'expériences, de meilleures signatures moléculaires ou encore une métrique de comparaison plus appropriée que les AUC moyens. D'autre part, la littérature n'est pas sans équivoque quant à la pertinence d'effectuer la correction de fond. Smyth et al.[96] ont récemment démontré par exemple que les méthodes de correction de fond s'inscrivent toutes dans un « spectre » du compromis entre spécificité et sensibilité, et que l'omission de correction de fond reste tout de même une option envisageable.

Normalisation

Pour une même expérience, la distribution des intensités varie grandement d'une puce à une autre et la nécessité de normaliser est indiscutable, avant même de poser la question de la comparaison des différentes méthodes. En ce sens, l'importance de la normalisation observée à la figure 21 offre un contrôle positif qui valide la méthodologie de comparaison proposée. En effet, les AUC moyens calculés sur les pipelines sans aucune normalisation se classent loin derrière les autres méthodes, avec un rang moyen de 3,61 et 3,16 respectivement sur les deux plateformes, avec des valeurs-p du test de Wilcoxon largement significatives. Étrangement, on observe aussi pour quelques expériences, des AUC moyens bien en deçà des ordonnancements aléatoires en l'absence de normalisation. L'importance de normaliser est de plus observée sur pratiquement toutes les sous-collections de MSigDB.

Les AUC moyens ne parviennent pas à départager *quantiles* et *constant*, mais *loess* affiche une légère supériorité avec un rang moyen qui s'en détache ($P_{Wilcoxon} = 1,6 \times 10^{-06}$ et 4×10^{-02}). Malgré cela, *quantiles* est tout de même la normalisation effectuée par le meilleur pipeline sur la plateforme HG-U133A, suivie de très près par *loess*. En somme, les résultats obtenus ne permettent pas de trancher sans équivoque entre *quantiles* et *loess*, deux méthodes populaires en pratique.

Correction PM

Rappelons que la correction PM a pour but de retirer la portion non spécifique du signal, par exemple due à l'hybridation croisée, en exploitant les sondes MM propres à la technologie GeneChip®. Comme il a été observé ailleurs[36], la AUC moyens montrés à la figure 22 suggèrent que le fait d'ignorer les intensités MM mène à de meilleures performances, dans le cas présent pour 68 % ($P_{Wilcoxon} = 6,02 \times 10^{-5}$) et 84 % ($P_{Wilcoxon} = 2.01 \times 10^{-6}$) des expériences et pour toutes les sous-collections MsigDB. Ces résultats suggèrent que l'utilisation des intensités MM introduit un bruit supplémentaire, du moins dans le cadre de l'algorithme *mas* proposé par Affymetrix.

Sommarisation

Les AUC moyens par méthode de sommarisation montrés à la figure 23 sont sans équivoque quant à la supériorité des deux versions de la sommarisation FARMS sur *avgdiff* ou *medianpolish,* peu importe la sous-collection MSigDB et sur écrasante majorité des expériences, peu importe la plateforme. Ce résultat est d'ailleurs en parfait accord avec les résultats du banc d'essai Affycomp[91], sur lequel le pipeline employant la sommarisation FARMS avec filtrage I/NI termine premier, et les pipelines *l.farms* et *q.farms* (pas de correction de fond, normalisation *loess* ou *quantiles*) terminent respectivement en troisième et quatrième places.

L'explication probable de l'écrasante domination de FARMS accompagnée du filtrage I/NI est en lien avec la statistique d'ordonnancement et sera discutée à la section suivante.

Il est aussi fort curieux de constater que *medianpolish* ne parvient pas, en moyenne, à surpasser *avgdiff,* une méthode qui non seulement n'est pas robuste[97], mais n'exploite pas l'architecture en probesets des GeneChip, comme le font FARMS ou *medianpolish*. Ce résultat est contre-intuitif sans toutefois être alarmant. En effet, la seule autre comparaison directe (c'est à dire en fixant les autres étapes du pré-traitement) entre *medianpolish* et *avgdiff* disponible dans la littérature conclue aussi à un léger avantage pour *avgdiff* (Bolstad[60], section 5.3). Toutefois, cette étude fut réalisée sur une expérience de dilution et non des données réelles. De plus, pour le meilleur pipeline (*none, loess, pmonly, treat*), les AUC moyens de *medianpolish* dépassent largement ceux de *avgdiff* (données non montrées).

Statistique d'ordonnancement

La figure 24 rapporte les AUC moyens, départagés cette fois par statistique d'ordonnancement. Tel que prévu par la littérature[98], la statistique *t* ordinaire et son homologue du modèle linéaire *t (lm)* génèrent les AUC moyens les plus faibles suivies des versions régularisées, qui par ordre de performance sont *SAMr*, *ebayes (lm)*, *ebayes* et *CyberT*, qui se détache manifestement des autres statistiques-t, comme d'autres l'ont observé[33,99,98].

Un constat crucial est que les ordonnancements par simple fold-change (FC) produisent des AUC systématiquement supérieurs aux statistiques-t régularisées, à l'exception d'une seule expérience sur la plateforme HG-U133A , et d'une seule sous-collection MSigDB, *c1* (en l'occurrence l'expérience et la sous-collection pour lesquelles le signal est le moins élevé). Ce résultat est fort intéressant puisqu'il fait écho à une série d'études récentes insistant sur l'importance de considérer la grandeur du changement

(problème de classement et de sélection) et non seulement sa significativité (problème de test). La section suivante traitera de ce problème plus en détail.

Dans la même veine, il est fort intéressant de constater que la statistique TREAT, un test d'hypothèse avec seuil sur le fold-change, fait figure de meilleure méthode si l'on combine les deux plateformes. Encore une fois, ce dernier résultat réaffirme l'importance que la considération de la grandeur du changement est déterminante pour la performance d'une statistique de détection de l'expression différentielle, et non seulement sa significativtié (rapport signal/bruit). La supériorité de TREAT sur *ebayes* pour des données simulées et réelles a d'ailleurs aussi été observée par les auteurs de ces même deux méthodes[100]. Assez curieusement, ce même article, en comparant différentes stratégies de sélection (TREAT, *ebayes*, FC, FC+seuil sur *ebaye*s, *t* et autres) passe relativement sous silence le fait que le simple fold-change performe alors lui aussi mieux que *ebayes*.

On remarquera que les statistiques *t* et *t (lm)* sont indistinguables en termes d'AUC moyens ($P_{Wilcoxon} = 0,186$ et $0,229$). Pourtant, un meilleur estimé de la variance mène théoriquement à des tests pour l'expression différentielle plus puissants. De plus, on observe même une diminution significative des AUC moyens pour *ebayes (lm)* ($P_{Wilcoxon} = 5,03 \times 10^{-10}$ et $2,32 \times 10^{-3}$). Une explication possible[3] serait que la procédure de régularisation d'*ebayes* attribue un poids moindre à la variance mise en commun (*pooled variance*) pour la version *lm*. Les statistiques régularisées du modèle linéaire complet (de la forme $t_{ebayes(lm)} = \frac{FC}{s+s_0}$) tendent par conséquent moins vers le fold-change : plus la régularisation est importante, plus la statistique tend vers le FC. Il est connu que *CyberT* est la statistique la plus régularisée (avec les paramètres par défaut), suivie de *ebayes*, suivie de SAMr. De toute évidence, les AUC moyens semblent favoriser les statistiques fortement régularisées, allant de t vers FC, les ordonnancements résultants de FC étant essentiellement les même qu'une statistique-t régularisée à l'infini ($s_0 \rightarrow \infty$).

3 Suggérée par un arbitre lors de la soumission d'un article.

Retour sur FARMS-I/NI; le filtrage non spécifique

Jusqu'ici, ce mémoire n'a pas encore fait allusion à la pratique pourtant courante du filtrage *non spécifique* des gènes. Par non spécifique, on entend de façon non informée par le design expérimental, donc indépendamment des variables qui sont testées pour l'expression différentielle. Ce filtrage peut se faire de différentes façons, les plus courantes étant de ne retenir que les gènes dont la variance et/ou l'intensité moyenne excèdent un certain seuil. Le principal argument en faveur du filtrage est que les gènes non différentiellement exprimés sont peu variables et/ou peu exprimés. En les filtrant, on diminue le nombre d'hypothèses subséquentes à tester et par conséquent le nombre de faux positifs (pour un même nombre de gènes sélectionnés, cela se manifeste par de meilleures valeurs-p corrigées pour tests multiples). Récemment, Huber et al.[101] ont offert un exposé rigoureux sur la procédure de filtrage. Ils y concluent qu'afin de maintenir le taux de contrôle des faux positifs, la paire filtre/statistique doit être indépendante sous l'hypothèse nulle et corrélée sous l'hypothèse alternative. Ainsi, le simple test-t gagne par exemple beaucoup d'un filtre sur la variance alors que par exemple, *ebayes*, qui utilise la variance globale dans le calcul de régularisation, risque théoriquement d'être affectée par la procédure de filtrage. Huber et al. recommandent soit d'effectuer un test-t suivit d'un filtre non spécifique ou 2) d'employer directement une statistique-t régularisée comme *ebayes*.

Les résultats du présent projet sont en accord avec la littérature en ce qui concerne la performance relative du test-t ordinaire, dont les AUC moyens sont les plus faibles. À la lumière des travaux de Huber et al., on s'attendrait donc à une augmentation importante des AUC moyens du test-t s'il est précédé d'un filtrage non spécifique. À défaut d'avoir explicitement inclus l'étape du filtre parmi les étapes d'analyse, la sommarisation FARMS I/NI implémente une forme de filtre : les sondes non informatives, plutôt que d'être mises de côté, sont adjointes aux sondes informatives, les deux types de sondes étant séparément triées selon la statistique d'ordonnancement. La figure 25 illustre bien l'effet important sur les AUC moyens qu'à la procédure de filtrage, et comment le filtrage a pour effet de « réhabiliter » la statistique-t ordinaire. D'une certaine façon, il s'agit d'une validation supplémentaire de la méthodologie de comparaison proposée. On remarquera aussi que

ebayes et *SAMr* bénéficient aussi du filtrage, ce qui vient appuyer l'idée que la régularisation implémentée par défaut par ces statistiques n'est pas optimale.

Huber et al. offrent une discussion intéressante sur la relation entre la procédure filtrage-variance/t et la procédure de double filtrage t/FC (volcano plot) promue par le MAQC comme maximisant la reproductibilité[102]. Huber et al. avançent que le filtrage non spécifique sur la variance revient à imposer une limite inférieure au fold-change. Il très intéressant de constater que le « débat » sur le double filtrage t/FC[56], le filtrage non-spécifique et la régularisation de la statistique-t convergent vers le fait qu'une bonne statistique ne devrait pas que considérer la significativité du changement, mais aussi sa grandeur.

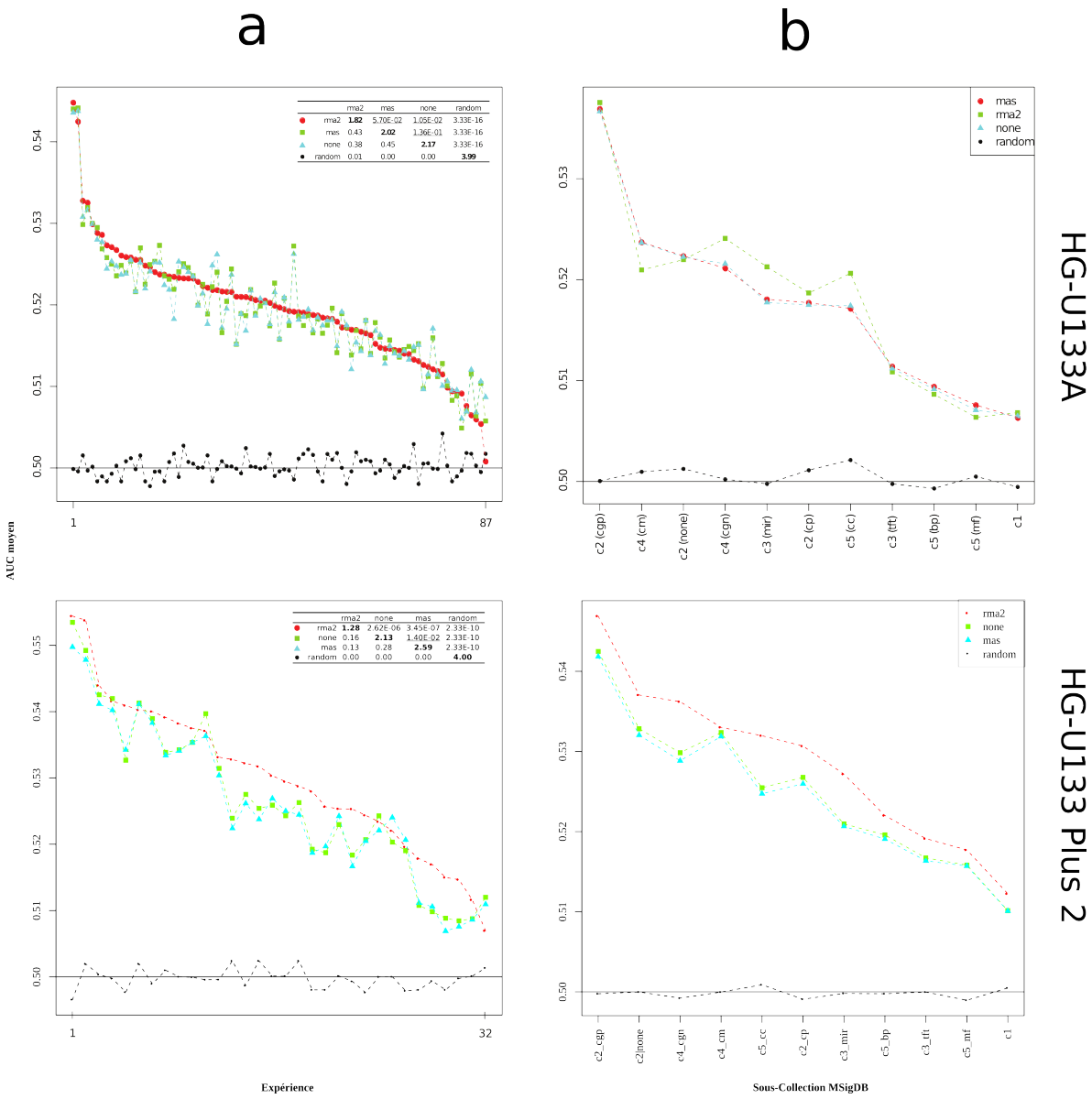


Figure 20 : Correction de fond. Série d'AUC moyens par **(a)** expérience ou **(b)** sous-collection MSigDB, et ce pour les deux plateformes Affymetrix. Le tableau en encart pour **(a)** présente trois statistiques différentes sur les séries. En position *i* de la diagonale, la moyenne des rangs dans chaque expérience de la méthode *i*. Sous la diagonale, en position (*i*,*j*), la proportion des expériences pour lesquelles l'AUC moyen de la méthode *i* est supérieur à la méthode *j*. Sur la diagonale, la valeur-p d'un test de Wilcoxon pour échantillons appariés entre les méthodes *i* et *j*. Les expériences et sous-collections MSigDB en abscisse sont triées selon l'AUC moyen pour la méthode dont le rang moyen maximal.

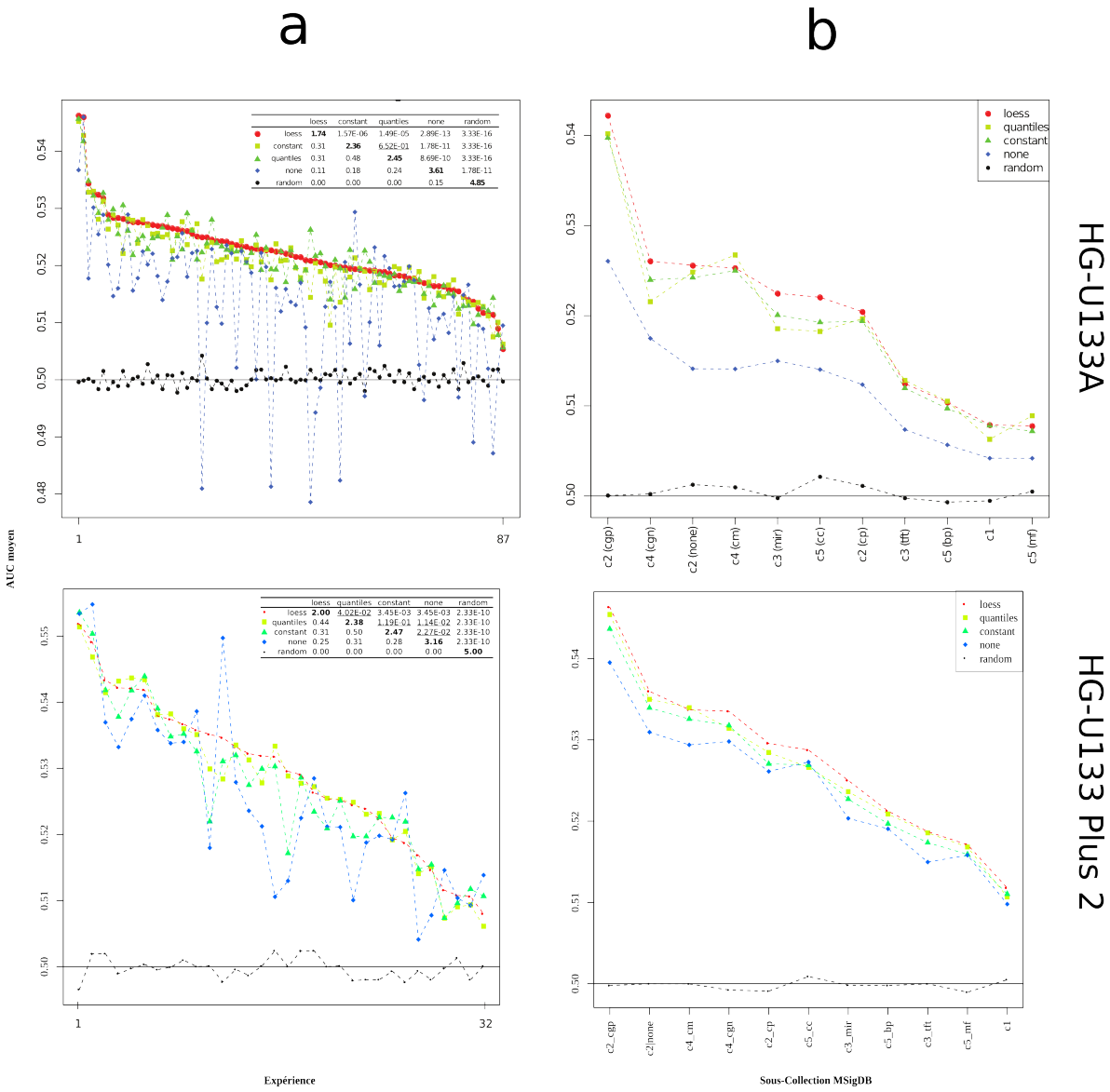


Figure 21 : Normalisation. cf. figure 20.

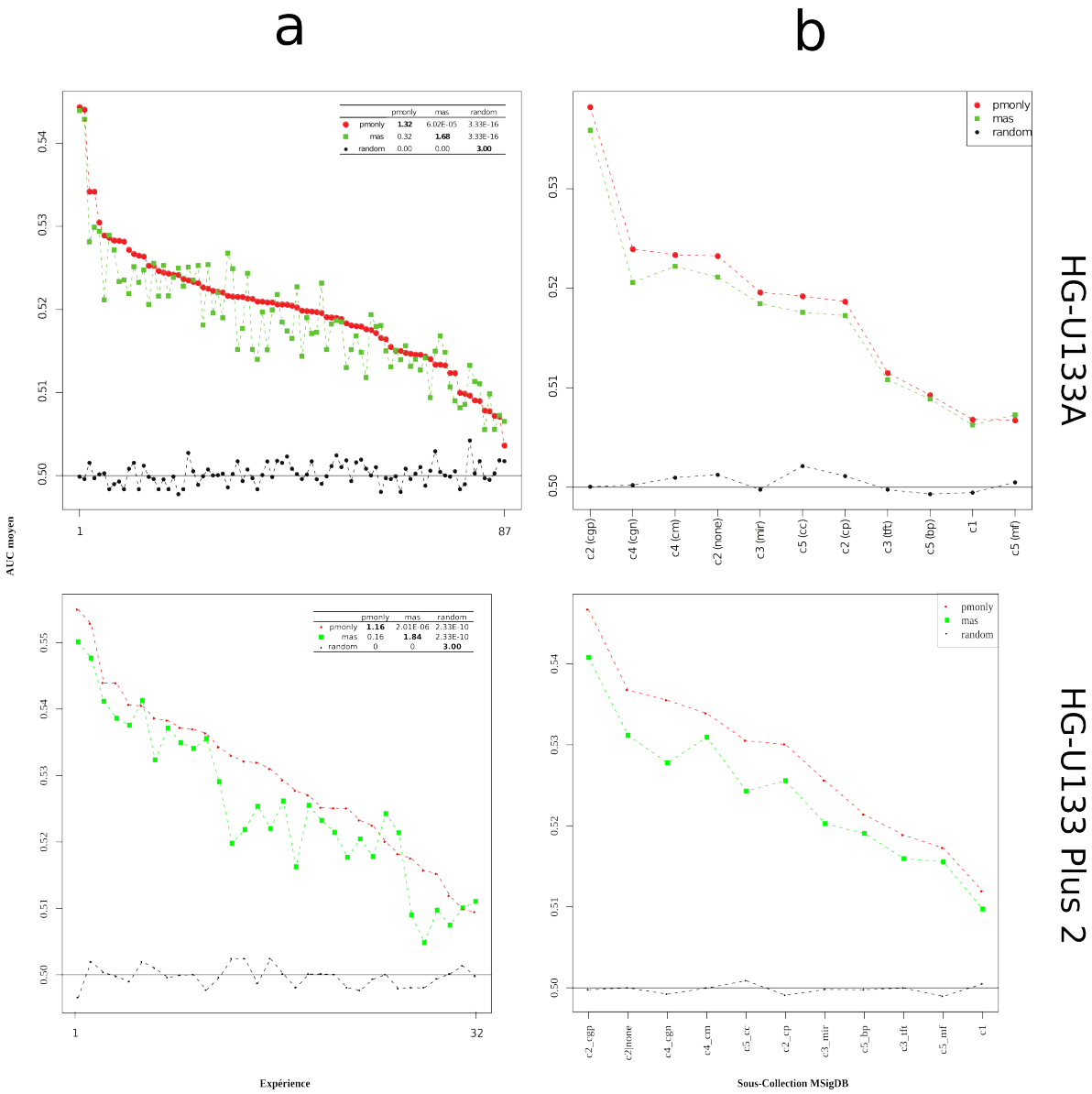


Figure 22 : Correction PM. cf. figure 20.

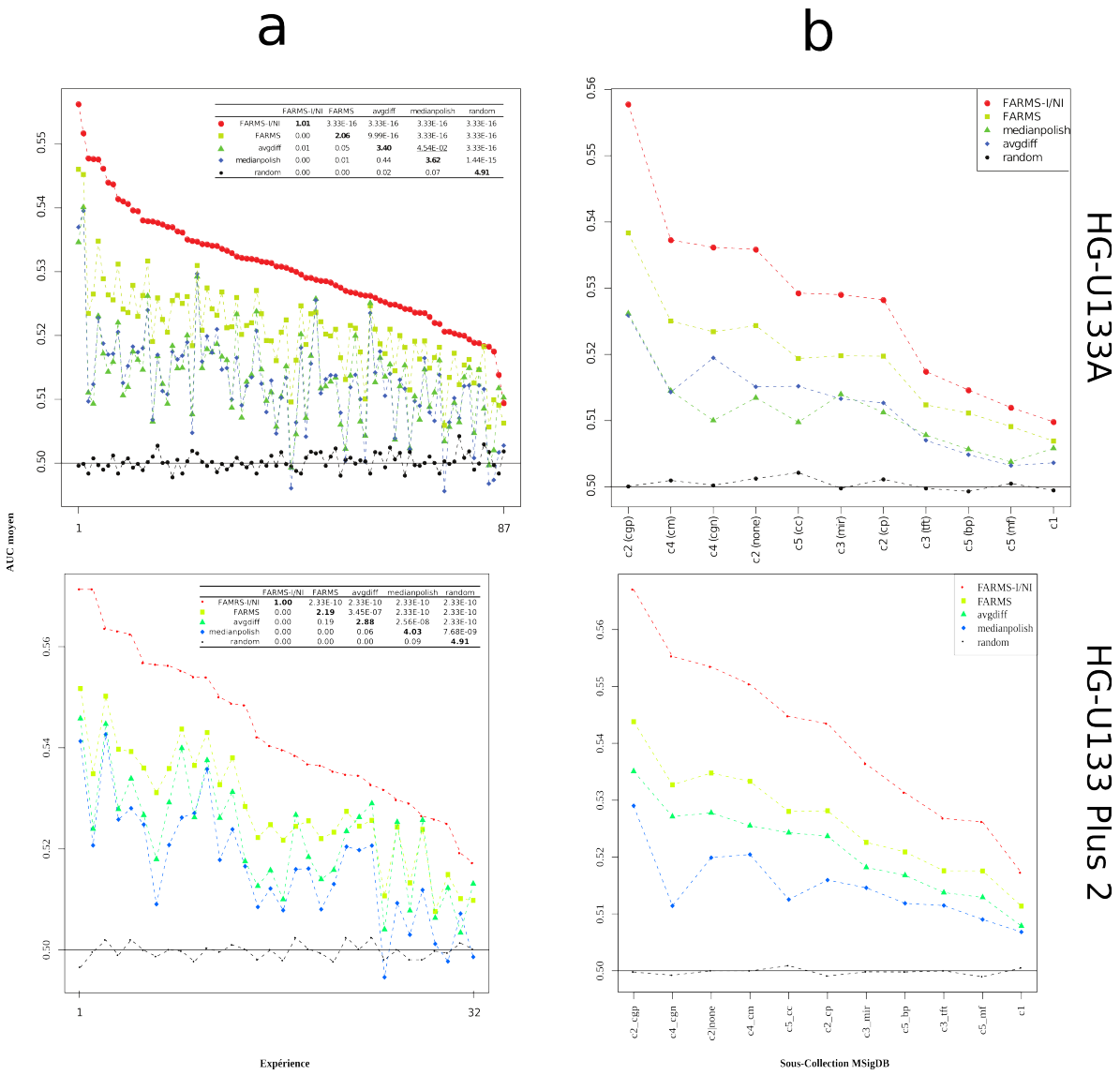
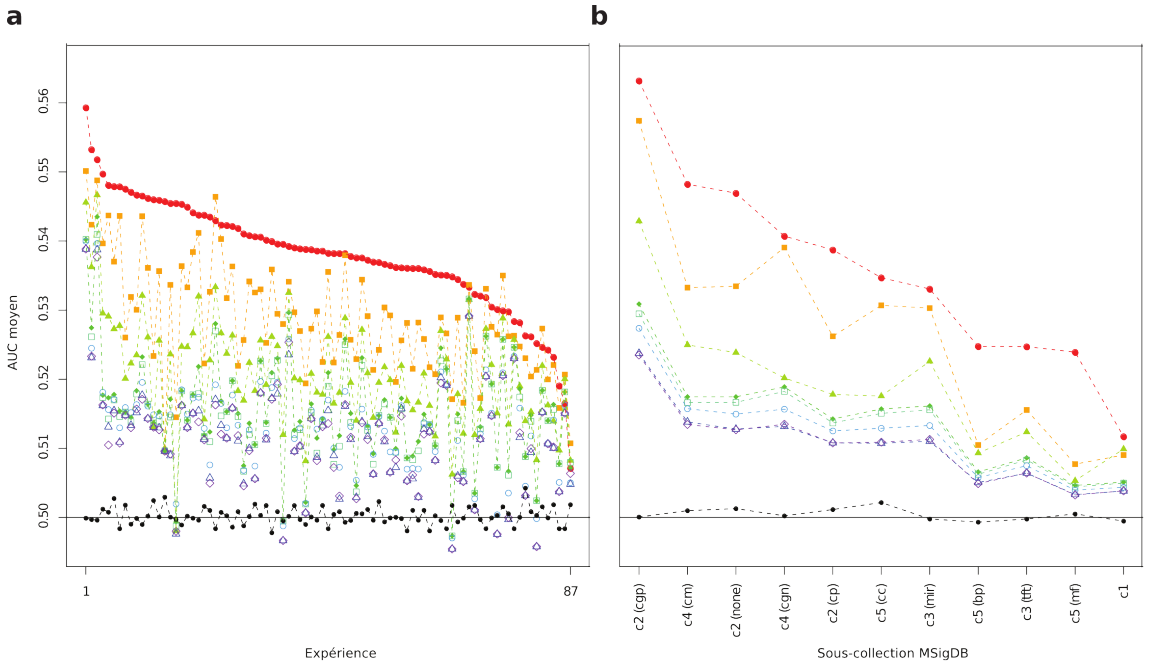
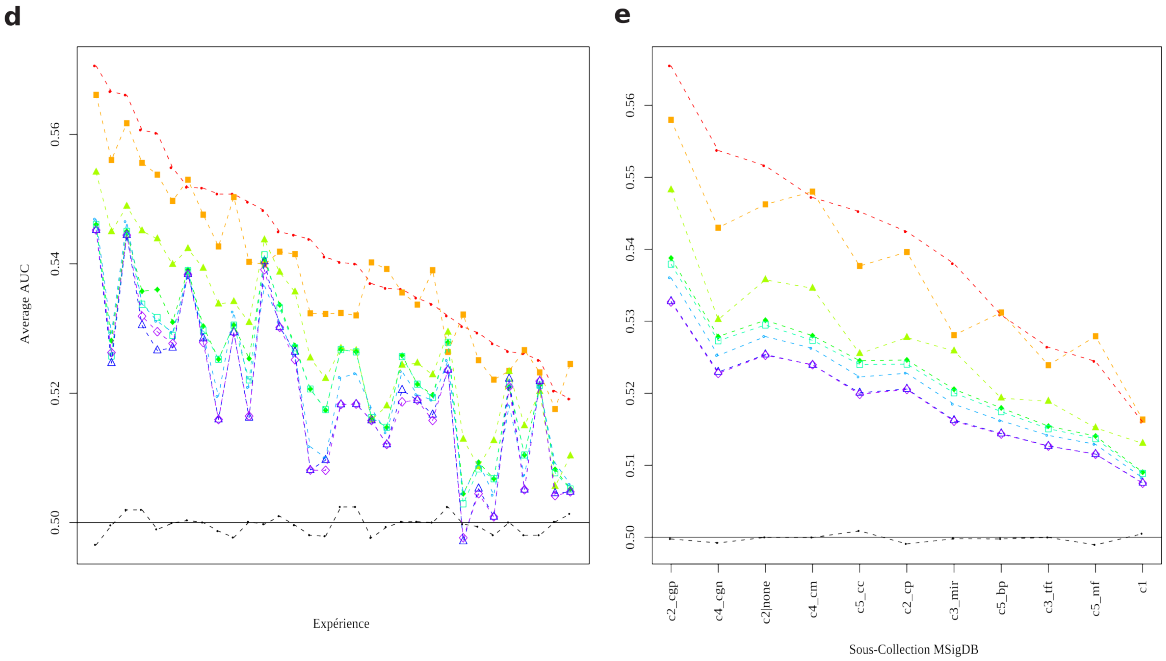


Figure 23 : Sommarisation. cf. figure 20.



c

	TREAT	FC	Cyber-T	ebayes	ebayes (lm)	SAMr	t	t (lm)	random
TREAT	**1.15**	7.99E-15	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16
FC	0.08	**1.99**	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16	3.33E-16
Cyber-T	0.02	0.01	**3.13**	8.88E-16	5.55E-16	5.55E-16	3.33E-16	3.33E-16	3.33E-16
ebayes	0.02	0.01	0.05	**4.30**	5.03E-10	2.58E-12	3.33E-16	3.33E-16	4.44E-16
ebayes (lm)	0.02	0.01	0.05	0.02	**5.08**	5.36E-07	5.55E-16	3.33E-16	4.44E-16
SAMr	0.00	0.01	0.03	0.14	0.31	**5.55**	3.33E-16	7.77E-16	5.55E-16
t	0.00	0.01	0.01	0.00	0.02	0.00	**7.53**	1.86E-01	2.22E-15
t (lm)	0.00	0.01	0.01	0.00	0.00	0.01	0.47	**7.57**	2.22E-15
random	0.00	0.00	0.01	0.03	0.03	0.03	0.09	0.09	**8.70**



f

	FC	TREAT	Cyber-T	ebayes	ebayes (lm)	SAMr	t (lm)	t	random
FC	**1.22**	1.53E-04	2.33E-10	2.33E-10	2.33E-10	2.33E-10	2.33E-10	2.33E-10	2.33E-10
TREAT	0.22	**2.00**	5.82E-09	1.16E-09	1.16E-09	2.33E-10	4.66E-10	2.33E-10	2.33E-10
Cyber-T	0.00	0.06	**3.41**	2.94E-07	1.77E-07	3.93E-08	1.63E-09	1.16E-09	2.33E-10
ebayes	0.00	0.06	0.16	**4.45**	2.32E-03	8.89E-05	3.19E-08	3.26E-09	2.33E-10
ebayes (lm)	0.00	0.06	0.13	0.18	**4.95**	1.03E-02	2.05E-08	1.00E-08	2.33E-10
SAMr	0.00	0.00	0.13	0.22	0.34	**5.53**	1.77E-05	2.62E-06	2.33E-10
t (lm)	0.00	0.03	0.03	0.06	0.06	0.13	**7.14**	2.29E-01	4.66E-10
t	0.00	0.00	0.03	0.06	0.06	0.09	0.35	**7.36**	4.66E-10
random	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	**8.94**

Figure 24 : Statistique de l'expression différentielle. cf. figure 20. (a, b, c) HG-U133A. (d, e, f) HG-U133 Plus 2.

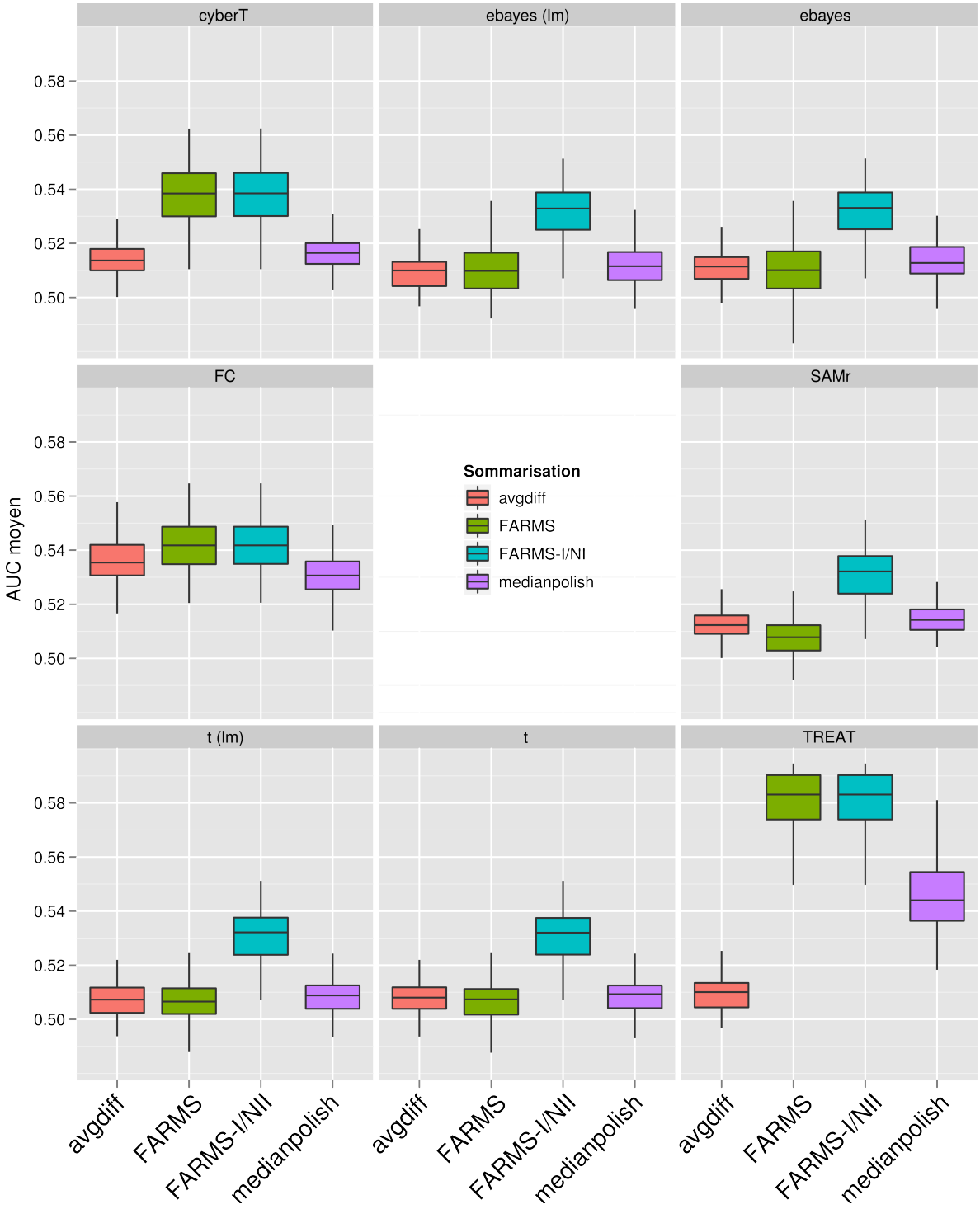


Figure 25: Boxplot comparant les AUC moyens pour les différentes méthodes de sommarisation et statistiques d'ordonnancements. La correction de fond, normalisation et correction PM sont respectivement

none, *loess* et *pmonly*, soit celles du pipeline d'AUC moyen maximal pour la plateforme HG-U133A. L'effet d'interaction entre le filtrage I/NI et les statistiques de type *t* et *ebayes* explique la supériorité de la sommarisation FARMS-I/NI observée à la figure 24.

Discussion et Conclusion

1. La « controverse » du fold-change

Cette section spécule à propos d'un paradigme différent et possiblement nouveau au problème de l'expression différentielle, débutant par la mise en situation suivante.

Supposons un instant qu'il existerait une expérience de microarrays, ou une expérience de toute autre technologie posant le problème des inférences multiples, comparant deux groupes, et que la réplication de cette expérience tende vers l'infini. Supposons, pour simplifier, que l'expression de chaque gène soit distribuée normalement avec une variance résiduelle (biologique) σ_j, affectée d'une fold-change (réel) β_j avec des nombres égaux de degrés de liberté pour chaque gène. La question de la « liste » des gènes différentiellement exprimés est alors plutôt facile à répondre : il ne suffit que de vérifier la condition $\beta_j \neq 0$ pour chaque gène et de retourner la liste correspondante. Mais qu'advient-il si l'on souhaite ordonner cette liste, par exemple, si en amont de l'expérience on ne peut n'utiliser que les 100 premiers gènes, pour validation par exemple? Il est évident que la notion de classement (ordonnancement) entre alors en jeu, ou du moins la notion de critère de classement. En fait, il semble être beaucoup plus utile d'aborder le problème de l'expression différentielle comme un problème général de classement plutôt que comme un problème de classification binaire, qui ne semble être qu'un cas particulier. La question du classement se poserait de toute façon dans un contexte de test : afin de fixer un seuil pour obtenir une classification en liste binaire, il faut bien d'abord classer les gènes selon une statistique échantillonnale quelconque, pour ensuite fixer (ou non) un seuil sur cette statistique

De toute évidence, puisque le problème n'est spécifié que par β et σ, le critère de classement ne dépendra que de ces deux variables, et prendra la forme fonctionnelle $f(\beta, \sigma)$ où f devrait normalement être une fonction continue. Il est probablement possible

d'aller plus en détail dans les formes fonctionnelles que pourrait assumer f, mais intuitivement deux types extrêmes de classements semblent pouvoir en résulter. Le premier est un classement par la « grandeur du signal » β. Ce type de classement sera noté Q_{FC}. Le second est un classement par le rapport signal/bruit (SNR), $\frac{\beta}{\sigma}$. Ce type de classement sera noté Q_{SNR}.

En principe, les deux types de questions sont potentiellement intéressantes, même si le biologiste exprimera probablement une préférence pour Q_{FC} : Entre un gène dont l'expression moyenne varie de 200% avec une variabilité biologique de 50%, et un gène variant de 2% avec une variabilité biologique de 0.01%, le premier cas est probablement le plus pertinent au phénotype étudié.

De plus, comme Q_{FC} et Q_{SNR} sont deux problèmes d'inférences différents, rien ne garantit que, étant donné le niveau de bruit et la faible réplication typiques des expériences de microarrays, notre aptitude à répondre à l'une ou l'autre des questions est la même. En d'autres mots, rien ne garantit que l'exactitude des estimés de $\frac{\beta}{\sigma}$, et par conséquent du classement résultant, sera la même que l'exactitude des estimés de β et de leur classement. Par exemple, les 100 premiers gènes de Q_{FC} ou de Q_{SNR} pour une expérience quelconque auraient beau être tous deux intéressants d'un point de vue biologique, l'exactitude de la réponse obtenue pour Q_{SNR} pourrait très bien être systématiquement inférieure à celle obtenue en réponse à Q_{FC} (et possiblement moins reproductibles).

Cette distinction cruciale entre les deux types de classements a déjà été discrètement soulignée en 2007 par Witten et Tibshirani[103]. Pour eux, le choix entre le fold-change et une statistique-t modérée revient en fait au choix entre deux estimateurs de deux critères de classement différents, soit le fold-change réel et le rapport signal bruit réel. Witten et Tibshirani démontrent, par une série de simulations simples (figure 26), que la statistique-t ordinaire est sous-optimale afin de retrouver le classement par rapport signal/bruit. Ils effleurent aussi l'idée que le fold-change échantillonal ne soit pas aussi la meilleure statistique pour retrouver le classement par fold-change réel.

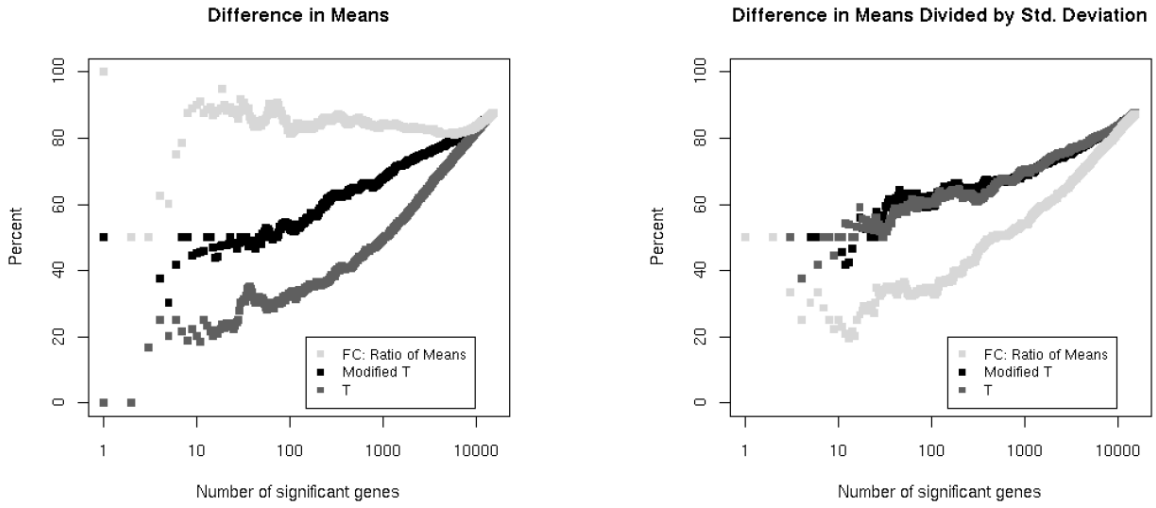


Figure 26 : Illustration de l'importance de définir quel type de classement, entre celui par fold-change réel et celui par rapport signal/bruit, est recherché. Tiré de la figure 3 de Witten et Tibshirani[103]. Sur des données simulées, le fold-change produit les meilleurs pourcentages d'exactitude (*accuracy*) pour le premier type de classement, la statistique modérée, pour le deuxième type. Il est aussi fortement intéressant de constater (les auteurs eux-mêmes ne le font pas) que l'exactitude du classement obtenu par fold-change réel est, même pour les meilleures statistiques, plus élevé dans le cas du classement par fold-change réel (par exemple 90 % à 100 gènes) que par SNR (60 % à 100 gènes).

C'est dans ce contexte que l'on peut tenter de clarifier la situation autour de la controverse et l'incertitude qui sévit encore à ce jour dans la communauté statistique de l'analyse des microarrays.

Les toutes premières publications de résultats de microarrays utilisant un seuil sur FC, la réaction de la communauté statistique fut évidemment de rappeler l'importance de la significativité. S'en suivit le développement d'un arsenal de statistiques plus aptes à tester $H_0 : \beta \neq 0$ que la statistique-t ordinaire, par exemple en tentant d'améliorer l'estimé en le régularisant. Toutefois, en passant de l'utilisation du simple FC vers une mesure de significativité, on se déplace de Q_{FC}, dont FC est l'estimateur à maximum de vraisemblance (MLE), vers Q_{SNR}, soit d'ordonner par SNR. En fait, il est remarquable que plus la régularisation est forte, plus la statistique-t régularisée tend vers le FC (par exemple en augmentant s_0 dans $\frac{FC_j}{s_j+s_o}$).

En 2006, les auteurs du MAQC jetèrent le proverbial pavé dans la marre en concluant qu'un classement par FC résulte en des listes de gènes plus reproductibles[104]. Du même coup, ils y recommandèrent, la pratique du « Volcano plot » soit un classement par FC accompagné d'un seuil permissif sur la valeur-p pour l'identification des gènes différentiellement exprimés[105,106]. D'autres auteurs ont par la suite proposé des statistiques basée sur le FC, c'est à dire qui prennent en compte la grandeur du changement, et non seulement sa significativité. Voir par exemple Kadota et al.[107], Deng et al.[108], Smyth et al.[100], Zhang et al.[109] et particulièrement Bickel et al.[110].

Une hypothèse qu'il serait intéressant d'explorer est que la plus grande reproductibilité des classements basés sur le FC s'explique par le fait qu'on tente alors d'estimer le fold-change réel (répondre à Q_{FC}) et que, considérant la faible puissance statistique des expériences typiques de microarrays, l'exactitude (et par conséquent la pertinence biologique) de ces estimés est supérieure à celle que l'on peut obtenir en tentant d'estimer le SNR avec un test d'hypothèse (répondre à Q_{SNR}). De plus, comme il fut déjà mentionné plus haut, la quantité d'intérêt biologique ici semble véritablement être la grandeur du changement. Comme plusieurs l'ont affirmé par le passé (par exemple récemment Nadon et al.[98]), il est vrai que le FC ne considère pas la variance, et qu'en standardisant par cette dernière, on ramène les FCs vers une métrique commune. Mais cela ne semble pas justifier le fait de se rabattre uniquement sur la valeur-p de ces tests statistiques comme critère de classement.

Finalement, il est opportun de mentionner que les résultats présentés pour ce mémoire s'accordent avec ces développements récents concernant l'importance de prendre en compte la grandeur du changement. En effet, même si les stratégies plus récentes, n'ont pas été testées, la supériorité de FC, de TREAT et l'effet spectaculaire de l'application d'un filtre non spécifique sur les AUC moyens sont sans équivoque. Le développement de nouvelles statistiques de l'expression différentielle reste un problème ouvert et l'outil de comparaison proposé, ou l'une de ses variantes, pourrait y jouer un rôle important.

2. Limites, Perspectives

Pertinence des signatures moléculaire. Dans sa version présentée pour ce mémoire, l'outil de comparaison calcule un AUC moyen sur toutes les signatures moléculaires ou encore sur un sous-ensemble défini. Cette approche est justifiée en avançant qu'une signature non pertinente, par définition, ne devrait pas favoriser une méthode plutôt qu'une autre puisque qu'elle n'est porteuse d'aucun «signal» de co-régulation ou de co-expression; son inclusion diminuant simplement le pouvoir de discrimination de la méthode de comparaison. Quoi qu'il en soit, il serait fort intéressant de reconduire l'analyse en utilisant une sélection de signatures jugées pertinentes à chaque expérience prise individuellement. Cette présélection pourrait être effectuée automatiquement, par exemple en ne retenant que quelques signatures dont l'enrichissement fait consensus parmi les méthodes comparées. Alternativement, la pertinence d'une signature moléculaire pourrait être décidée par un expert ou toute autre méthode de sélection biologiquement motivée.

L'AUC moyen sous l'hypothèse nulle. Même si le départage par expérience et sous-collections supportent avec robustesse les observations les plus importantes, il subsiste néanmoins un certain doute quant à l'effet de la dépendance (redondance) des signatures moléculaires et la possibilité d'un biais d'annotation de ces dernières. Par dépendance, on entend que le contenu de certaines signatures puisse se recouper avec d'autres. L'effet de la dépendance serait de gonfler la variance de l'AUC moyen sous l'hypothèse nulle de l'absence d'enrichissement : $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$. Par biais d'annotation, on entend que la valeur attendue de l'AUC moyen sous l'hypothèse nulle de l'absence d'enrichissement puisse ne pas être de 0,5. Plusieurs facteurs sont susceptibles de causer un tel biais. Par exemple, le design d'une puce pourrait contenir plusieurs sondes non spécifiques, donc non annotées. Étant peu exprimées, de telles sondes seraient alors moins susceptibles de générer des faux positifs, créant ainsi un biais à la hausse sur le rang moyen (donc l'AUC des signatures) des sondes annotées. La solution au problème de dépendance et de biais pourrait être d'employer une stratégie de ré-échantillonnage sur les signatures moléculaires. Les AUC pour une méthode donnée pourraient alors être comparés aux AUC

obtenus sur des signatures moléculaires produites par ré-échantillonnage afin d'obtenir des valeurs-p empiriques. Toutefois, l'impact réel de la dépendance et d'un éventuel de l'AUC moyen dans le contexte actuel de la *comparaison* des méthodes demanderait plus de réflexion. Si le biais et la violation de l'indépendance affectent également toutes les méthodes, importe-t-il vraiment de s'y attarder?

Étendre à d'autres plateformes, expériences et méthodes. Le nombre d'expériences de microarrays disponible dans les bases de données publiques augmentant de jour en jour, il serait possible et évidemment souhaitable de les intégrer dans une nouvelle itération du projet. Cette tâche n'est cependant pas triviale et demanderait un effort de curation considérable. Quant à l'inclusion de nouvelles méthodes, les orientations les plus prometteuses semblent être au niveau des nouvelles statistiques d'ordonnancement, de l'étape de filtre non spécifique et de l'effet des différentes transformations stabilisatrices de variances. Il serait aussi intéressant d'étudier l'impact du contrôle de qualité qui retirerait les quelques puces aberrantes plutôt que de les conserver. De plus, l'étude de la statistique d'ordonnancement, avec le filtre non spécifique sont les seules deux étapes qui ne sont pas limitées par la disponibilité des données brutes, le facteur le plus limitant en terme de nombre d'expériences disponibles. Aussi, afin de maintenir le temps de calcul et la taille des données raisonnables, les étapes de pré-traitement pourraient être fusionnées en solutions complètes.

Soulignons que l'utilité des données accumulées (les expériences, leur curation) et les calculs effectués (expression différentielle, enrichissement) ne se limitent pas qu'à la simple comparaison des méthodes d'analyse. Il est évident que les données elles-mêmes ont une valeur pour d'éventuelles applications d'extraction de données (*data mining*). Par exemple, le simple fait de savoir si tel gène est différentiellement exprimé ou telle voie moléculaire est activée dans telle ou telle expérience peut être d'une immense valeur pour un biologiste. En somme, le travail accompli pour le présent projet constitue la première étape vers la construction d'une base de données intégrative de données transcriptomiques.

3. En Conclusion

Ce mémoire a proposé une approche entièrement nouvelle à l'évaluation de différentes méthodes pour l'analyse de l'expression différentielle sur des données de puces à ADN. Cette approche repose sur l'hypothèse que la probabilité de co-expression différentielle de gènes associés dans la littérature (co-régulation, même fonction, etc.) est supérieure à celle de gènes choisis au hasard. Ainsi, les résultats produits par la meilleure méthode d'analyse devraient, par conséquent, le mieux refléter les associations tirées de la littérature (signatures moléculaires).

Une implémentation particulière de l'approche, utilisant les AUC moyens (un score d'enrichissement), a été testée sur un grand nombre d'expériences et de signatures moléculaires afin de comparer une sélection de pipelines d'analyse propres à la plateforme GeneChip d'Affymetrix. En plus de corroborer certaines observations précédentes à propos des étapes de correction de fond, normalisation et de sommarisation, les résultats obtenus appuient fortement l'idée controversée que le simple fold-change devrait être préféré aux statistiques-t régularisées telles que proposées actuellement (p.e. SAM, LIMMA). La supériorité de TREAT, une méthode basée sur un test-t avec seuil sur le fold-change, suggère aussi que la communauté statistique des microarrays devrait investir plus d'effort dans le développement et la promotion de statistiques priorisant la grandeur du changement plutôt qu'uniquement la significativité. Cela semble d'autant plus important que les données provenant de technologies qui succèderont (succèdent déjà) aux microarrays, telles que le séquençage haut débit, posent le même problème d'inférences multiples de faible puissance statistique.

Bibliographie

1. Voet, D. & Voet, J. *Biochimie*. (De Boeck: 2005).
2. Latchman, D. *Gene Regulation*. (Taylor & Francis: 2005).
3. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet* **21**, 33-37 (1999).
4. Kohane, I.S., Kho, A. & Butte, A.J. *Microarrays for an Integrative Genomics*. (The MIT Press: 2005).
5. Draghici, S. *Data Analysis Tools for DNA Microarrays*. (Chapman & Hall/CRC: 2003).
6. Gunderson, K.L. et al. Decoding Randomly Ordered DNA Arrays. *Genome Research* **14**, 870-877 (2004).
7. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat Genet*
8. Pease, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U.S.A* **91**, 5022-5026 (1994).
9. Mei, R. et al. Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A* **100**, 11237-11242 (2003).
10. Kohane, I.S., Kho, A. & Butte, A.J. *Microarrays for an Integrative Genomics*. (The MIT Press: 2005).
11. Shi, L. et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* **9**, S10-S10
12. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. (Springer: 2005).
13. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. (Springer: 2005).
14. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

15. Bittner, M.L., Chen, Y., Dorsel, A.N. & Dougherty, E.R. Microarrays: Optical Technologies and Informatics. (2001).at <http://adsabs.harvard.edu/abs/2001SPIE.4266.....B>
16. Dudoit, S. & Laan, M.J.V.D. *Multiple Testing Procedures with Applications to Genomics*. (Springer: 2010).
17. Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 47-47
18. Goeman, J.J., van de Geer, S.A., de Kort, F. & van Houwelingen, H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**, 93-99 (2004).
19. Mansmann, U. & Meister, R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med* **44**, 449-453 (2005).
20. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *The Annals of Applied Statistics* **1**, 107-129 (2007).
21. Wu, D. et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* (2010).doi:10.1093/bioinformatics/btq401
22. Irizarry, R.A., Wang, C., Zhou, Y. & Speed, T.P. Gene set enrichment analysis made simple. *Stat Methods Med Res* **18**, 565-575 (2009).
23. Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 47-47
24. Irizarry, R.A., Wang, C., Zhou, Y. & Speed, T.P. Gene set enrichment analysis made simple. *Stat Methods Med Res* **18**, 565-575 (2009).
25. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *The Annals of Applied Statistics* **1**, 107-129 (2007).
26. Rocke, D.M., Ideker, T., Troyanskaya, O., Quackenbush, J. & Dopazo, J. Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics* **25**, 701-702 (2009).

27. Shi, L. et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6 Suppl 2**, S12 (2005).
28. Wood, J.R. et al. Valproate-induced alterations in human theca cell gene expression: clues to the association between valproate use and metabolic side effects. *Physiol. Genomics* **20**, 233-243 (2005).
29. Tan, P.K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684 (2003).
30. Shi, L. et al. Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6**, S12-S12
31. De Hertogh, B. et al. A benchmark for statistical microarray data analysis that preserves actual biological and technical variance. *BMC Bioinformatics* **11**, 17 (2010).
32. Affymetrix - Latin Square Data. at <http://www.affymetrix.com/support/technical/sample_data/datasets.affx>
33. Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. & Halfon, M.S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* **6**, R16 (2005).
34. Murie, C., Woody, O., Lee, A.Y. & Nadon, R. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics.* **10**, 45 (2009).
35. Charting Pathways of Life. at <http://www.biocarta.com/>
36. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
37. Chen, Z. et al. A Distribution-Free Convolution Model for background correction of oligonucleotide microarray data. *BMC Genomics* **10 Suppl 1**, S19 (2009).
38. McGee, M. & Chen, Z. Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. *Stat Appl Genet Mol Biol* **5**, Article24 (2006).
39. Statistical algorithms description document. Technical report, Affymetrix, Santa Clara, CA. (2002).

40. Schadt, E.E., Li, C., Ellis, B. & Wong, W.H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem. Suppl* **Suppl 37**, 120-125 (2001).
41. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2**, RESEARCH0032 (2001).
42. Cleveland, W.S. & Devlin, S.J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* **83**, 596-610 (1988).
43. Tukey, J.W. *Exploratory Data Analysis*. (Addison Wesley: 1977).
44. Hochreiter, S., Clevert, D. & Obermayer, K. A new summarization method for affymetrix probe level data. *Bioinformatics* **22**, 943-949 (2006).
45. Talloen, W. et al. I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics* **23**, 2897-2902 (2007).
46. plier_technote.pdf. at <http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf>
47. Wu, Z., Irizarry, R.A., Gentleman, R., Martinez, F. & Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Dept. of Biostatistics Working Paper* **1**, (2003).
48. Lin, S.M., Du, P., Huber, W. & Kibbe, W.A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* **36**, e11-e11 (2008).
49. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364-374 (1997).
50. Huang, S. et al. At what scale should microarray data be analyzed? *Am J Pharmacogenomics* **4**, 129-139 (2004).
51. Durbin, B.P., Hardin, J.S., Hawkins, D.M. & Rocke, D.M. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18 Suppl 1**, S105-110 (2002).

52. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96-104 (2002).
53. Rocke, D.M. & Durbin, B. A model for measurement error for gene expression arrays. *J. Comput. Biol* **8**, 557-569 (2001).
54. Kutner, M., Nachtsheim, C., Neter, J. & Li, W. *Applied Linear Statistical Models*. (McGraw-Hill/Irwin: 2004).
55. K, S.G. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, (2004).
56. Zhang, S. & Cao, J. A close examination of double filtering with fold change and t test in microarray analysis. *BMC Bioinformatics* **10**, 402 (2009).
57. Long, A.D. et al. Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and A Bayesian Statistical Framework. *Journal of Biological Chemistry* **276**, 19937 -19944 (2001).
58. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A* **98**, 5116-5121 (2001).
59. McCarthy, D.J. & Smyth, G.K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25**, 765-771 (2009).
60. Bolstad, B. Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. (2004).
61. Lemieux, S. Probe-level linear model fitting and mixture modeling results in high accuracy detection of differential gene expression. *BMC Bioinformatics* **7**, 391 (2006).
62. Kim, S. & Volsky, D.J. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* **6**, 144-144
63. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).
64. Micklos, D. & Freyer, G.A. *DNA Science: A First Course, Second Edition*. (Cold Spring Harbor Laboratory Press: 2003).

65. BioCarta. at <http://www.biocarta.com/>
66. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucl. Acids Res.* **36**, D480-484 (2008).
67. Saunders, B. et al. The Molecule Pages database. *Nucl. Acids Res.* **36**, D700-706 (2008).
68. GSEA | MSigDB. at <http://www.broadinstitute.org/gsea/msigdb/index.jsp>
69. Abbud, R.A., Kelleher, R. & Melmed, S. Cell-Specific Pituitary Gene Expression Profiles after Treatment with Leukemia Inhibitory Factor Reveal Novel Modulators for Proopiomelanocortin Expression. *Endocrinology* **145**, 867-880 (2004).
70. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-378 (2003).
71. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res* **32**, D109-111 (2004).
72. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. *Nature* **434**, 338-345 (2005).
73. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet* **36**, 1090-1098 (2004).
74. Demeter, J. et al. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **35**, D766-770 (2007).
75. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol* **5**, e1000431 (2009).
76. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet* **31**, 19-20 (2002).
77. Su, A.I. et al. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 4465-4470 (2002).
78. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A* **95**, 14863-14868 (1998).

79. Brentani, H. et al. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. U.S.A* **100**, 13418-13423 (2003).
80. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A* **101**, 6062-6067 (2004).
81. Su, A.I. et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* **61**, 7388-7393 (2001).
82. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A* **98**, 15149-15154 (2001).
83. Choi, S. *Introduction to Systems Biology*. (Humana Press: 2007).
84. Davis, S. & Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846-1847 (2007).
85. Zhang, L., Yoder, S. & Enkemann, S. Identical probes on different high-density oligonucleotide microarrays can produce different measurements of gene expression. *BMC Genomics* **7**, 153 (2006).
86. Barrett, T. et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucl. Acids Res.* **37**, D885-890 (2009).
87. Brettschneider, J., Collin, F., Bolstad, B. & Speed, T. Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics* **50**, 264, 241 (2008).
88. Team, R.D.C. *R: A Language and Environment for Statistical Computing*. (Vienna, Austria, 2009).at <http://www.R-project.org>
89. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
90. O'Grady, E.P., Mulcahy, H., O'Callaghan, J., Adams, C. & O'Gara, F. Pseudomonas aeruginosa Infection of Airway Epithelial Cells Modulates Expression of Kruppel-Like Factors 2 and 6 via RsmA-Mediated Regulation of Type III Exoenzymes S and Y. *Infection and Immunity* **74**, 5893-5902 (2006).
91. Irizarry, R.A., Wu, Z. & Jaffee, H.A. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**, 789-794 (2006).

92. Lee, M.S., Hanspers, K., Barker, C.S., Korn, A.P. & McCune, J.M. Gene expression profiles during human CD4+ T cell differentiation. *Int. Immunol* **16**, 1109-1124 (2004).
93. Blalock, E.M. et al. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U.S.A* **101**, 2173-2178 (2004).
94. Xie, Y., Wang, X. & Story, M. Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics* **25**, 751 -757 (2009).
95. Kooperberg, C., Fazzio, T.G., Delrow, J.J. & Tsukiyama, T. Improved background correction for spotted DNA microarrays. *J. Comput. Biol* **9**, 55-66 (2002).
96. Shi, W., Oshlack, A. & Smyth, G.K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Research* (2010).doi:10.1093/nar/gkq871
97. de Leeuw, W., Rauwerda, H., Jonker, M. & Breit, T. Salvaging Affymetrix probes after probe-level re-annotation. *BMC Research Notes* **1**, 66 (2008).
98. Murie, C., Woody, O., Lee, A.Y. & Nadon, R. Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics.* **10**, 45 (2009).
99. Broberg, P. Statistical methods for ranking differentially expressed genes. *Genome Biol* **4**, R41 (2003).
100. McCarthy, D.J. & Smyth, G.K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics.* **25**, 765–771 (2009).
101. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U.S.A* **107**, 9546-9551 (2010).
102. Guo, L. et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol* **24**, 1162-1169 (2006).
103. Witten, D.M. & Tibshirani, R. A comparison of fold-change and the t-statistic for microarray data analysis. (2007).

104. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol* **24**, 1151-1161 (2006).

105. Fan, X. et al. Investigation of reproducibility of differentially expressed genes in DNA microarrays through statistical simulation. *BMC Proc.* **3**, S4 (2009).

106. Chen, J.J., Hsueh, H., Delongchamp, R.R., Lin, C. & Tsai, C. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics* **8**, 412 (2007).

107. Kadota, K., Nakai, Y. & Shimizu, K. Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. *Algorithms Mol Biol.* **4**, 7 (2009).

108. Deng, X., Xu, J., Hui, J. & Wang, C. Probability fold change: A robust computational approach for identifying differentially expressed gene lists. *Computer Methods and Programs in Biomedicine* **93**, 124-139 (2009).

109. Zhang, S. & Cao, J. A close examination of double filtering with fold change and t test in microarray analysis. *BMC Bioinformatics* **10**, 402 (2009).

110. Montazeri, Z., Yanofsky, C.M. & Bickel, D.R. Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: applications to differential gene expression. *Stat Appl Genet Mol Biol* **9**, Article23 (2010).

Annexe 1 : Liste des expériences

Plateforme	Accession GEO	Nombre de... puces	groupes	contrastes	Réplication des groupes Min.	Max.	Méd.	Titre de l'expérience	PMID
Affymetrix Human Genome U133A (GPL96)	GDS1050	13	3	3	4	5	4	Valproic acid effect on theca cells	15598877
	GDS1062	27	3	3	5	14	8	Squamous cell carcinoma of the oral cavity with lymph node metastasis	15558013
	GDS1064	42	7	21	2	10	7	Acute myeloid leukemia subclasses	15674361
	GDS1065	15	4	6	3	4	4	Mitochondrial encephalomyopathies associated with mitochondrial DNA mutations	15728662
	GDS1067	52	3	3	6	39	7	Plasma cell dyscrasias	15735737
	GDS1204	18	6	15	3	3	3	Lung cancer cell line response to motexafin gadolinium: time course	15867382
	GDS1230	18	4	6	4	5	4.5	Hematopoietic stem cell and progenitor cell comparison	16089502
	GDS1286	6	3	3	2	2	2	Esophageal cell response to low pH: time course	16113055
	GDS1288	5	2	1	2	3	2.5	Mesenchymal precursor cells derived from embryonic stem cells	15971941
	GDS1317	8	4	6	2	2	2	Umbilical vein endothelial cell response to shear stress and intraluminal pressure	16155357
	GDS1321	24	3	3	8	8	8	Barrett's metaplasia progression to adenocarcinoma	15833844
	GDS1329	49	3	3	6	27	16	Molecular apocrine breast tumors	15897907
	GDS1362	25	2	1	10	15	12.5	Ischemic and nonischemic cardiomyopathy comparison	15769906
	GDS1380	18	2	1	9	9	9	Glioblastoma pseudopalisading cells	16254489
	GDS1384	8	4	6	2	2	2	c-Myb and its oncogenic variant v-Myb	16205643
	GDS1390	20	2	1	10	10	10	Prostate cancer progression after androgen ablation	16203770
	GDS1476	22	2	1	7	15	11	Uterine fibroids with fumarate hydratase mutations	16319128
	GDS1479	60	6	15	5	15	11	Carcinoma in situ lesions of the urinary bladder	15173019
	GDS1503	18	6	15	3	3	3	Hutchinson-Gilford progeria syndrome: fibroblast	15268757
	GDS1542	8	2	1	4	4	4	Tumor necrosis factor effect on macrovascular umbilical vein endothelial cells	16617158
	GDS1543	6	2	1	3	3	3	Tumor necrosis factor effect on microvascular endothelial cells	16617158
	*GDS1556	25	2	1	5	20	12.5	Atrial and ventricular myocardium comparison	15877233
	*GDS1558	35	3	3	5	20	10	Permanent atrial fibrillation	15817885
	*GDS1574	46	12	66	3	4	4	Endotoxin effect on leukocytes: time course	16136080
	GDS1584	20	2	1	4	16	10	Oral squamous cell carcinoma	15381369
	GDS1618	20	2	1	10	10	10	Lymph node and tonsil comparison	16440291
	GDS1630	16	8	28	2	2	2	Immortalized endothelial cell line response to atorvastatin	16575254
	GDS1637	10	5	10	2	2	2	Oncogene-induced sensescence in vitro model	16079833
	GDS1663	25	8	28	3	4	3	Expression data from different research centers	ND
	GDS1672	16	5	10	2	5	2	Keratinocyte stem cell–enriched hair follicle bulge cells	16395407
	GDS1688	29	3	3	9	10	10	Various lung cancer cell lines	16813650
	GDS1736	8	2	1	4	4	4	Arachidonic acid effect on prostate cancer cells	16452198
	GDS1758	12	2	1	4	8	6	Pterygium	16488932
	GDS1780	5	2	1	2	3	2.5	Colorectal cancer progression: polysomal mRNA profiles	16531451
	GDS1873	18	6	15	3	3	3	Antiestrogen and aromatase inhibitor effect on breast cancer cells	15831674
	GDS1902	26	13	78	2	2	2	Cyanobacterial metabolite apratoxin A cytotoxic effect on colon adenocarcinoma Cells: time course and dose response	16474387
	GDS1975	85	4	6	7	59	9.5	Gliomas of grades III and IV	15374961
	GDS2014	8	2	1	3	5	4	Ulcerative colitis	ND
	GDS2021	9	3	3	3	3	3	Coronary smooth muscle cell response to beta-1 receptor blockers metoprolol and nebivolol	17467819
	GDS2057	18	6	15	3	3	3	Androgen receptor modulator effect: time course	16574741
	GDS2084	15	2	1	7	8	7.5	Polycystic ovary syndrome: adipose tissue	17062763
	GDS2090	6	2	1	3	3	3	Sphingosine 1-phosphate effect on glioblastoma cells	16901352
	GDS2095	24	8	28	3	3	3	Glucocorticoid receptor activation effect on breast cancer cells: time course	16690749
	GDS2142	19	3	3	4	10	5	Cystic fibrosis patients with mild and severe lung Disease: nasal respiratory epithelium	16614352
	GDS2153	9	2	1	4	5	4.5	Dermatomyositis	16504012
	GDS2175	8	3	3	2	3	3	Cockayne syndrome group B protein-null fibroblast rescue	16772382
	GDS2190	61	2	1	30	31	30.5	Bipolar disorder: dorsolateral prefrontal cortex	16894394
	GDS2191	21	2	1	10	11	10.5	Bipolar disorder: orbitofrontal cortex	16894394
	GDS2201	37	2	1	8	29	18.5	Serrated and conventional adenocarcinomas	16819509
	GDS2205	12	2	1	5	7	6	Dilated cardiomyopathy: left ventricle	17045896
	GDS2241	7	2	1	3	4	3.5	Trophoblast cell lines	16797695
	GDS2245	18	2	1	7	11	9	Uterine fibroids with mutated fumarate hydratase (I)	16319128
	GDS2246	10	3	3	2	5	3	Uterine fibroids with mutated fumarate hydratase (II)	16319128
	GDS2287	6	3	3	2	2	2	Airway epithelial cell response to Pseudomonas aeruginosa rsmA mutant infection	16988269
	GDS2332	11	4	6	2	3	3	Neisseria meningitidis non-adherent mutant infected umbilical vein endothelial cells	16958858
	GDS2333	12	6	15	2	2	2	Neisseria meningitidis non-adherent mutant infected umbilical Veinendothelial cells: time course	16958858
	GDS2362	71	4	6	12	22	18.5	Presymptomatic and symptomatic malaria: peripheral blood mononuclear cells	16988231
	GDS2366	24	8	28	3	3	3	Preadipocytes from anatomically separate fat depots	16985259

Affymetrix Human Genome U133A (GPL96)	GDS2381	17	3	3	5	6	6	Atopic dermatitis	17181634
	GDS2617	22	3	3	3	14	5	Tumorigenic breast cancer cells	17229949
	GDS2643	56	6	15	5	12	10	Waldenstrom's macroglobulinemia: B lymphocytes and plasma cells	17252022
	GDS2649	40	8	28	5	5	5	HIV infection effect on CD4+ and CD8+ T cells	17251300
	GDS2655	28	2	1	14	14	14	Fetal and adult reticulocytes	17405831
	GDS266	29	5	10	2	13	5	Asthma and atopy	18269679
	GDS268	24	3	3	8	8	8	Obesity and fatty acid oxidation	16849634
	GDS2705	8	4	6	2	2	2	PPARgamma agonist and platinum-based drug effect on adenocarcinoma cell line	17482130
	GDS2744	6	2	1	3	3	3	Dioxin effect on breast cancer cell line	17517823
	GDS2852	15	5	10	3	3	3	Interleukin 13 effect on bronchial cell line: time course	ND
	GDS289	7	2	1	3	4	3.5	Chronic obstructive pulmonary disease	ND
	GDS493	11	3	3	3	5	3	Cystic fibrosis pathology and 4-phenylbutyrate	14583596
	GDS505	17	2	1	8	9	8.5	Renal clear cell carcinoma	14641932
	GDS525	5	2	1	2	3	2.5	Extraocular and limb muscle comparison	15855387
	GDS534	75	3	3	18	34	23	Smoking-induced changes in airway transcriptome	15210990
	GDS556	8	4	6	2	2	2	Extraocular muscle layer profiles	15326121
	*GDS558	38	6	15	5	8	6	Myocardial remodeling in response to LVAD	15326121
	GDS737	30	2	1	12	18	15	Lung tissue from smokers with severe emphysema	15374838
	GDS738	11	3	3	3	4	4	Intervertebral disc cells and osmotic loading	16133915
	*GDS749	22	2	1	10	12	11	Sarcopenia expression profiling	15687482
	GDS756	6	2	1	3	3	3	Colon cancer progression	16531451
	*GDS760	29	3	3	9	10	10	T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma comparison	16358311
	GDS785	15	6	15	2	3	2.5	CD4+ T cell differentiation	15210650
	GDS810	31	4	6	7	9	7.5	Alzheimer's disease at various stages of severity	14769913
	GDS885	7	2	1	3	4	3.5	Tumor cell response to topoisomerase poison camptothecin	15026349
	*GDS914	24	8	28	3	3	3	Metabolic syndrome response to exercise intervention	15347626
	GDS962	14	3	3	4	5	5	Peripheral blood mononuclear cells and the effect of exercise	15194674
	GDS992	6	2	1	3	3	3	Endoplasmic reticulum membrane-associated genes in breast cancer cell line MCF-7	15574777
	GDS999	34	2	1	7	27	17	Bronchoalveolar lavage cells of lung transplant recipients with acute rejection	12958056
Affymetrix Human Genome U133 Plus 2.0 (GPL570)	GDS1369	4	2	1	2	2	2	Autophagy effect on the MHC class II antigenic peptide repertoire	15894616
	GDS1427	6	3	3	2	2	2	c-Myb and oncogenic variant v-Myb transcriptional activities	16205643
	GDS1685	4	2	1	2	2	2	Hereditary gingival fibromatosis-	ND
	GDS1807	4	2	1	2	2	2	Hypoxia effect on B lymphocyte cell line	16517405
	GDS2125	11	4	6	2	3	3	Methyl-CpG-binding protein 2 binding disruption during neuronal maturation	16682435
	GDS2250	47	4	6	2	20	12.5	Basal-like breast cancer tumors-	16473279
	GDS2414	14	5	10	2	3	3	Decidual stromal cell response to trophoblast conditioned medium: time course	17021345
	GDS2611	25	9	36	2	3	3	Interleukin-20 subfamily cytokines effect on epidermal keratinocytes	17277128
	GDS2653	6	3	3	2	2	2	Sequestosome 1 and dipeptidylpeptidase III overexpression Effect on neuroblastoma cell line	17360324
	GDS1237	6	2	1	3	3	3	Promyelocytic cell response to Anaplasma phagocytophilum infection	16005178
	GDS2052	27	5	10	3	8	6	Endometrium throughout the menstrual cycle	16306079
	GDS2089	6	2	1	3	3	3	Skeletal muscle response to weight loss	16849634
	GDS2221	12	4	6	3	3	3	Galectin-1 maturative effect on monocyte-derived dendritic cells	16785517
	GDS2307	9	3	3	3	3	3	Oxidatively modified LDL effect on retinal pigment epithelial cell line	18182060
	GDS2499	12	4	6	3	3	3	Anti-cancer agent sapphyrin PCI-2050 effect on lung cancer cell line: dose response	17233922
	GDS2628	6	2	1	3	3	3	Vitamin D effect on bronchial smooth muscle cells	17213369
	GDS2724	15	5	10	3	3	3	Transcriptional regulators Bmi-1 and Mel-18 depletion Effect on medulloblastoma cell line	17452456
	GDS2728	6	2	1	3	3	3	CD133+ and CD133- glioblastoma-derived cancer stem cell lines	17483311
	GDS2737	37	6	15	3	9	6	Endometriosis	17510236
	GDS1869	8	2	1	4	4	4	Heregulin and forskolin mitogenic effect on cultured Schwann cells	ND
	GDS2652	12	3	3	4	4	4	Neonatal brain response to docosahexaenoic acid supplemented formulas	17426818
	GDS2083	10	2	1	5	5	5	Limb immobilization effect on skeletal muscle	16763108
	GDS2635	30	4	6	5	10	7.5	Invasive ductal and lobular breast carcinomas	17389037
	GDS1439	19	3	3	6	7	6	Prostate cancer progression	16286247
	GDS1579	18	3	3	6	6	6	Endotoxin effect on leukocytes: time course (U133 2.0)	16136080
	GDS1732	14	2	1	7	7	7	Papillary thyroid cancer	AACR Abstracts 2006, 1:463
	GDS2432	16	2	1	7	9	8	Pituitary adenoma predisposition: whole blood	16728643
	GDS2697	21	2	1	8	13	10.5	Teratozoospermia (HG-U133 2.0)	17327269
	GDS2418	18	2	1	9	9	9	Vulvar intraepithelial neoplasia	17471573
	GDS2609	22	2	1	10	12	11	Early onset colorectal cancer: normal-appearing colonic mucosa	17317818
	GDS1917	28	2	1	14	14	14	Cerebellar cortex in schizophrenia	ND
	**GDS2204	12	4	6	2	6	2	Analysis of 1, 2, 5 or 10 ug of Stratagene universal human reference RNA as starting material. Results provide insight into the influence of starting RNA quantity on the subsequent gene expression signal measured from a microarray.	16776839

* Expérience retirée plus tard par GEO.

** Pas de biologie sous-jacente. Utilisé comme contrôle à la méthodologie

Annexe 2 : Modifications apportées aux expériences

Accession GEO	Modification au design expérimental
GDS1064	Le niveau factoriel "acute erythroid leukemia" n'est pas répliqué, le microarray correspondant (GSM30312) est donc exclu de l'analyse.
GDS1286	Les fichiers .CEL pour GSM38763, GSM38764, GSM38771, GSM38772 ne sont pas fournis.
GDS1288	Le niveau factoriel "bone marrow-derived MS" n'est pas répliqué, le microarray correspondant (GSM38627) est donc exclu de l'analyse.
GDS1321	Le facteur "individual" est ignoré.
GDS1362	Les fichiers CEL pour GSM33109, GSM33110, GSM33111, GSM33112, GSM33113, GSM33069, GSM33088, GSM33089, GSM33087, GSM33091, GSM33090 ne sont pas fournis.
GDS1556	GSM40995 est retiré, car identique à GSM40994
GDS1558	GSM40995 est retiré, car identique à GSM40994
GDS1574	Le facteur"Subjects" est ignoré.
GDS1780	Le microarrays GSM47876 a des intensités manquantes, il est donc exclu de l'analyse.
GDS1902	La combinaison de facteurs "12 h" et "200 J per m2" n'est pas répliquée, le microarray correspondant (GSM92815) est donc exclu de l'analyse.
GDS2014	Le facteur "Gender" est ignoré.
GDS2095	GDS2096 et GDS2097 ont été fusionnés avec cette expérience (réplicats biologiques)
GDS2241	Le fichier .CEL pour GSM48273 n'est pas fourni.
GDS2381	Le facteur "individual" est ignoré.
GDS505	Le facteur "individual" est ignoré.
GDS558	Le facteur "individual" est ignoré.
GDS738	Le facteur "age" est ignoré.
GDS760	GSM27088 est identique à GSM27087.
GDS914	GSM19162, GSM20659 sont identiques.
GDS962	Le fichier .CEL GSM38763 n'est pas fourni.
GDS1369	Le niveau factoriel “6 hour” est retiré, car non répliqué.
GDS2125	Le fichier GSM102825.CEL est endommagé donc retiré de l'analyse.
GDS2089	Le facteur “patient” est ignoré.
GDS2083	Le facteur “subject” est ignoré.
GDS1579	Le facteur “subject” est ignoré.