From microarray data acquisition to their interpretation: the importance of raw data processing.

Nolwenn Le Meur

UNIVERSITE DE NANTES
FACULTE DE MEDECINE

De l'Acquisition des Données de Puces à ADN vers leur Interprétation : Importance du Traitement des Données Primaires

THESE DE DOCTORAT

Ecole Doctorale CHIMIE BIOLOGIE
Discipline Sciences de la Vie et de la Santé
Spécialité Bio-informatique

Présentée
et soutenue publiquement par

LE MEUR Nolwenn

Le 13 juin 2005, devant le jury ci-dessous

Président	**M. ESCANDE Denis**, Professeur, Université de Nantes
Rapporteurs	**M. FROIDEVAUX Christine**, Professeur, Université Paris Sud
	M. LOREAL Olivier, Chargé de Recherche, Université de Rennes I
Examinateur	**M. GENTLEMAN Robert**, Professeur, Université de Washington
Directeur de thèse	**M. LEGER Jean**, Directeur de Recherche, Université de Nantes

TITRE : De l'acquisition des données de puces à ADN vers leur interprétation : importance du traitement des données primaires.

Les outils bio-informatiques sont devenus indispensables aux traitements et à l'analyse des données de puces à ADN. De l'extraction des données primaires par les logiciels d'analyse d'images à la recherche des réseaux moléculaires en passant par la normalisation et la validation des données, les méthodes mathématiques et statistiques sont incontournables. Ce travail s'intéresse aux méthodes d'analyse d'images des puces à ADN, à la métrologie et à la transformation des données primaires en données « consolidées ». Au cours de cette étude, un service Web nommé MADSCAN (*MicroArray Data Suite of Computed Analysis)* a été développé pour traiter les données primaires d'expression. Cet outil permet de filtrer, normaliser et valider statistiquement les données primaires. Les enjeux suivants sont l'analyse (mise en évidence des gènes d'intérêts, méthodes de classifications) et l'intégration des données d'expression avec les méta-données (ontologies, littérature…) pour une meilleure compréhension des mécanismes de fonctionnement des gènes.

MOTS CLES Puce à ADN, Transcriptome, Bio-informatique, Traitement des données, Analyse des données, Intégration des données, langage R

TITLE: From microarray data acquisition to their interpretation: the importance of raw data processing.

Bioinformatic tools have become essential for microarray processing. From raw data extraction by image analysis software to the research of molecular networks via data normalization and validation, mathematical and statistical methods are crucial. This work deals with the performance of microarray image analysis software, with metrology and the transformation of raw microarray data to consolidated gene expression data matrices. A Web service, MADSCAN (MicroArray Data Suite of Computed Analysis) has been developed to process raw expression data. This tool filters, normalizes, and statistically validates raw data. The next challenges are the analysis (molecular marker identification, classification approaches) and the integration of expression data with metadata (ontology, literature…) for a better understanding of gene mechanisms.

KEYWORDS :Microarray, Transcriptome, Bioinformatic, Data processing, Data analysis, Data integration, R language

LE MEUR Nolwenn

Unité INSERM U533

Faculté de Médecine

1, rue Gaston Veil

44035 Nantes cedex 1, France

J'exprime toute ma reconnaissance à Monsieur Jean Léger pour m'avoir accueillie il y a quatre ans au sein de l'équipe Génomique fonctionnelle de l'unité INSERM U533, pour la confiance qu'il m'a accordé et sa passion pour la recherche qu'il a su partager.

Je tiens à adresser mes plus sincères remerciements à Monsieur Olivier Loréal et Madame Christine Froidevaux, pour avoir accepté de juger et critiquer ce travail en assumant les rôles de rapporteurs ; ainsi qu'à Monsieur Robert Gentleman d'assumer le rôle d'examinateur, et Monsieur Denis Escande de présider ce jury.

Je remercie chaleureusement Audrey Bihouée, Raluca Teusan, Guillaume Lamirault, et Gérard Ramstein pour leurs précieux conseils et leur participation essentielle à ce travail.

Je tiens également à remercier Martine Le Cunff, Isabelle Guisle, Marja Steenman et Hélène Bédrine-Ferran pour leur soutien en ayant accepté d'être les « cobayes » de mes premiers essais en bio-informatiques.

Merci à Catherine Chevalier, Lise Caron et Alice Le Bars pour les discussions que nous avons pu avoir.

Enfin, je tiens également à remercier tous les membres de l'unité U533 pour leur accueil chaleureux et leur sympathie.

Table des matières

AVANT PROPOS

AVANT PROPOS

Ces travaux ont été effectués sous la direction de Jean Léger, responsable de l'équipe « Génomique fonctionnelle » au sein de l'unité INSERM U533 et de la plate-forme puce à ADN d'OUEST-genopole®, à Nantes.

Depuis 1999, l'équipe développe et pratique la génomique fonctionnelle dans le domaine cardiovasculaire et neuromusculaire. Les puces à ADN sont l'outil privilégié pour cette stratégie globale d'analyse de la complexité du fonctionnement du muscle strié (squelettique et cardiaque) et de quelques unes de ses perturbations pathologiques sur le plan transcriptomal. En effet, les puces à ADN permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Cependant, cette technologie génère une grande diversité de données qui implique un important travail de bio-informatique. Aussi, un grand nombre de techniques liées à l'informatique sont nécessaires à l'analyse des données issues de cette technologie : analyse d'images, stockage et gestion des informations, techniques de normalisation, analyses statistiques, représentation graphiques, techniques d'extraction de connaissances…Au cours de mon DEA et de ma thèse, je me suis attachée à mettre en place l'ensemble des procédures indispensables à l'acquisition et à la validation des mesures issues des expériences de puces à ADN.

Ce manuscrit se compose de trois parties. Le premier chapitre est une introduction consacrée au développement de la génomique fonctionnelle, à la technologie des puces à ADN et à l'avènement de la bio-informatique dans ce domaine. La seconde partie décrit les procédures que j'ai mises en place et l'outil informatique que j'ai développé pour le traitement et la transformation des données issues des images de puces à ADN en données « consolidées ». Le dernier chapitre présente une étude bibliographie sur les principales méthodes d'analyse de données d'expression à savoir les techniques statistiques pour la mise en évidence des gènes d'intérêt, les méthodes de classification des données pour regrouper les gènes et/ou les échantillons suivant leur profil d'expression et les outils pour l'intégration des données du transcriptome avec d'autres informations biologiques.

Chapitre I.

Bases Biologiques et Bio-Informatiques

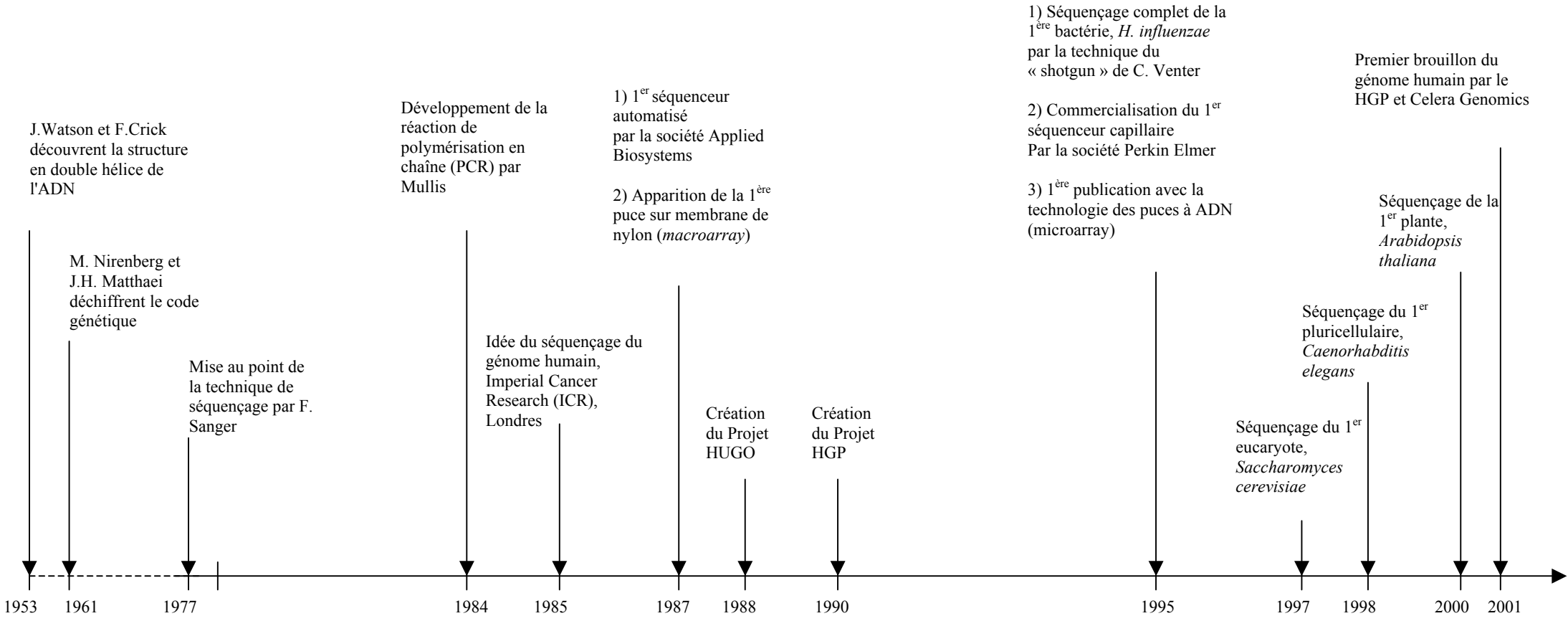


Figure 1. Avènement de la génomique et du séquençage des génomes (source : Infobiogen, http://www.infobiogen.fr)

Chapitre I. Bases biologiques et bio-informatiques

I. Avènement de la génomique fonctionnelle et puces à ADN

Le génie génétique comprend l'ensemble des techniques de la biologie moléculaire. Cette discipline s'est structurée et définie autour de la « traduction » de l'information génétique et de ses mécanismes de régulation. Depuis la découverte de la structure de la molécule d'ADN par F. Crick et J. Watson en 1953 et le décryptage du code génétique par M. Nrenberg et J.H. Matthaei en 1961, la biologie moléculaire a vu son histoire s'accélérer (Fig. 1). Les techniques de digestion enzymatique ont notamment permis d'isoler, cloner et séquencer les gènes.

1. Séquençage des génomes

1.1 Historique

En 1977, deux techniques de séquençage des acides nucléiques apparaissent à peu près simultanément : la méthode enzymatique de Frédérick Sanger et l'approche chimique de Walter Gilbert et Allan Maxam. La première, grâce aux connaissances qui seront acquises sur les enzymes, va prendre le pas sur la seconde (trop toxique). Dans les laboratoires, chacun se met à séquencer son « gène ». En 1984, la mise au point de la technique d'amplification génétique, ou PCR (***P**olymerase **C**hain **R**eaction*), est un progrès technique important pour le développement des méthodes de séquençage (Mullis *et al.*, 1986). Cette technique, qui permet d'amplifier sélectivement toute séquence d'ADN, devient rapidement un outil puissant et indispensable au séquençage des génomes. En 1985, à l'*Imperial Cancer Research* de Londres, naît pour la première fois l'idée de décrypter les trois milliards de bases du génome humain. L'objectif du déchiffrage de notre « patrimoine génétique » et ses retombées scientifiques et médicales annoncées (fonctionnement de l'organisme, évolution, diagnostic génétique et thérapie géniques, nouveaux médicaments...) décident rapidement les parlementaires du Congrès américain à affecter les 100 ou 200 millions de dollars annuels nécessaires à ce projet.

En 1988, la fondation *Human Genome Organization* (HUGO) est créée afin de coordonner les efforts de cartographie et de séquençage entrepris dans le monde. Cependant, le projet échoue en raison du coût financier supérieur à celui évalué. **En 1990**, une nouvelle initiative, le **Projet Génome Humain ou *Human Genome Project*** (**HGP**) voit le jour. Ce projet international, coordonné par le département américain pour l'énergie (DOE) et l'institut national américain de la santé (NIH), établi un plan sur 15 ans pour décrypter le génome humain (séquençage, annotation), adresser les questions éthiques, légales et sociales issues du projet et analyser les génomes des nombreux autres organismes afin de comprendre les fonctions des gènes. En juillet 1995 au TIGR (*The Institute for Genome Research),* l'équipe de Craig Venter publie la première séquence complète du génome de la bactérie *Haemophilus influenzae* et présente la méthode « *shotgun* », méthode de séquençage aléatoire et de reconstitution *in silico* du génome (Fleischmann *et al.*, 1995). Cette même année, l'apparition du premier séquenceur à capillaire constitue une avancée technique considérable, autorisant les traitements à haut débit, la reproductibilité des résultats et la diminution des coûts. Cette technologie, qui ne cesse de s'améliorer, a augmenté les performances de séquençage des laboratoires d'un facteur dix entre 1995 et la fin de 1997. Ainsi, les génomes de la levure de boulanger, *Saccharomyces cerevisiae* (The yeast genome directory, 1997), du ver nématode, *Caenorhabditis elegans* (The C.elegans Sequencing Consortium, 1998), de la drosophile, *Drosophila melanogaster* (Adams *et al.*, 2000), ou encore de l'arabette, *Arabidopsis thaliana (The Arabidopsis genome initiative, 2000),* ont rapidement suivi celui de *H. influenzae*.

En 1998, Craig Venter crée la société Celera Genomics® dont le but est de séquencer le génome humain en trois ans. En réponse à cette annonce, et afin de faire face aux menaces d'appropriation du génome humain, les organismes publics de financement anglo-saxons (NIH, DOE et *Wellcome Trust*) annoncent une augmentation importante des budgets et un nouvel objectif intermédiaire : l'assemblage, pour le printemps 2000, d'une ébauche préliminaire de la séquence du génome humain. Ainsi, avec 5 ans d'avance sur le programme, le premier brouillon (*working draft*) du **génome humain est publié en 2001,** simultanément par Celera Genomics® et le HGP (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Le travail de finition s'est achevé en avril 2003: une version complète et précise à 99,99% de la séquence du génome humain est aujourd'hui librement accessible (Schmutz *et al.*, 2004).

1.2 Conséquences

Les conséquences scientifiques et médicales du séquençage des génomes sont considérables : compréhension du fonctionnement et de l'évolution des organismes, diagnostic génétique et thérapie génique, nouveaux médicaments…Mais au-delà de ces retombées, le séquençage des génomes a engendré d'importants développements techniques et technologiques. La pratique de la biologie moléculaire s'est généralisée. La miniaturisation et l'automatisation des procédés ont conduit aux expériences en multiplexes, dites à haut débit, générant simultanément un grand nombre d'informations. Enfin, les outils et méthodes informatiques sont dès lors devenus indispensables pour automatiser les expériences comme pour sauvegarder et analyser les résultats.

1.3 Prochain enjeu

L'achèvement du séquençage du génome humain correspond plutôt au début d'une aventure qu'à l'avènement d'une connaissance. **Le prochain enjeu est l'annotation des génomes**. En effet, nous ignorons encore beaucoup de la structure des génomes, des mécanismes de régulation des gènes ou encore du fonctionnement des produits des gènes.

Pareil à Champollion, il va falloir traduire les quelques 3 milliards de base du génome humain en le comparant, par exemple, aux génomes d'organismes modèles, telles la drosophile ou la souris. L'analyse des séquences a, par exemple, permis la mise en évidence de motifs conservés entre les organismes et a ainsi offert d'émettre des hypothèses sur leur fonction. Certains de ces motifs très conservés (1% du génome humain) correspondent notamment à des séquences non-codantes (***C**onserved **N**on-**G**enic **s**equences* - *CNGs*) dont nous connaissons peu de choses (Dermitzakis *et al.*, 2005). Elles semblent être impliquées dans les mécanismes de régulation des gènes comme dans la conformation et l'interaction des chromosomes. De plus, il est probable que des altérations au niveau de ces motifs (mutation, délétion) sont à l'origine de variations phénotypiques, voire de pathologies. Dans le cas de motifs conservés représentant des séquences codantes, les techniques d'alignement de séquences ont également permis d'inférer la fonction de quelques gènes et le fonctionnement de certains de leurs produits (Henikoff *et al.*, 1997). Cependant, la capacité d'identifier les gènes au niveau des acides nucléiques, et plus particulièrement des ARNm, permet non seulement d'aider à l'annotation des gènes mais aussi de mettre en évidence leurs niveaux et modes d'expression dans des conditions données. Par conséquent, l'approche transcriptomique a rapidement été privilégiée pour étudier les mécanismes de l'expression des gènes.

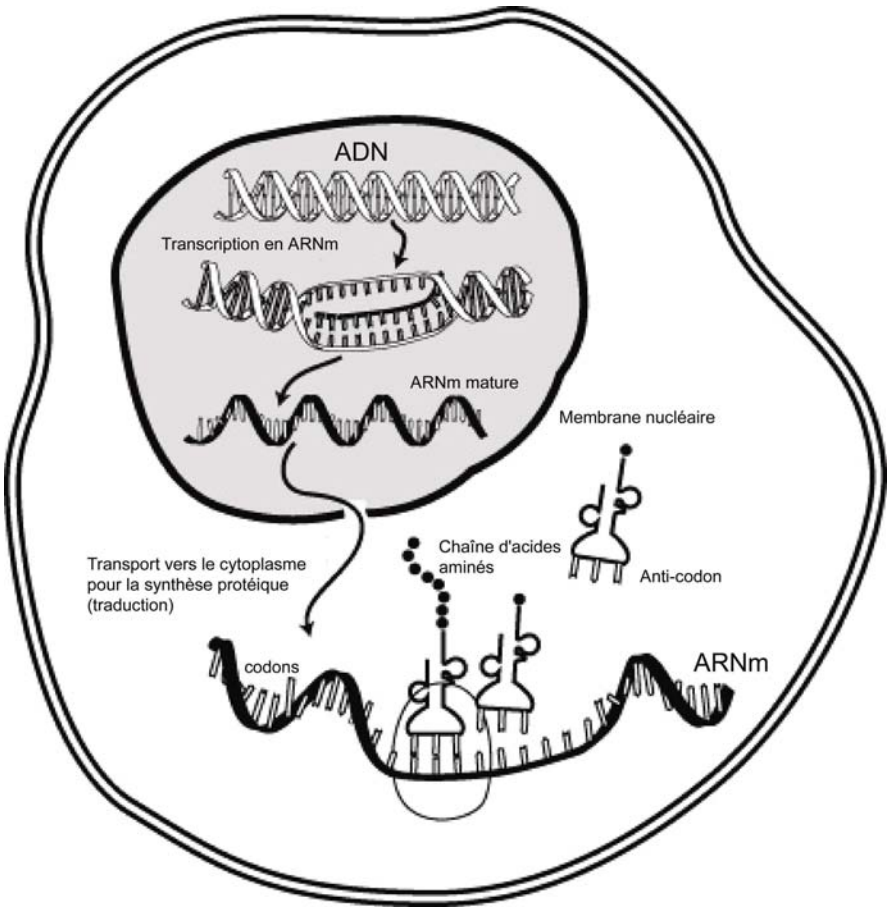


Figure 2. Du gène à la protéine : transcription des gènes en ARNm (transcriptome) et traduction des ARNm en protéines (protéome). Les ARNm sont produits dans le noyau par l'ARN polymérase II à partir d'une matrice d'ADN. Cette transcription est déclenchée par des facteurs de transcription capables d'activer spécifiquement certains gènes. Les ARNm sont ensuite exportés dans le cytoplasme pour être éventuellement traduits en protéine.

2. Etude du transcriptome

A l'issu du séquençage du génome humain, le nombre de gènes constituant notre génome a été estimé à 35 000 gènes (Hogenesch *et al.*, 2001; Pennisi, 2003). Aujourd'hui, de nouvelles estimations réduisent ce chiffre à 20 000 - 25 000 gènes (International Human Genome Sequencing Consortium, 2004), à peine plus que la drosophile qui en compte 13 601 (Adams *et al.*, 2000). Aussi, l'origine des différents niveaux de complexité des organismes est la diversité des mécanismes de régulation et de transcription des gènes (Fig. 2). Chez l'homme, sur ~ 200 000 ARNm transcrits (différents) seuls 10 000 à 20 000 sont exprimés dans une cellule spécialisée. Ces derniers constituent le **transcriptome de la cellule**, *i.e*. **l'ensemble des ARNm présents dans un type cellulaire donné à un moment donné et dans une condition biologique précise**. De plus, parmi ces transcrits, 4 000 à 6 000 semblent spécifiques de ce type cellulaire. L'étude de cette population de transcrits offre donc la possibilité de mieux comprendre le fonctionnement de ces cellules.

Depuis une vingtaine d'années, plusieurs techniques de biologie moléculaire ont été développées afin d'étudier le transcriptome (Liang et Pardee, 1992; Adams *et al.*, 2000). Les premières approches proposées, le *Southern blot* et le *Northern blot,* permettent d'identifier et localiser une séquence particulière (sonde d'ARNm ou ADNc) dans un génome entier (cible) ou tout autre mélange complexe d'ADN. Ces techniques se limitent à l'analyse d'un petit nombre de gènes à la fois et ne permettent pas d'appréhender la complexité du phénomène de la transcription. Plus récemment, la technique SAGE *(Serial Analysis of Genes Expression)*, permet d'identifier et quantifier, simultanément, le niveau d'expression de plusieurs milliers de gènes, dans un type cellulaire donné (Velculescu *et al.*, 1995). Cette méthode consiste à réaliser un inventaire des transcrits par séquençage en série de courts fragments d'ADNc (9 à 14 pb) ou *sequence tags*. Cette méthode est très sensible mais aussi très longue à mettre en œuvre, coûteuse et se limite à l'évaluation des niveaux d'expression des gènes. Parallèlement à la méthode SAGE, s'est développée la technologie des puces à ADN (Schena *et al.*, 1995; Lockhart *et al.*, 1996), moins coûteuse et surtout plus évolutive en terme d'applications. En effet, les puces à ADN permettent non seulement de visualiser, simultanément, le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier ; mais aussi d'étudier la séquence des gènes dans un échantillon, les mutations ou le polymorphisme (Mantripragada *et al.*, 2004). Elles sont donc rapidement devenues les outils privilégiés pour l'analyse du transcriptome.

En bref

Les premières publications sur le séquençage du génome humain sont apparues en 2001. Aujourd'hui, le nombre de gènes du génome humain est estimé à ~25 000.

Le projet de séquençage des génomes a été accompagné par d'importants progrès techniques et technologiques dans les domaines de (i) la manipulation de l'ADN, (ii) de la miniaturisation et de l'automatisation des expériences, qui a permis le développement des technologies dites de « haut débit », ainsi que (iii) dans les méthodes informatiques pour la gestion et l'analyse des données.

L'enjeu suivant est l'annotation des gènes, *i.e.* définir leurs fonctions et leurs produits. Le transcriptome est défini comme l'ensemble des ARNm présents dans un type cellulaire donné à un moment donné et dans une condition biologique précise. L'analyse du transcriptome offre la possibilité d'annoter les gènes par la mise en évidence de leurs modes et niveaux d'expression.

II. Puces à ADN pour l'étude du transcriptome

Les puces à ADN permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Elles appartiennent à un ensemble de nouvelles techniques développées depuis quelques années à l'interface de nombreuses spécialités comme la biologie moléculaire, la chimie, l'informatique, l'électronique et la robotique. Le concept de puce à ADN date du début des années 1990. Toutefois, le principe fondateur remonte à 1975. En effet, la technologie des puces à ADN se base sur la technique d'hybridation entre des séquences complémentaires d'ADN, conformément aux observations de E. Southern en 1975 (Fig. 3). De ces observations sont nées les techniques de *Southern* et *Northern blot* qui sont à l'origine des premières puces à ADN (Lander, 1999).

Les puces à ADN ont d'abord été conçues sur de grandes membranes poreuses en nylon ou *macroarrays* (Gress *et al.*, 1992; Nguyen *et al.*, 1995). La miniaturisation, rendue possible par les progrès de la robotique, a ensuite permis le développement des *microarrays*. Comme leur nom l'indique, ces puces à ADN sont de plus petites surfaces telles une lame de microscope (Schena *et al.*, 1995) ou une petite membrane nylon (Jordan, 1998). Elles présentent également l'avantage de pouvoir être de très haute densité et par conséquent sont susceptibles de recouvrir l'intégralité du génome d'un organisme.

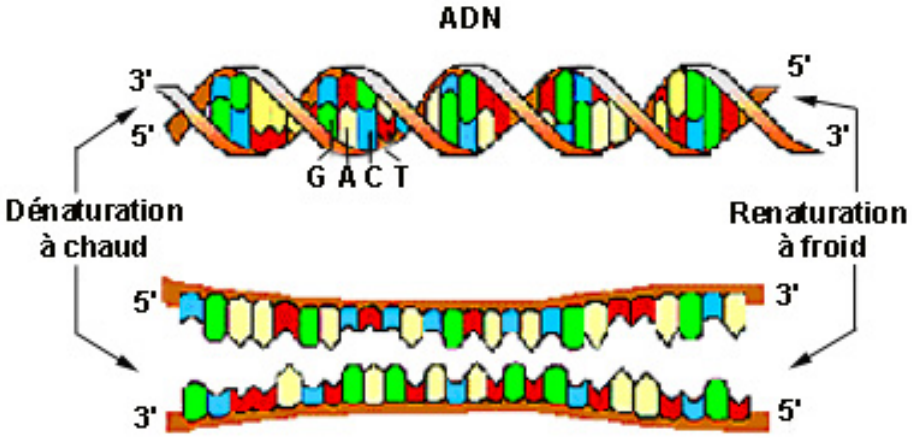


Figure 3. Complémentarité des acides nucléiques (bases).

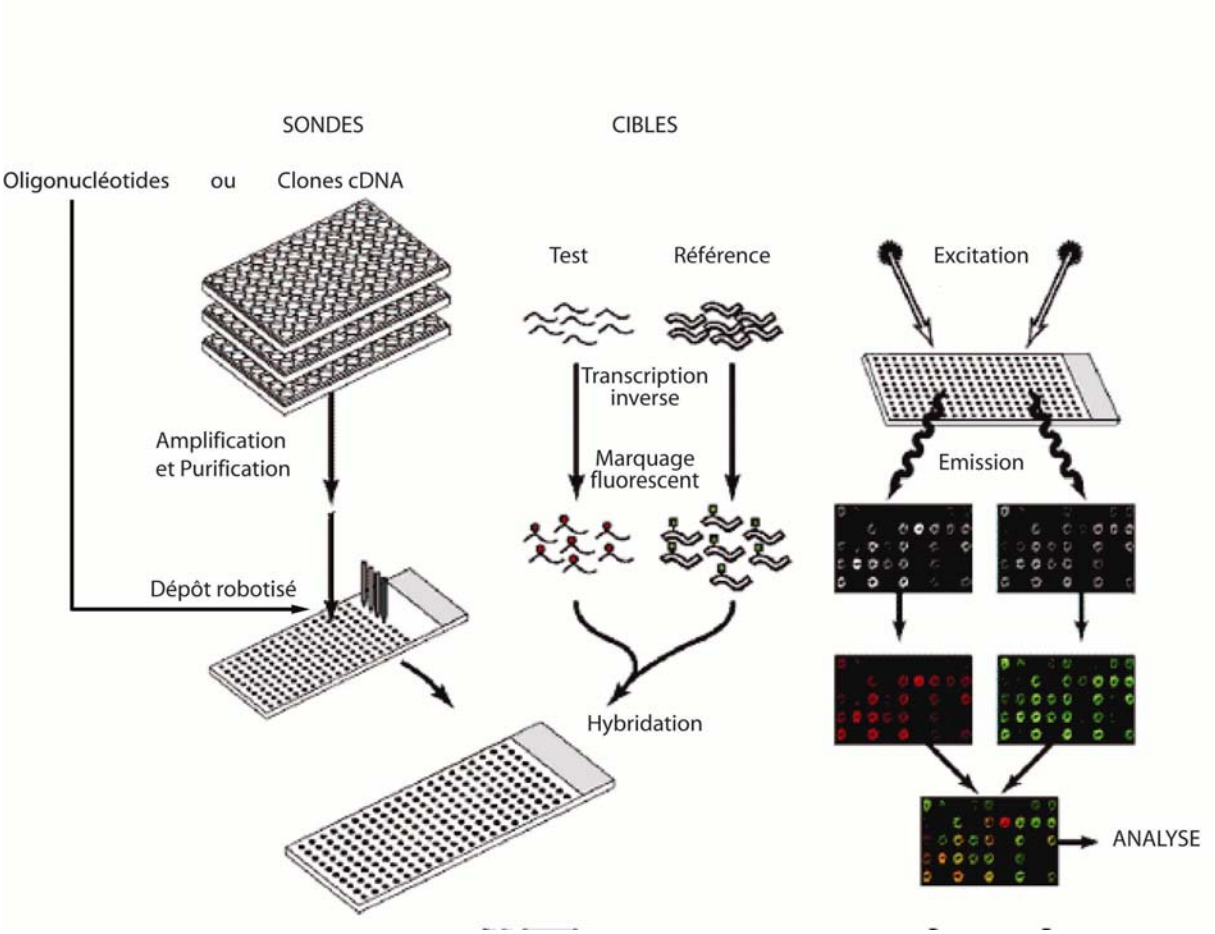


Figure 4. Schématisation de la technique d'analyse du transcriptome par la technologie des puces à DNA (d'après Duggan *et al.*, 1999).

Les sondes (oligonucléotides ou clones d'ADNc purifiés et amplifiés) sont déposées mécaniquement sur une lame de verre. Parallèlement, les cibles sont couplées à des marqueurs fluorescents (parfois amplifiés) par transcription inverse. Par exemple, la cible test est marquée par une Cyanine 5 (Cy5) rouge et la cible de référence par une Cyanine 3 (Cy3) verte. Les cibles sont assemblées pour former un mélange complexe. Ce mélange pourra s'hybrider, dans des conditions de stringence particulières, avec les sondes présentes sur la puce. La lecture est réalisée par un scanner muni d'un microscope confocal, couplé à deux lasers. Ces lasers possèdent des longueurs d 'ondes d'excitation spécifiques, correspondant à celles des deux marqueurs fluorescents. L'excitation et l'émission (amplifiée par des photomultiplicateurs) des fluorochromes permet l'obtention de deux images (une pour chaque marqueur) en niveau de gris. Ces images sont ensuite converties en pseudo-couleur et fusionnées pour être analysées par un logiciel d'analyse d'images.

1. Principe

Sur une puce à ADN, des dizaines de milliers d'hybridations peuvent être réalisées simultanément. Les hybridations se font entre des sondes nucléotidiques (*probe* ou *reporters*) ordonnées sur un support solide et des cibles (*target*) marquées, présentes dans un mélange complexe (Duggan *et al.*, 1999) (Fig. 4). Les sondes et les cibles représentent respectivement les gènes du transcriptome à analyser. Le signal d'intensité, recueilli pour chaque hybridation spécifique « sonde-cible », permet d'apprécier le niveau d'expression de chaque gène étudié dans le tissu analysé. Un profil d'expression est obtenu pour chaque échantillon.

2. Technologies

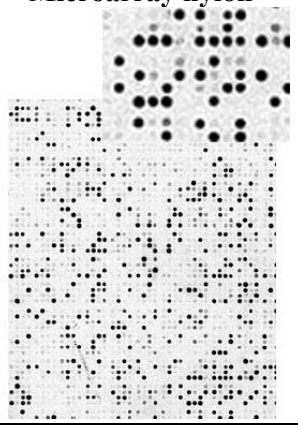
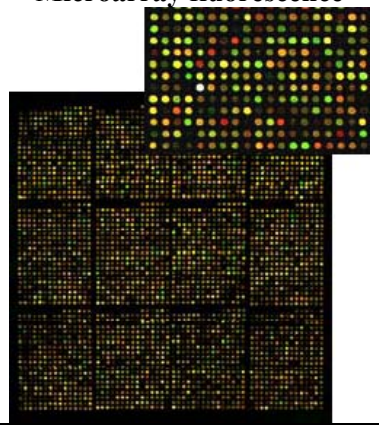
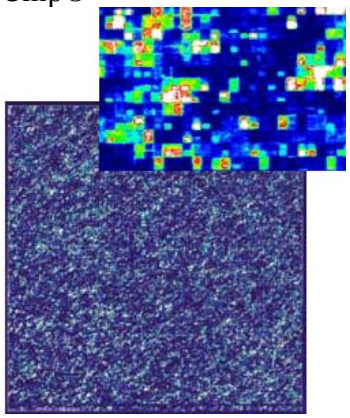
La particularité des puces à ADN, par rapport aux *macroarrays*, réside dans la miniaturisation du procédé permettant l'utilisation d'une moindre quantité de matériel génétique pour une densité plus importante de sondes. Plusieurs types de puces à ADN existent selon le support, la nature des fragments fixés à la surface, le mode de fabrication, la densité, le mode de marquage des cibles et les méthodes d'hybridation (Tab. 1).

Les supports sur lesquels sont fixées les sondes sont des supports solides, de surface plane généralement inférieure à 1cm². Les matériaux qui les composent peuvent être du verre, des polymères, du silicium, de l'or ou encore du platine. Quel que soit le support choisi, il est traité pour former un réseau dense et régulier de micro-surfaces où seront greffées les sondes.

Les sondes sont qualifiées de « *gene reporter* » car elles représentent des fragments de gènes et rapportent leur niveau d'expression. Ces *gene reporters,* ordonnés sur les lames, peuvent être des produits de PCR (puce à ADNc) (Schena *et al.*, 1995) ou des oligonucléotides plus ou moins longs (25 à 70 mers). Les produits de PCR et les oligonucléotides issus de synthèses chimiques (50-70 mers) sont greffés sur les puces à ADN par adressage mécanique ou électrochimique (Leung et Cavalieri, 2003). Les oligonucléotides peuvent également être synthétisés *in situ.* Breveté par la société Affymetrix®[1], la synthèse *in situ* par photolithographie, ou adressage photochimique, rappelle une technique couramment utilisée pour la fabrication des puces électroniques (Lockhart *et al.*, 1996).

[1] http://www.affymetrix.com/index.affx

Tableau 1. Exemples de technologies de puces à ADN

	Microarray nylon	**Microarray fluorescence**	GeneChip®
			
Support	Membrane de nylon	Verre ou silice avec revêtement chimique	Verre avec revêtement chimique
Densité	quelques centaines des spots/cm^2	1000-10000 spots/cm^2	~ 250 000 spots/cm^2
Sonde	ADNc	ADNc ou oligonucléotides	oligonucléotides (synthèse *in situ*)
Longueur de la sonde (nucléotides)	100 à 1000	ADNc ~100-1000 oligonucléotides ~ 30-70	25-60
Cibles	ADNc	ADNc	ARNc
Marquage de l'échantillon	Radioactivité (^{33}P)	Fluorescence (double : cyanine3 et 5)	Fluorescence (simple : biotine-streptavidine)
Quantité d'échantillon nécessaire (µg)	1-100	10-100	0.05-5
Sensibilité	+++	++	++
Spécificité	++	+++	+++
Principales applications	Analyse du niveau d'expression des gènes	Analyse du niveau d'expression des gènes, ChIP-Chip, CGH-array …	Analyse du niveau d'expression des gènes, étude des polymorphismes…

Les cibles sont les échantillons à étudier. Elles peuvent avoir différentes origines (tissu, une culture cellulaire…) et de différentes natures (ARNm, ADNc…). Selon la technologie de puce utilisée, les cibles sont identifiées par un marquage radioactif ou fluorescent. Bien que moins sensibles que les marquages radioactifs (Tab. 1.), certains systèmes de marquages fluorescents présentent l'avantage de pouvoir identifier plusieurs cibles sur la même puce. Par exemple, un tissu « anormal » peut être marqué par une cyanine verte (Cy3) et un tissu « sain » peut être identifié par une cyanine rouge (Cy5) (Fig. 4). Le rapport (*ratio)* des intensités obtenues pour chaque fluorochrome offre une comparaison directe des variations d'expression entre les deux échantillons.

La lecture des résultats d'hybridation se fait grâce à un scanner. Dans le cas des technologies à fluorescence, son principe est celui d'un microscope confocal couplé à un ou plusieurs lasers. Chaque laser excite spécifiquement un fluorochrome. L'émission est amplifiée par un photomultiplicateur et transformée en signal digital, *i.e.* en image. Chaque pixel de l'image scannée représente une mesure de fluorescence. Pour les puces à ADN deux couleurs, deux images en niveau de gris sont générées (une pour chaque fluorochrome). Ces images sont converties en fausses couleurs (allant généralement du vert au rouge) et superposées. Un logiciel d'analyse d'images extrait des informations qualitatives (diamètre, niveau de saturation) et semi quantitatives (intensité du signal et du bruit de fond) pour chaque complexe sonde-cible (*spot*) dans chacun des fluorochromes. Des méthodes et outils informatiques sont ensuite nécessaires pour analyser et extraire la connaissance des données.

Le choix de l'unité INSERM U533, au sein de l'IFR26 et de la plate-forme puce à ADN Ouest Genopole® de Nantes, s'est tout d'abord porté sur l'emploi des puces à ADNc puis à oligonucléotides longs (50 mers). Ces oligonucléotides sont issus d'une synthèse chimique et sont adressés mécaniquement sur les lames de verre. Le marquage des cibles se fait au moyen de deux fluorochromes, les cyanines 3 et 5. Par conséquent, dans la suite du présent manuscrit, seules les questions concernant cette technologie seront abordées.

3. "Array of hope"

Comme souligné par Lander (1999), dans son article « *Array of hope* », les puces à ADN offrent de nombreuses perspectives. Leur principale application est l'**étude du niveau d'expression des gènes et les mécanismes génétiques** qui leur sont associés au niveau cellulaire. De nombreuses études ont notamment été réalisées pour étudier **la cinétique des phénomènes cellulaires** comme la différenciation ou le cycle cellulaire (DeRisi *et al.*, 1997; Chu *et al.*, 1998; Spellman *et al.*, 1998; Schaffer *et al.*, 2001). Par exemple, Spellman *et al.* (1998) identifient chez *Saccharomyces cerevisae* 800 gènes exprimés au cours du cycle cellulaire. Des travaux ont également offert une meilleure description de certaines **voies métaboliques** chez la levure (DeRisi *et al.*, 1997; Eisen *et al.*, 1998). Enfin, ces études ont permis d'**inférer les fonctions de gènes inconnus** grâce à l'observation de leur co-régulation avec des gènes annotés.

Une autre application de l'analyse des profils d'expression est l'**amélioration du diagnostic et pronostic clinique**. Les études du transcriptome humain réalisées sur les **cancers** (Golub *et al.*, 1999; Sorlie *et al.*, 2001; van de Vijver *et al.*, 2002) et les **hémopathies** (Alizadeh *et al.*, 2000) ont apporté des résultats essentiels à la compréhension de ces pathologies. Des profils d'expressions, caractéristiques de certaines tumeurs, ont permis d'**améliorer les classifications cliniques,** parfois insuffisantes (Sorlie *et al.*, 2001; Vey *et al.*, 2004; Bertucci *et al.*, 2004). Grâce aux puces à ADN, il est désormais possible de distinguer différents types d'hémopathies aux pronostics de survie différents (Alizadeh *et al.*, 2000). De la même manière, les portraits moléculaires de patientes atteintes du cancer du sein ont mis en évidence 70 marqueurs pronostiques (gènes) jugés plus pertinents que les paramètres clinico-biologiques existants (van de Vijver *et al.*, 2002). Plus récemment, l'utilisation des puces à ADN pour étudier les **maladies cardio-vasculaires** apporte des résultats encourageants (Steenman *et al.*, 2003; Napoli *et al.*, 2003; Liew et Dzau, 2004). Les profils moléculaires de patients en insuffisance cardiaque semblent pouvoir affiner les classifications cliniques (Liew et Dzau, 2004; Steenman *et al.*, 2005). Ces études ont également permis d'identifier de nouveaux facteurs génétiques et environnementaux impliqués dans les problèmes d'insuffisance cardiaque comme l'hypertrophie, l'infarctus ou l'ischémie. Parmi les facteurs environnementaux, ces travaux citent entre autres les thérapies et l'alcool. Les facteurs génétiques ont des origines héréditaires ou sont les conséquences d'autres dysfonctionnements (maladie des artères coronaires, dysfonctionnement valvulaire, hypertension …). Finalement, ces nouveaux marqueurs offrent de nouvelles pistes pour le

diagnostic et le pronostic de ces pathologies ainsi que pour la recherche de nouvelles cibles thérapeutiques (Liew et Dzau, 2004).

Les puces à ADN sont également de puissants outils **pour la découverte et la validation des médicaments** *via* la compréhension de leurs mécanismes d'action au niveau cellulaire (Hughes, 2002; Howbrook *et al.*, 2003). Par exemple, Nguyen *et al.* (2004) ont mis en évidence un effet inhibiteur de la curcumine sur le mécanisme de re-sténose suite à la pose de stent coronarien. Des différences d'efficacité d'un même médicament, entre différents patients, ont également été montrés grâce aux puces à ADN (Kim *et al.*, 2004).En effet, certains patients atteints d'un cancer gastrique possèdent (ou acquièrent) une résistance à leur traitement. De telles analyses, faites de manière prospective, pourront peut-être aider au choix de la meilleure thérapie pour chaque patient.

Finalement, outre l'analyse du niveau d'expression des gènes, les puces à ADN sont utilisées avec succès dans divers domaines (Mantripragada *et al.*, 2004) tels que le **séquençage et la détection de polymorphismes**, ou SNP (*Single Nucleotide Polymorphism*). Cette technologie est également à l'origine de nouveaux types de bio-puces (Tab. 2) parmi lesquelles les **Array-CGH** (*microarray-based Comparative Genomic Hybridization*) pour la recherche d'altérations chromosomiques, les ***Tissu array*** pour localiser l'expression des gènes et produits de gènes au niveau tissulaire ou encore les **ChIP-Chip** (*ChromatoImmunoPrecipitation-chip*) pour l'étude des facteurs de transcription.

Aujourd'hui, la technologie des puces à ADN a atteint une certaine « maturité » (Kenyon *et al.*, 2002). De nombreuses améliorations techniques et technologiques ont été apportées pour valider la qualité des cibles à hybrider (Auer *et al.*, 2003), améliorer leur marquage (Manduchi *et al.*, 2002) ainsi qu'optimiser l'acquisition et le traitement du signal (Le Meur, 2001; Yang *et al.*, 2001a). Dans l'optique d'applications cliniques, des progrès ont également été faits dans les domaines de la miniaturisation et de l'amplification des cibles (Eberwine, 1996). Ces améliorations permettent de travailler avec des quantités toujours plus faibles telles que des biopsies (~ quelques milligrammes)(Wang *et al.*, 2003) ou différents types de petites cellules (~10^5 cellules) (Whitney *et al.*, 2003; Ma et Liew, 2003; Xiang *et al.*, 2003). Enfin, de nombreux outils bio-informatiques ont été développés pour améliorer la gestion, le traitement, l'analyse et l'intégration de cette pléthore de données.

Tableau 2. Quelques applications des bio-puces.

Bio-puces	Interactions	Applications (exemples)	Références
Puce à ADN	ADN-ADN,ADN-ARN	Analyse du niveau d'expression des gènes Recherche de polymorphisme (SNP) Séquençage	(Schena *et al.*, 1995) (Wang *et al.*, 1998) (Hurowitz et Brown, 2003)
Tissu array	ADN, ARN, protéine	Détection en parallèle de réarrangements génétiques (ADN), de l'expression des gènes (ARN) et des produits de gènes (protéines)	(Kononen *et al.*, 1998)
Array-CGH Comparative Genomic Hybridization)	ADN-ADN	Hybridation génomique comparative à haute résolution pour la rechercher des altérations chromosomiques	(Kallioniemi *et al.*, 1992) (Pinkel *et al.*, 1998)
Protein array	protéine–ADN, protéine–petite molécule protéine–ARN protéine–protéine protéine–récepteur	Analyse de la spécificité des anticorps Expression des protéines Diagnostic à partir de sérum Immuno-essay	(Haab *et al.*, 2001)
ChIP-chip (Chromatin-ImmunoPrecipitation on Chip),	ADN–protéine	chromatine-immunoprécipitation pour l'identification des sites d'interaction protéines-ADN génomique comme les facteurs de transcription	(Ren *et al.*, 2000) (Lieb *et al.*, 2001) (Iyer *et al.*, 2001)
Carbohydrates array	Sucre-anticorps Sucre protéine	Mécanisme d'ancrage, signal	(Houseman et Mrksich, 2002) (Wang *et al.*, 2002)

En bref

Les puces à ADN sont des outils puissants pour l'analyse du transcriptome. Elles permettent, entre autres, de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Elles offrent des perspectives d'applications dans les domaines du diagnostique et pronostic médical.

Cette technologie est pluridisciplinaire. Elle intègre la biologie moléculaire, la chimie, l'informatique, l'électronique et la robotique. La production de données en masse, avec une fiabilité de plus en plus grande, ne cesse de s'accélérer. Le recours aux moyens informatiques pour gérer, exploiter, analyser cette pléthore de données est devenu indispensable.

III. Bio-informatique

1. Définition

La bio-informatique moderne est née de la convergence de deux aspects de la recherche en biologie : le stockage des séquences moléculaires sur ordinateurs sous la forme de bases données et l'application d'algorithmes mathématiques pour l'alignement des séquences d'acides nucléiques et protéiques. Discipline hybride en constante évolution, la bio-informatique et ses domaines d'applications se précisent.

La plupart des définitions de la bio-informatique suggèrent l'interaction entre la biologie, les technologies de l'information et les sciences informatiques (les mathématiques). D'après Claverie *et al.* (1999), « la bio-informatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de structures de protéines …C'est le décryptage de la « bio-information » (« *Computational Biology* » en anglais) ». Andrade et Sander, dans *Bioinformatics : from genome data to biological knowledge*, Current Opinion in Biotechnology (1997), présentent une définition plus large de la bio-informatique. Selon ces auteurs, *« Bioinformatics is a science of recent creation that uses biological data, completed by computational methods, to derive new biological knowledge* ». Cette définition, plus moderne, sous-entend que la bio-informatique ne se limite évidemment pas à l'analyse des séquences. Un objectif fondamental est la volonté d'intégration de données de différentes natures, celles relatives aux séquences mais aussi celles concernant les marqueurs moléculaires, les données phénotypiques, *etc.* La bio-informatique est une approche *in silico* de la biologie traditionnelle qui vient compléter les approches classiques *in situ* (dans le milieu naturel), *in vivo* (dans l'organisme vivant) et *in vitro* (en éprouvette).

La bio-informatique est une branche théorique et pratique de la biologie. Sur le plan théorique, sa finalité est la synthèse des données biologiques à l'aide de modèles et de théories en énonçant des hypothèses généralisatrices et en formulant des prédictions. Sur le plan pratique, son but est de proposer des méthodes et des logiciels pour la sauvegarde, la gestion et le traitement de données biologiques. Par souci de clarté, les Anglo-saxons, utilisent deux termes pour distinguer ces deux aspects de la bio-informatique. Associé au terme de "bioinformatics" pour l'aspect pratique, ils utilisent le terme générique de « *biocomputing* » ("*computational biology*" pour les Américains) pour désigner l'aspect théorique.

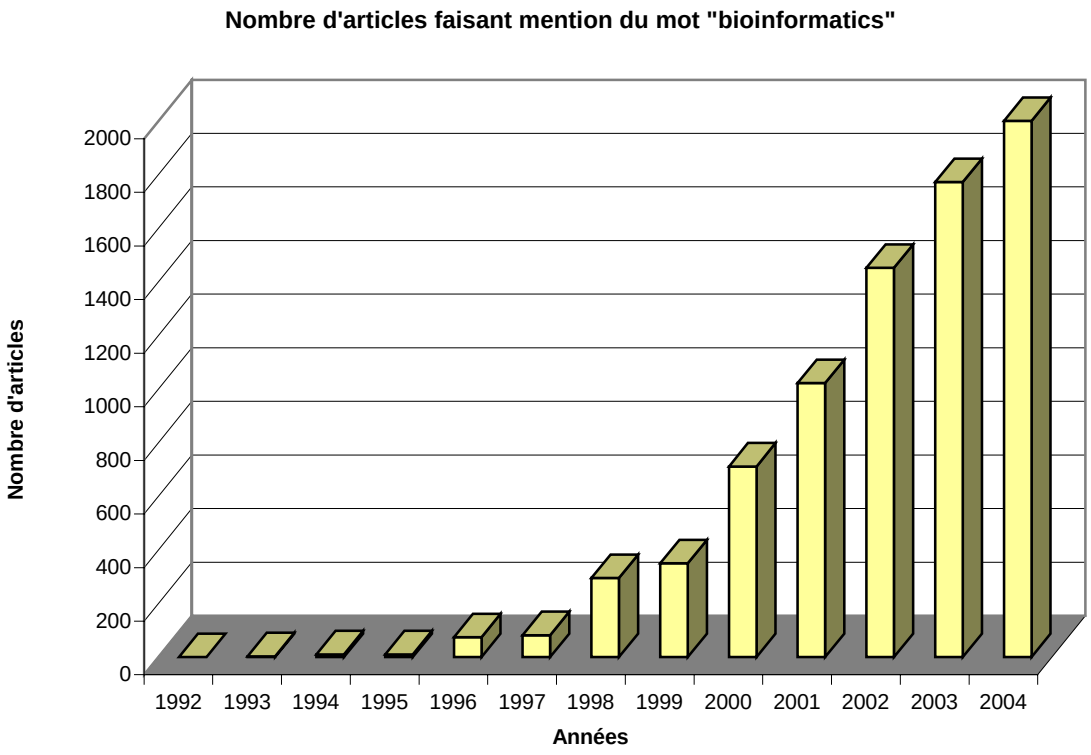


Figure 5. La bio-informatique dans la littérature scientifique de 1992 à nos jours (source : PubMed). De 1992 à 2004, croissance exponentielle du nombre d'articles référencés dans PubMed sous le terme « *bioinformatics* ».

2. Historique

2.1 Emergence de la bio-informatique

Dès 1965, quelques dizaines de laboratoires dans le monde travaillent avec les bio-mathématiques, disciplines constituées pour répondre aux besoins de la phylogénie moléculaire, de la modélisation et de la génétique des populations. La même année est publiée le premier atlas sur les séquences et structures protéiques, par Margaret Dayhoff. Dès lors, les bases de données biologiques se développent. En 1979 et 1980, les premières bases de données de séquences d'ADN apparaissent avec la *Los Alamos Sequence library* du DOE, qui devient en 1982 GenBank, et EMBL du Laboratoire Européen de Biologie Moléculaire (Brown, 2003). Enfin, la base théorique de la plupart des algorithmes qui constituent aujourd'hui le cœur de nombreux outils bio-informatiques (modèles de Markov, échantillonneur de Gibbs…) date également de cette époque.

Cependant, le terme « *Bioinformatics* » n'est apparu dans la littérature scientifique qu'au tout début des années 1990 (Brown, 2003). Longtemps cantonné dans les articles aux matériels et méthodes, l'emploi du terme « bio-informatique » n'apparaît que très tardivement dans les bases de données bibliographiques. Avant 1985, le terme « bio-informatique » n'est pas indexé comme mot clé par la base de données de références bibliographiques médicales Medline. Jusqu'en 1992, il n'apparaît pas non plus dans les titres ou les résumés référencés. En 1993, le terme apparaît enfin 3 fois puis 9 et 10 fois en 1994-95 pour ensuite augmenter de façon exponentielle (Fig. 5). Les premiers articles dans le domaine ont le plus souvent été publiés dans les journaux *The journal of Molecular Biology, Nucleic Acids Research* et *Computer Applications in Biological Sciences*. Ce dernier, fondé en 1985, devient en 1998 *Bioinformatics*, aujourd'hui journal de référence de la discipline. Désormais, plus d'une dizaine de journaux consacrés à la bio-informatique existent. Avec le développement de l'Internet, certains de ces journaux tel que *BMC bioinformatics* ne paraissent même plus sous format papier.

2.2 La communauté bio-informatique

La communauté bio-informatique s'est largement développée au cours de ces 10 dernières années. Des nombreux groupes de travail nationaux, européens et internationaux ont vu le jour.

En 1997, né le consortium ISCB ou *The International Society for Computational Biology.* Issu des conférences du ISMB (*Intelligent Systems for Molecular Biology*) initiées en

1993, le ISCB se consacre à l'avancement de la compréhension des systèmes vivants par le calcul. Cette organisation compte plus de 1300 membres et a pour journal officiel *Bioinformatics.* De cette initiative sont nées de nombreuses conférences pour le développement de la bio-informatique, parmi lesquelles nous pouvons citer JOBIM (Journées Ouvertes à la Biologie, Informatique et Mathématiques), ECCB (*European Conference on Computational Biology*), *Computer Society Bioinformatics Conference* ou encore *Pacific Symposium on Biocomputing*. Les thèmes abordés lors de ces conférences sont très variés allant de l'analyse de séquences aux traitements des données issues des technologies « haut débit » (SAGE, CGH, puces à ADN) en passant par la représentation des connaissances et les ontologies.

Plus spécifiquement, dans le domaine des puces à ADN, nous pouvons mentionner le groupe de travail MGED[2] (*Microarray Gene Expression Data Society*) et la conférence CAMDA (*Critical Assessment of MicroarrArray Data Analysis*). MGED est une organisation internationale composée de biologistes et bio-informaticiens dont le but est la standardisation et le partage des données issues des expériences de génomique fonctionnelle et protéomique. La conférence CAMDA, quant à elle, a vu le jour au département de ressources informatiques de l'université de Duke (Johnson et Lin, 2001). Elle vise à établir l'état actuel des connaissances concernant les méthodes d'exploitation des données de puces à ADN, identifier les progrès et définir les nouvelles orientations. Dans ce but, CAMDA a adoptée une approche originale. Elle propose une expérience à l'échelle internationale, laissant les scientifiques analyser le même jeu de données avec différentes méthodes. Les techniques sont ensuite présentées et discutées.

3. Outils de la bio-informatique

3.1 Internet

L'innovation technique représentée par l'informatique, en particulier le réseau Internet et les différents services qui lui sont associés, a changé le processus de traitement des données et des documents issus de la biologie moléculaire. Les rapports à l'information et aux méthodes de travail individuelles et collectives ont été profondément bouleversés. En effet, Internet constitue un outil puissant pour la collecte, le traitement et la diffusion des informations scientifiques (données ou documents numériques). Des portails, tel Google[3], offrent désormais des moteurs de recherche dédiés à la documentation scientifique (Butler,

[2] http://www.mged.org/
[3] http://scholar.google.com/

2004). Internet a contribué à "populariser" la bio-informatique par le biais de grands centres de ressources et banques de données tels que Infobiogen[4], l'Institut européen de bio-informatique[5] (*European bioinformatic Institut* - EBI), *The National Center for Biotechnology Information*[6] (NCBI) et Ensembl[7]. Les banques et les bases de données publiques sont rapidement devenues des sources précieuses d'informations et des outils d'analyse pour la recherche en biologie et bio-informatique. Ainsi, chaque année, le journal *Nucleic Acids Research* présente deux éditions spéciales : « Database Issue » consacrée aux banques et bases de données pour biologie moléculaire et « Web Service Issue » dédiée aux outils bio-informatiques. Finalement, associé à Internet, un nouveau concept de développement informatique a également participé à l'avènement de la bio-informatique : la philosophie du « monde du libre ».

3.2 Le « monde du libre » et *open-source*

Au milieu des années 1980, Richard Stallman crée la *Free Software Foundation* ainsi que le projet GNU[8] (acronyme récursif de "GNU's Not UNIX") dans le but d'offrir une implémentation gratuite, libre et ouverte du système d'exploitation UNIX. La motivation principale de cette initiative était, selon Stallman, que les créations logicielles des chercheurs en informatique devaient pouvoir être évaluées, justifiées, reproduites et améliorées afin de dynamiser l'innovation scientifique. C'est ainsi que né le noyau Linux dont l'association avec le projet GNU a initié le développement des logiciels *open-source*.

La bio-informatique a tiré profit du développement des logiciels *open-source*. Les logiciels dits *open-source* sont des logiciels gratuits et libres d'utilisation. Leur diffusion suit les termes des licences publiques, ou *General Public Licences* (GPL), qui spécifient la libre distribution et utilisation des codes sources de ces programmes. Grâce à ce mode de diffusion, les logiciels *open-source* ont rapidement été adoptés. Ils ont également et surtout modifié la façon de penser et d'appréhender les sciences de l'information. Ainsi, de nombreuses organisations internationales, notamment dans le domaine de la bio-informatique, se sont développées dans le but de coordonner les efforts de développement des logiciels *open-source.*

[4] http://www.infobiogen.fr/
[5] http://www.ebi.ac.uk/
[6] http://www.ncbi.nlm.nih.gov/
[7] http://www.ensembl.org/
[8] http://www.gnu.org

The Bioinformatics Organization[9], *Inc.*, par exemple, a été fondée en 1998 afin de faciliter la communication et les collaborations internationales entre les bio-informaticiens confirmés et les néophytes. Le site *bioinformatics.org*[8] est aujourd'hui le portail d'une organisation internationale destinée au développement de projets de bio-informatique. Cette organisation compte actuellement plus de 11000 membres et 180 projets. Le site propose un libre accès aux matériels et méthodes pour la recherche scientifique et le développement de logiciels *open-source*.

The Open Bioinformatics Foundation[10] propose, quant à elle, les projets Bio* qui sont une série d'initiatives *open-source* (Stein, 2002). Ces projets sont réalisés par une centaine de développeurs qui créent des librairies de codes réutilisables dans différents langages tels que PERL, Java ou Python. Ces librairies sont encore connues sous les noms de BioPerl, BioJava et BioPython. BioPerl, par exemple, est une collection de modules qui facilitent le développement de scripts PERL pour l'interfaçage entre plusieurs applications de bio-informatique. Ces modules aident notamment à l'interrogation de diverses bases de données ainsi qu'à la manipulation des données de séquences nucléiques et protéiques. Enfin, ils permettent l'exécution et l'analyse automatique des résultats générés par différents programmes de biologie moléculaire comme BLAST, Clustal W, T-Coffee, Genscan ou HMMER.

L'ensemble de ces développements (Internet, logiciels *open-source*) et de ces projets permet le partage des connaissances. Ces outils informatiques facilitent et accélèrent la recherche en bio-informatique et, par conséquent, en biologie. Ils sont devenus une aide précieuse pour l'analyse des données biologiques et, tout particulièrement, pour le traitement des données et l'extraction de connaissances des expériences de puces à ADN.

[9] http://www.bioinformatics.org
[10] http://www.open-bio.org

En bref

La bio-informatique est le domaine de la science où la biologie, l'informatique, les mathématiques et la technologie de l'information convergent comme une discipline unique.

Reconnues tardivement dans la littérature scientifique, les techniques et les outils bio-informatiques d'analyse des données biologiques font aujourd'hui l'objet de nombreux articles et manifestations scientifiques. Le dynamisme de la communauté bio-informatique est principalement liée à l'Internet et à la communauté du « monde du libre » qui participent à la diffusion de outils *open-source*.

IV. Bio-informatique et puces à ADN

1. Besoins

Les techniques bio-informatiques sont essentielles à la mise en place des méthodes d'analyse du transcriptome ainsi qu'à la gestion et l'exploitation des données qui en résultent. Les paragraphes suivants font état de quelques uns des besoins dans le domaine des puces à ADN.

Le **choix des *gene reporters*** à déposer sur les puces à oligonucléotides n'est pas trivial. De leurs propriétés physico-chimiques dépendent leur sensibilité et spécificité. De nombreux outils bio-informatiques[11] ont donc été développés dans le but d'optimiser ce rapport. Les algorithmes s'attachent notamment à valider les alignements de séquences (pas de structures secondaires internes, ni d'hybridation croisée…), établir la distance par rapport à l'extrémité 3'UTR de la séquence ou encore calculer des paramètres thermodynamiques tels que l'enthalpie et l'entropie du complexe sonde-cible pour estimer la température de dénaturation (Stekel 2003).

Les **bases de données** sont devenues des outils informatiques indispensables pour **sauvegarder, structurer, sécuriser et manipuler les données**. En effet, les puces à ADN appartiennent à ces nouvelles technologies dites à « haut débit » qui génèrent une masse considérable de données qu'il faut savoir gérer. Les informations enregistrées font référence non seulement aux résultats mais aussi à l'ensemble des étapes mises en œuvre pour concevoir les puces. Il existe un grand nombre de bases de données dédiées aux expériences de puces à ADN (Dudoit *et al.*, 2003) parmi lesquelles BASE (Saal *et al.*, 2002), ArrayDB (NHGRI), Acuity® (Axon *Inc.*) ou encore Rosetta Resolver® (Rosetta).

Une autre finalité des bases de données est la **standardisation des informations** à sauvegarder pour un meilleur partage des connaissances. Ainsi, le consortium MGED[12] propose MIAME (*Minimun Information About Microarray Experiment*) qui correspond à la liste des informations minimales à enregistrer pour décrire une expérience de puces à ADN (Brazma *et al.*, 2001). MIAME est aujourd'hui la référence pour diffuser les données de puces à ADN sur les banques de données publiques (*repository)* telles que ArrrayExpress[13] à l'EBI

[11] http://genomicshome.com/
[12] http://www.mged.org/
[13] http://www.ebi.ac.uk/arrayexpress/

ou *Gene Expression Omnibus*[14] au NCBI. Par ailleurs, ce mode diffusion est devenu, pour de nombreux journaux la condition *sine qua non* à la publication des travaux issus de cette technologie.

Les **méthodes mathématiques et statistiques** sont aussi devenues incontournables **pour le traitement et l'interprétation des données** de puces à ADN. En effet, les matrices de données d'expression présentent généralement des caractéristiques atypiques : les données sont le plus souvent bruitées et les matrices sont généralement dissymétriques (plus de *gene reporters* que d'échantillons). Aussi, la nécessité de valider la qualité des données et la difficulté d'analyser des matrices dissymétriques sont à l'origine de nombreuses recherches et d'un grand nombre de développement mathématiques et statistiques. De plus, compte tenu de la quantité des informations générées, une analyse manuelle devient très rapidement fastidieuse et source d'erreurs. L'exploitation des données ne peut se faire sans l'aide de procédures automatiques, *i.e.* d'outils logiciels. Ainsi, de nombreux algorithmes et outils, à commencer par les logiciels d'analyse d'images, sont développés pour l'acquisition, le traitement et l'analyse des données de puces à ADN.

Enfin, le traitement et l'interprétation des données issues des expériences de puces à ADN (et de manière plus générale des données génomiquees) évoluent constamment. Les outils pour le développement des méthodes de traitement et d'analyse doivent donc être flexibles. Ce besoin a incité les chercheurs en bio-informatiques à s'orienter vers des logiciels possédant des environnements de développement tels Microsoft Excel®, SAS®, S-plus®, Matlab® ou R[15] (Ihaka et Gentleman, 1996). Les principaux avantages de ces logiciels sont leur souplesse et leur interopérabilité avec d'autres outils informatiques comme les banques et les bases de données accessibles sur le Web. Ils offrent ainsi de nombreuses possibilités d'analyses avec une perspective d'intégration des diverses sources de données pour une meilleure interprétation.

[14] http://www.ncbi.nlm.nih.gov/geo/
[15] http://www.r-project.org

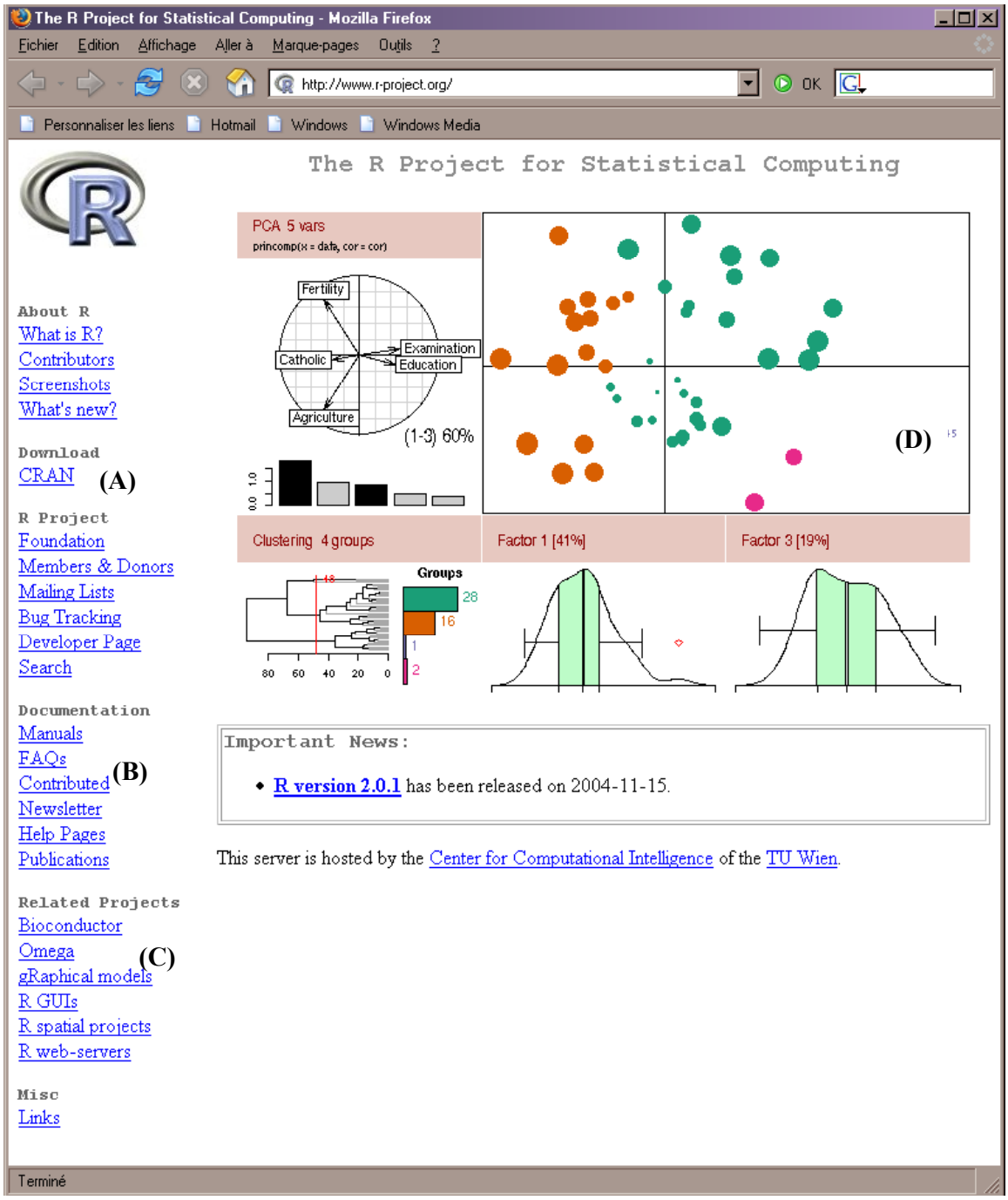


Figure 6. Site Web du projet R (www.r-project.org). La page d'accueil présente la version courante de R (ici R 2.0.1). De nombreux liens hypertextes (menu de gauche) donnent notamment accès aux (A) sites miroirs pour le téléchargement de R et de ses modules, (B) aux différentes documentations dont des manuels imprimables et des listes de diffusion, et (C) au projet *Bioconductor* pour l'analyse des données issues des expériences de génomique (puces à ADN, SAGE...). (D) Exemples des fonctionnalités graphiques de R.

2. R et BioConductor

D'après *The Bioinformatics Organization*[16], *Inc*., R est actuellement l'outil le plus utilisé pour le traitement numérique des données biologiques (soit 24% des 1136 votants contre 19% pour Matlab et 6% pour SAS®).

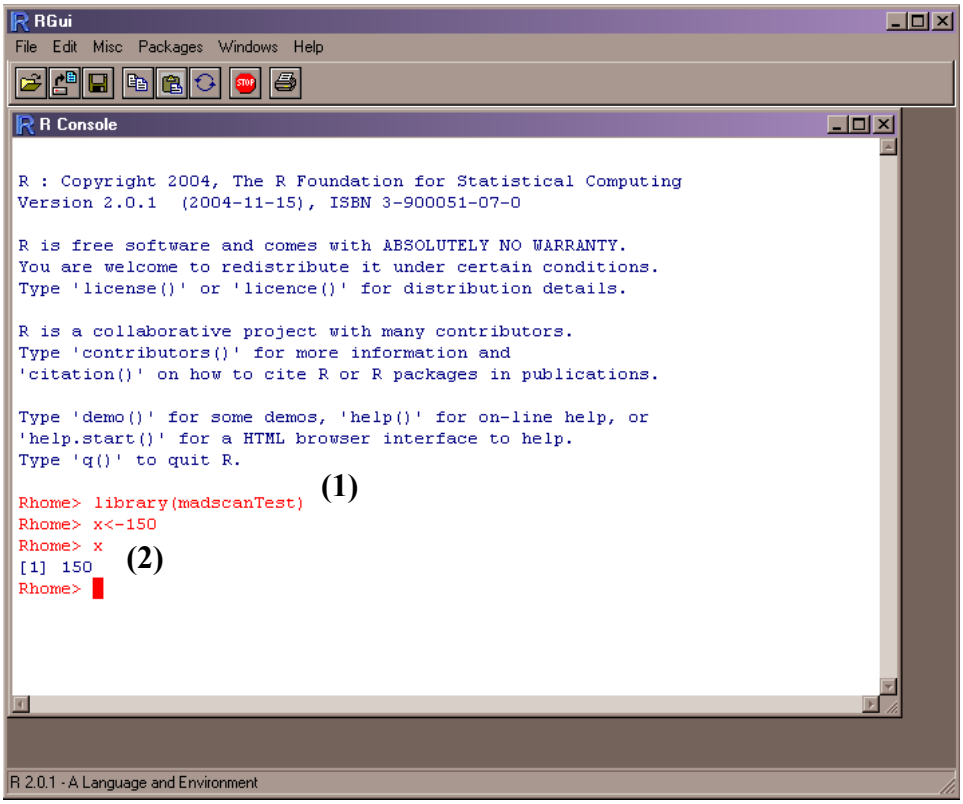
2.1 Historique

R est un outil d'analyses statistiques et graphiques qui possède son propre langage de programmation. Nommé ainsi en référence à ses deux auteurs, Ross Ihaka et Robert Gentleman (1996), son nom est aussi un clin d'œil au langage S de *AT&T Bell Laboratoires* dont il est un dialecte. Contrairement au langage S et à l'outil d'analyse statistiques S-plus, commercialisés par Insightful®, R est distribué gratuitement suivant les termes des licences publiques (*GPL*). Les codes sources et modules d'applications sont donc librement mis à la disposition de l'ensemble de la communauté scientifique.

Dans un premier temps développé pour les systèmes d'exploitation libres (et gratuits) à savoir UNIX et Linux, R est très vite devenu disponible pour les systèmes d'exploitation Windows et Mac-OS. Le noyau de R est implémenté essentiellement en langage C et FORTRAN. Ses versions sont distribuées sous la forme de codes sources binaires à compiler (UNIX et Linux) ou d'exécutables pré-compilés (Windows). Les fichiers d'installation sont disponibles à partir du site Web du CRAN[17] (*Comprehensive R Archive Network*) (Fig. 6). Ce site répertorie également une importante source de documentation pour l'installation et l'utilisation de R sur chaque système d'exploitation. Depuis 1997, un groupe de développeurs (*R Core Team)*, s'attache au maintien du bon développement des différentes versions de l'outil qui ne cesse de s'améliorer en terme de fonctionnalités graphiques et domaines d'applications (de l'exploitation des données géologiques à la génomique).

[16] http://www.bioinformatics.org
[17] http://www.r-project.org/

(A)



(B)

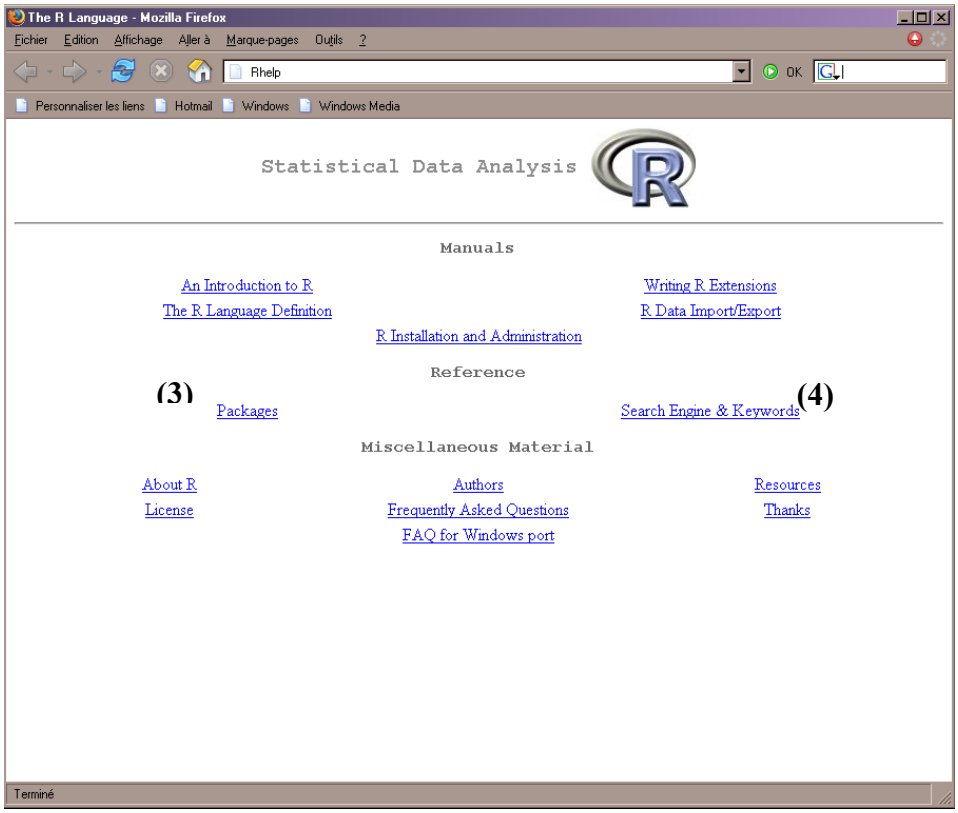


Figure 7. Logiciel R : (A) Environnement de développement (B) Documentation électronique. (1) Appel d'une librairie de fonctions (2) définition et affichage d'une variable numérique. (3) liste des librairies installées localement (4) Moteur de recherche pour l'aide (en local)

2.2 Propriétés de R

R est un langage orienté objet ce qui signifie que les variables, les données, les fonctions, les résultats (*etc.*) sont stockés dans la mémoire de l'ordinateur sous forme d'objets qui ont chacun un nom (Fig. 7A). R est également un langage interprété, i.e. non compilé. Les commandes entrées au clavier sont directement exécutées et, contrairement à la plupart des langages informatiques (C, FORTRAN, JAVA...), la construction d'un programme complet n'est pas nécessaire. Cette propriété permet d'évaluer rapidement la qualité des algorithmes et de les déboguer. Cependant, l'exécution d'un tel programme peut être plus coûteuse en temps machine qu'un programme équivalent compilé.

Outil d'analyses statistiques et graphiques, R possède un environnement graphique d'applications qui permet l'exécution de commandes non seulement en mode interactif mais aussi sous forme de programmes (*scripts*). Cette fonctionnalité permet aux développeurs de créer des librairies de fonctions. Ces modules sont dédiés à des analyses spécifiques telles la librairie *ctest* qui proposent de nombreux tests statistiques ou la librairie *blighty* qui permet de dessiner le contour des côtes britanniques. L'interface graphique offre donc également la possibilité de réaliser des représentations graphiques très sophistiquées (Fig. 6.). Toutefois, les fonctions associées sont généralement complexes et le résultat peu interactif.

Un autre atout de R est son interopérabilité. R peut dialoguer et interagir avec d'autres logiciels *open-source* écrits dans des langages différents. L'initiative Omegahat[18] a notamment contribué à promouvoir et développer cette interopérabilité. De nombreux scripts et API (***A****pplication* ***P****rogramming* ***I****nterface*) permettent ainsi une interaction bidirectionnelle de R avec les langages S, PERL, Python, Java et Visual Basic.

Enfin, R possède une importante communauté de développeurs et une documentation très riche (Fig. 7B). Les documents fournis pour l'installation et la création de librairies sont généralement très détaillés. Chaque librairie s'accompagne également d'une documentation qui décrit chaque fonction (paramètres d'entrée, format de sortie des résultats) et présente le plus souvent des exemples d'exécution.

2.3 R et la génomique : le projet BioConductor

Les projets BioPerl, BioJava, BioPython *etc.* (*cf.* p. 18) proposent différentes solutions pour le traitement et l'analyse des données biologiques. La plupart des algorithmes ont été

[18] http://www.omegahat.org

développés pour les analyses de séquences et très peu pour les analyses de données quantitatives telles que les données de bio-puces (ADN, protéines). R, langage et outil d'analyses statistiques, s'est avéré plus puissant pour le traitement des données numériques.

L'analyse des données de puces à ADN avec R a été initiée par un groupe de statisticiens dirigés par Terry Speed[19]. Leur première librairie, nommée *sma* (*statistical microarray analysis*), a été développée pour répondre aux problèmes de normalisation des données de puces à ADN deux couleurs. Cet outil et les fonctions associées ont eu un impact considérable dans le domaine de l'analyse des puces à ADN.

Compte tenu des résultats, des propriétés de R (fonctions et puissance de calcul) et du besoin croissant d'outils mathématiques pour l'analyse des données biologiques, des développeurs au sein de la communauté R ont proposé le projet BioConductor[20]. BioConductor est une initiative de collaboration entre statisticiens, mathématiciens, biologistes et développeurs afin de créer des outils informatiques (algorithmes, logiciels) pour résoudre des problèmes de biologie et de bio-informatique (Gentleman *et al.*, 2004). Les principaux buts de ce projet sont le développement, en collaboration, de logiciels innovants ainsi que leur vaste diffusion et utilisation, pour une reproductibilité des résultats de recherche.

Né en 2000, BioConductor, associé à R, reçoit en 2002 le titre de *Insightful Innovation Award Open Source & Open Development Software Project*. Insightful® commercialise aujourd'hui S^+ Analyser, un outil relativement convivial qui reprend en majorité les librairies de BioConductor. De même, GeneTrafic® d'Iobion, logiciel dédié à la gestion et au traitement des données de puces à ADN, utilise de nombreuses librairies de BioConductor. Enfin, dédiées à l'analyse des données de génomique, les librairies disponibles sur le site de BioConductor permettent non seulement l'analyse des données de puces à ADN (*e.g.* librairies *Affy, marray, limma*) mais aussi des expériences SAGE (*SAGElyzer*), de la spectrométrie de masse (*PROcess*) ou encore l'annotation des gènes (*GOstats*).

[19] http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html
[20] http://bioconductor.org

Issu du projet R, BioConductor en possède les avantages et les inconvénients. Des interfaces utilisateurs sont disponibles pour quelques librairies, telles *limmaGUI* ou *affylmGUI,* et facilitent leur emploi. Un système de « *vignettes »* documente le fonctionnement de certaines librairies et permet parfois l'exécution interactive d'exemples.Néanmoins, l'utilisation de la majorité des librairies nécessite une certaine expertise en R. De plus les modes de visualisation graphiques restent encore peu interactifs.

En Bref

R est un outil d'analyses statistiques et graphiques qui possède son propre langage de programmation. Ce logiciel, gratuit et *open-source*, offre de nombreuses possibilités d'analyses et de développement. R et sa suite BioConductor, dédiée aux méthodes d'analyse des données de génomiques, sont actuellement les outils les plus utilisés pour le développement de procédures mathématiques et statistiques en génomique.

Chapitre II.

Obtention de Données « Consolidées » à partir des Analyses d'Images de Puces à ADN

Chapitre II. Obtention de données « consolidées » à partir des analyses d'images de puces à ADN

I. Acquisition et propriétés des données de puces à ADN

Selon la finalité scientifique, différentes approches sont envisageables pour l'obtention des données d'expression. Elles reposent sur l'utilisation de différents types de puces : les puces pangénomiques ou les puces « dédiées » (Fig. 8).

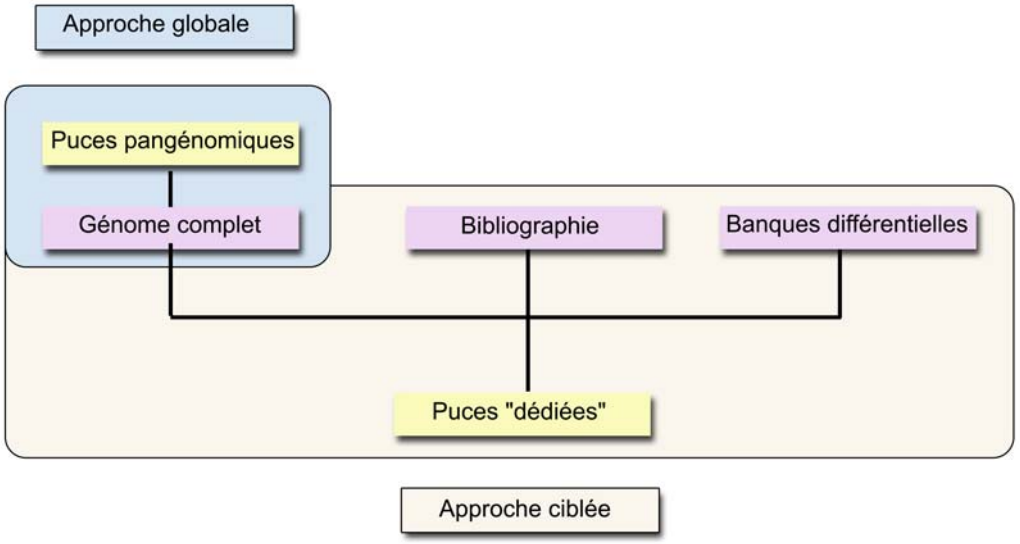


Figure 8. Puces à ADN pangénomiques ou dédiées, analyse globale *versus* analyse ciblée du transcriptome. Les puces pangénomiques proposent l'analyse d'un génome complet. Les puces « dédiées » ciblent un transcriptome. Les gènes d'intérêt des puces « dédiées » ont été identifiés par l'analyse de la littérature, le criblage de puces pangénomiques et/ou la construction de banques différentielles.

1. Puces à ADN pangénomiques ou dédiées

Les **puces pangénomiques** dites « généralistes » proposent des jeux de sondes oligonucléotidiques représentant l'ensemble des gènes d'un génome, *i.e.* séquences potentiellement transcrites. Ainsi, il existe des puces à ADN contenant plus de 30 000 *gene reporters* représentant les quelques 25 000 gènes du génome humain. L'avantage des puces

pangénomiques réside dans leur exhaustivité. Elles sont utiles pour une **analyse globale des génomes** et des classifications haut débit par hybridations systématiques. En revanche, sur ce type de puces, chaque sonde n'est déposée qu'une seule fois sur le support (limite essentiellement technologique). Aucune validation statistique des mesures ne peut donc être réalisée à l'intérieur de la puce, ce qui altère sensiblement la validité des résultats obtenus.

Les **puces dédiées** sont constituées d'une collection de gènes spécifiquement (voire exclusivement) liés à un tissu, une pathologie et/ou une thématique. Elles permettent de mieux **cibler les transcrits pertinents pour l'étude d'un type cellulaire donné**. En effet, parmi les 10 000 à 20 000 transcrits potentiellement exprimés dans une cellule spécialisée ; seuls 4 000 à 6 000 d'entre eux sont caractéristiques de ce type cellulaire. Les transcrits d'intérêt peuvent être obtenus de différentes manières, souvent complémentaires (Lamirault *et al.*, 2004):

(i) sélection expérimentale par **criblage** (*screening*) de puces à ADN pangénomiques,

(ii) connaissances biologiques ***a priori*** par analyse de la littérature et/ou interrogation des bases de données publiques,

(iii) constitution de **banques différentielles** par approches soustractives (SSH) (Diatchenko *et al.*, 1996; Tkatchenko *et al.*, 2000), séquençage systématique de banques d'EST et/ou analyse de banques SAGE (Velculescu *et al.*, 1995).

Par exemple, la comparaison des transcriptomes de patients insuffisants cardiaques ou non, au moyen de puces pangénomiques, nous a permis de montrer que seuls 2031 *gene reporters* sur 12626 (soit 16%) s'exprimaient dans au moins un des échantillons (Steenman *et al.*, 2003).

Les premières générations de puces à ADN dédiées viennent des institutions académiques, les sociétés commerciales n'étant pas convaincues d'une telle approche. Les *gene reporters* sont alors principalement des produits de PCR. Ces puces sont entre autres baptisées *CardioChip* (Barrans *et al.*, 2001), *IonChip* (Le Bouter *et al.*, 2003) ou *Myochip*[21]. La *Myochip*, développée dans notre laboratoire, est dédiée à l'exploration des pathologies cardiovasculaires et neuromusculaires telles que les valvulopathies, les arythmies, la myogenèse ou les dystrophinopathies (Fig. 9). La sélection des *marqueurs* des transcriptomes musculaires et cardiaques a été réalisée par un consortium de laboratoires d'OUEST genopole® qui travaillent sur différentes thématiques équivalentes. Aujourd'hui, la *Myochip* se compose de 4217 *gene reporters*, représentatifs de 4012 gènes.

[21] http://cardioserve.nantes.inserm.fr/ptf-puce/index.php

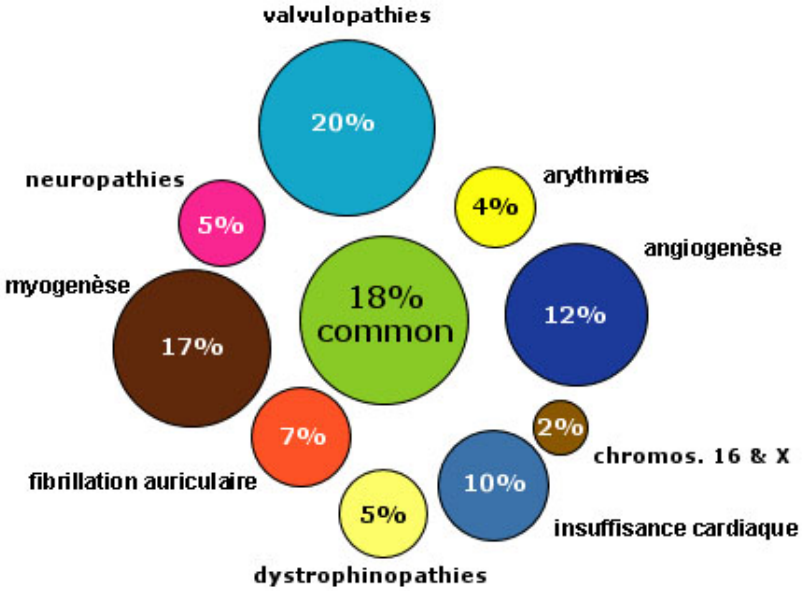


Figure 9. La *Myochip*, puce dédiée à l'exploration des pathologies cardiovasculaires et neuromusculaires. Les gènes d'intérêt ont été identifiés par l'analyse de la littérature, le criblage de puces pangénomiques et la construction de banques différentielles. Ce travail a été réalisé, dans le cadre d'OUEST *genopole* ®, par un consortium de laboratoires qui travaillent sur des thématiques équivalentes, autour du domaine du muscle strié cardiaque et squelettique. Cette puce est constituée d'environ 4,000 *gene reporters*, 18% d'entre eux étant communs à toutes les équipes du consortium.

Outre l'avantage d'une analyse plus ciblée des transcriptomes, les puces à ADN dédiées offrent la possibilité d'évaluer plusieurs fois le même *gene reporter* sur la même puce (*replicates*). En effet, les robots de dépôts ont actuellement la capacité de déposer sur une lame ~10.000 *spots* par cm² soit une densité de près de 40.000 *spots*. Ils atteignent donc leur limite avec les puces pangénomiques « humaines » et sont dans l'impossibilité de déposer plusieurs fois le même *gene reporter*. En revanche, les puces à ADN dédiées, composées de quelques milliers de *gene reporters*, permettent le dépôts d'un même *gene reporter* en plusieurs exemplaires ou plusieurs *gene reporters* pour un même gène (marqueur moléculaire). Elles offrent ainsi la possibilité d'une validation statistique de la mesure à l'intérieure de la puce et une meilleure analyse de la spécificité des *gene reporters* (et par extension des marqueurs moléculaires).

Finalement, la prise de conscience « tardive » par les entreprises commerciales de l'intérêt des puces à ADN dédiées, les a incitées à dessiner leurs propres puces dédiées (*e.g.* SuperArray bioscience Corporation[22]). Certaines sociétés proposent également de créer des puces à façon (*e.g.* Operon[23], Agilent[24]) et/ou de commercialiser les oligonucléotides à l'unité. Les puces à ADN dédiées (académiques) sont alors devenues des puces à oligonucléotides, telle la *Myochip* qui est aujourd'hui composée d'oligonucléotides 50 mers.

En bref

Les puces pangénomiques offrent une approche globale pour l'analyse du transcriptome.

Les puces à ADN dédiées permettent de cibler l'étude du transcriptome à un type cellulaire donné. Les *gene reporters* qui les composent sont sélectionnés à partir du criblage de puces pangénomiques, des connaissances biologiques *a priori* et/ou de banques différentielles. La *Myochip*, développée au sein du laboratoire, est un exemple de puce à ADN dédiée à l'exploration des pathologies cardiovasculaires et neuromusculaires.

[22] http://www.superarray.com/home.php
[23] http://www.operon.com/products_main.php
[24] http://www.home.agilent.com

Tableau 3. Exemples de sources de bruits dans les mesures de données d'expression issues des expériences de puces à ADN.

	Intensité des signaux	Bruid de fond	Forme des spots
Date	xx	x	
Expérimentateur	x	x	
Qualité du support (lame)	x	xxx	xxx
Préparation des cibles			
Marquage	xxx	xx	x
Amplification	xx		
Paramètre d'hybridation	xx	xx	
Préparation des sondes			
Aiguille du robot de dépôt	xx	xxx	xxx
Quantité déposée	x	x	xx
Acquisition des données par les logiciels d'analyse d'images			
Paramètres de lecture des lames	xx		
Superposition des images			xx
Segmentation des pixels	xx	xx	

Niveau d'impact : x faiblement, xx moyennement, xxx fortement

2. Importance du plan expérimental dans l'obtention des données de puce à ADN

Un plan expérimental décrit les expériences et conditions d'expérience à mettre en oeuvre pour répondre à une question. Cette étape est cruciale dans l'obtention de mesures de qualité. Dans le domaine des puces à ADN, de nombreux travaux ont montré la nécessité d'une meilleure définition des plans expérimentaux pour une estimation plus robuste des données d'expression (Churchill, 2002; Yang et Speed, 2002). Ainsi, lorsque deux niveaux d'expression sont comparés, l'enjeu principal est de faire la distinction entre les variabilités techniques et les variations biologiques. Ensuite, la définition du mode de comparaison des échantillons est également essentielle à la bonne interprétation de leur variation d'expression.

2.1 Variabilités techniques et/ou biologiques

La difficulté de distinguer les variabilités techniques et biologiques d'intérêt réside dans leurs origines multiples. Les **variabilités techniques** sont des **biais expérimentaux,** tels que la qualité des lots de lames, des réactifs ou des préparations (Tab. 3), qui entraînent des erreurs dans les mesures et nécessitent un traitement (nettoyage) des données. Les **variations biologiques** sont les **variabilités intra et inter-individus.** Les variabilités intra-individu font référence à l'hétérogénéité des mesures obtenues pour différents prélèvements d'un même tissu chez un individu. Les variations inter-individus correspondent aux variations biologiques entre individus et sont le plus souvent l'objectif de la recherche.

Une des solutions proposées pour distinguer les variabilités techniques et les variations biologiques est de **maximiser les *replicates* biologiques et optimiser les *replicates* techniques** (Lee *et al.*, 2000; Pavlidis *et al.*, 2003).

Réaliser des *replicates* biologiques consiste à analyser le plus grand nombre d'échantillons possibles. Si l'objectif de l'analyse est la découverte d'une nouvelle taxonomie au sein d'une maladie, les *replicates* biologiques peuvent être intra et inter-individus. Les *replicates* intra-individus correspondent à différents prélèvements du même tissu chez un même patient ; tandis que les *replicates* inter-individus sont différents prélèvements du même tissu chez différents individus. Cette approche aide à déterminer la consistance des groupes taxonomiques identifiés (Simon et Dobbin, 2003).

Les *replicates* techniques, quant à eux, peuvent correspondre au dépôt multiple d'un même *gene reporter* sur une même lame ou à l'hybridation de plusieurs lames avec les

mêmes échantillons et une inversion des marquages (*dye-swap*). Les *replicates* techniques permettent de valider la qualité des différentes étapes de conception d'une puce à ADN. Le nombre de *réplicates* nécessaires et suffisants à l'obtention de données de qualité dépend du nombre de facteurs étudiés, et plus particulièrement du type d'expérience et des modes d'analyses employés. La principale méthode pour estimer ce nombre est l'analyse de puissance (*cf.* article p. 45) (Pan *et al.*, 2002; Pavlidis *et al.*, 2003).

Un autre concept important du plan expérimental est la **randomisation** (choix au hasard) du plus grand nombre possible de paramètres capables d'influencer les résultats de l'expérience (expérimentateurs, dépôts multiples sur une lame, lot de lames...). Ces paramètres sont souvent qualifiés de facteurs de « nuisance ». Un exemple est l'utilisation de lames provenant de différents lots. En effet, si l'ensemble des échantillons contrôles est hybridé sur les lames d'un lot X et les échantillons traités sur des lames du lot Y, il sera impossible de distinguer les variations biologiques des variations liées aux lots des lames. Les deux facteurs seront confondus.

Enfin, un plan d'expérience en **« bloc »** (*block design)* permet également de gérer l'influence de certains facteurs de « nuisance » sur une expérience (Draghici 2003). Un bloc est un sous-ensemble de conditions expérimentales (*e.g.* une puce) pour lesquelles il est possible de bloquer l'effet des facteurs de « nuisance » sur la mesure. Par exemple, les spots d'une lame subissent les mêmes traitements (hybridation, lavage, séchage...) et les mesures au sein de cette lame sont plus homogènes entre elles qu'entre les autres lames de l'expérience. La comparaison entre deux échantillons hybridés sur une même lame est donc plus directe. Dans le cas d'une comparaison intra-lame, seule une normalisation des données à l'intérieure de la puce est nécessaire tandis qu'une comparaison entre lame nécessite également un ajustement des mesures entre les lames (*cf.* p. 40).

2.2 Plan expérimental : mode de comparaison des échantillons

Le choix du plan expérimental à mettre en œuvre dépend essentiellement de la question biologique posée (synthèse par Yang et Speed (2002)). Les plans expérimentaux se distinguent principalement par le mode de comparaison des échantillons qui peut être **direct et/ou indirect** (Fig. 10).

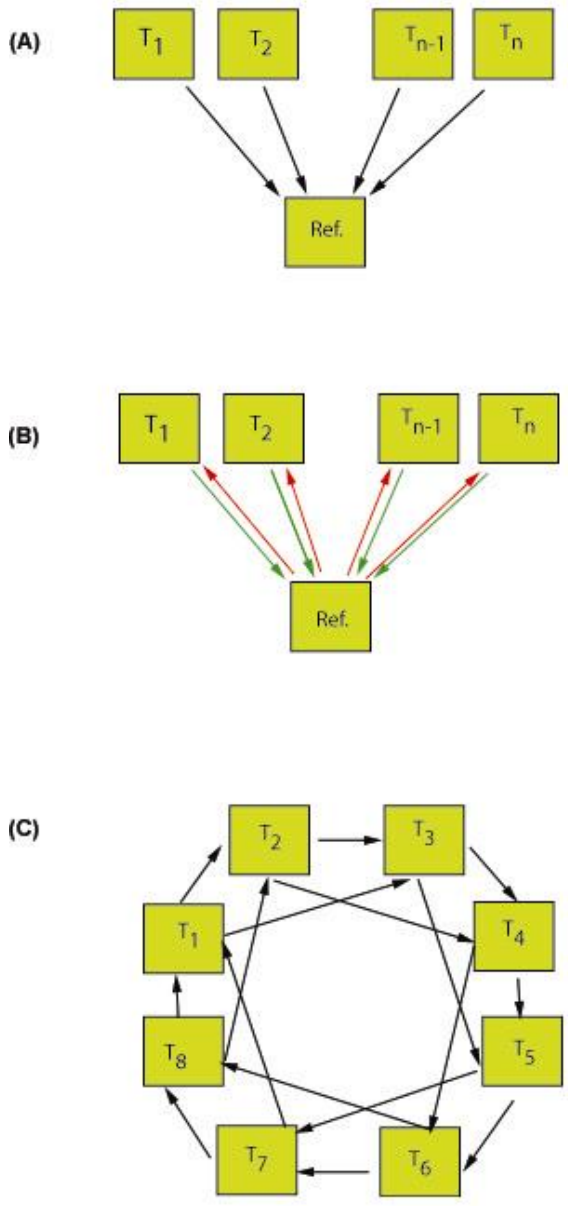
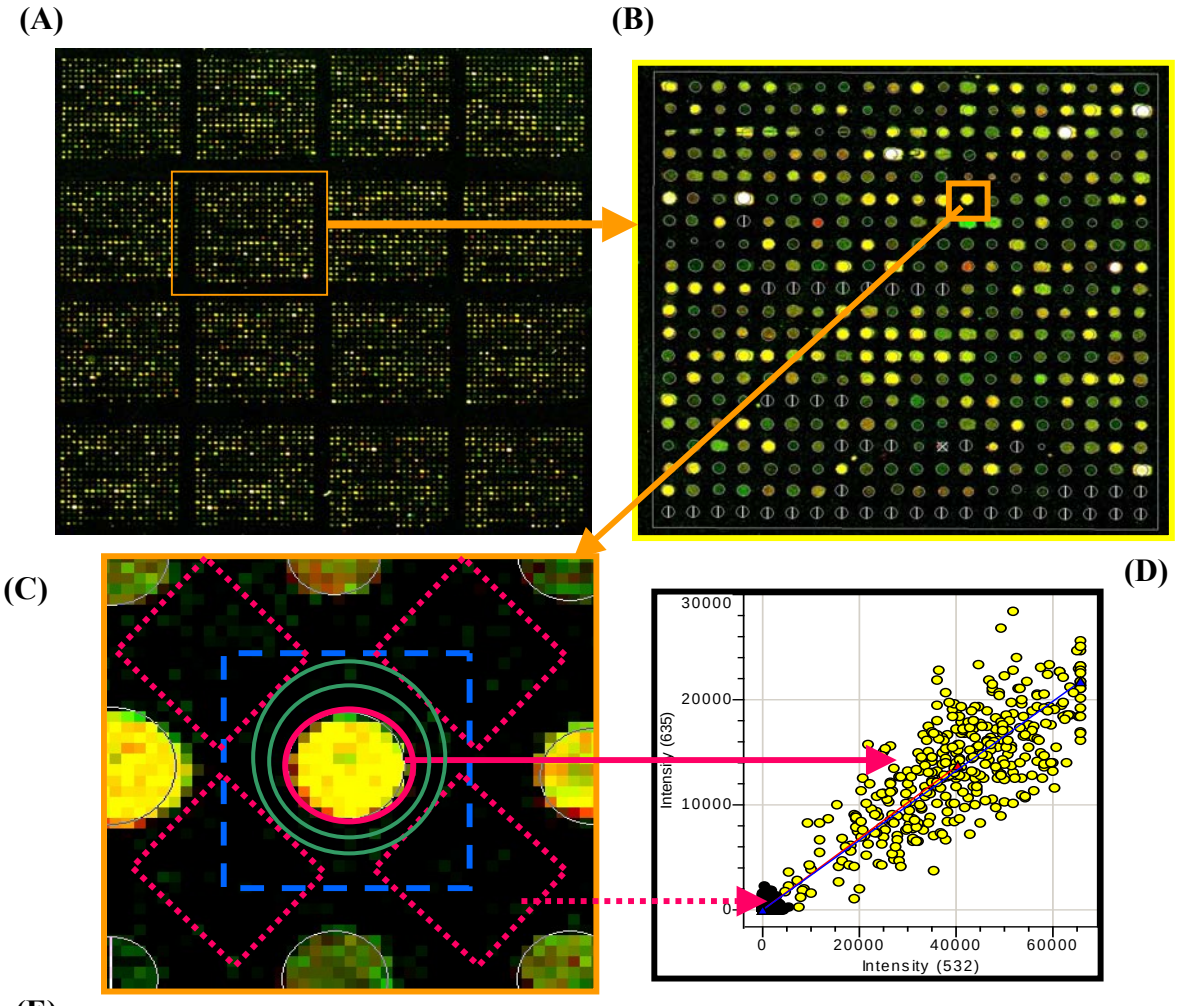


Figure 10. Exemples de plans expérimentaux : mode de comparaison des échantillons. T_x représentent les échantillons à comparer (tests), Ref. représente l'échantillon de référence pour les comparaisons indirectes. (A) Dessin à une référence (*Reference design*), tous les échantillons sont comparés à la même référence. (B) Variante du dessin à une référence avec une inversion des fluorochromes (*Dye-swap design*). (C) Dessin en boucle (*Loop design*) pour une comparaison direct des échantillons.

Pour construire une base de données de portraits moléculaires de patients atteints de différentes pathologies cardiovasculaires ou de différentes formes de cancers, il est intéressant d'analyser les échantillons indirectement par rapport à une référence commune. Dans ce cas le plan expérimental le plus employé est le « ***Reference design ».*** Ce mode consiste à utiliser pour l'un des échantillons hybridé sur l'ensemble des puces le même ARNm dit de « référence », par exemple un pool des ARNm « contrôles » (Fig. 10A). Généralement, l'échantillon de référence est marqué par le même fluorochrome pour toutes les puces. Les biais éventuels liés au marquage sont alors les mêmes pour l'ensemble des puces et n'affectent pas les comparaisons entre les échantillons. Toutefois, dans certaines conditions l'expérience nécessite une hybridation répétée des échantillons avec une inversion des fluorochromes ou « ***Dye swap*** » (Fig. 10B). En effet, si l'objectif est non seulement de comparer les échantillons entre eux mais aussi au pool de référence, une analyse en *dye swap* permet de prendre en compte la variabilité liée aux fluorochromes (Simon et Dobbin, 2003). Enfin, le plan expérimental en boucle simple ou « *simple* ***Loop design*** » (Fig. 10C) est une alternative au *reference design* qui permet une comparaison directe entre les échantillons. En générale, un *simple loop design* avec un petit nombre d'échantillon donne de bons résultats (Churchill, 2002). Cependant dès que le nombre d'échantillons augmente, le *loop design* simple est inapproprié. Si un expérimentateur désire évaluer toutes les paires de comparaisons entre 10 échantillons, le *loop design* simple est inefficace. De plus, la simple perte d'une lame pour cause de défauts, réduit considérablement la puissance de l'analyse. Les plans expérimentaux qui utilisent l'enchaînement de plusieurs boucles sont également possibles et peuvent être très puissants. Toutefois, l'interprétation, fonction de modèles mathématiques, devient plus complexe (Simon et Dobbin, 2003).

En bref

La distinction entre les variabilités techniques et les variations biologiques passe par l'optimisation du nombre de *replicates* biologiques et techniques ainsi que par la randomisation et/ou le blocage des facteurs de nuisances. Le plan expérimental conditionne la qualité des données d'expression et les interprétations qui en découlent.



(E)

Spot	Block	Column	Row	Dia.	F633 Median	F633 Mean	B633 Median	F633 % Sat.	F532 Median	F532 Mean	B532 Median	F532 % Sat.	Ratio of Medians	Rgn Ratio	F633 Median - B633	F532 Median - B532	Flags
1	1	1	1	120	6520	5783	0	0	12738	11534	0	0	0.512	0.469	6520	12738	-50
2	1	2	1	200	0	1238	0	0	75	2341	0	0	0.000	0.519	0	75	-100
3	1	3	1	200	0	399	0	0	34	721	0	0	0.000	0.459	0	34	-100
4	1	4	1	200	0	343	0	0	0	734	0	0	0.000	0.467	0	0	-100
5	1	5	1	200	0	183	0	0	1	475	0	0	0.000	0.238	0	1	-100
6	1	6	1	200	0	710	0	0	0	900	0	0	0.000	0.736	0	0	-100
7	1	7	1	200	0	207	0	0	133	510	0	0	0.000	0.054	0	133	-100
8	1	8	1	200	0	226	0	0	0	556	0	0	0.000	0.312	0	0	-100
9	1	9	1	200	0	183	0	0	3	368	0	0	0.000	0.271	0	3	-100
15	1	15	1	200	0	40	0	0	0	176	0	0	0.000	0.010	0	0	-100
16	1	16	1	200	0	28	0	0	0	196	0	0	0.000	0.005	0	0	-100
17	1	1	2	160	1067	1094	0	0	1629	1877	0	0	0.655	0.381	1067	1629	0
18	1	2	2	180	7800	8144	0	0	17947	18518	0	0	0.435	0.416	7800	17947	0
19	1	3	2	190	65535	64348	0	95	65535	63751	0	93	1.000	1.015	65535	65535	0
20	1	4	2	200	10781	10734	0	0	14067	14340	0	0	0.766	0.738	10781	14067	0
21	1	5	2	200	4219	4234	0	0	6116	6303	0	0	0.690	0.625	4219	6116	0
22	1	6	2	210	785	851	0	0	1297	1416	0	0	0.605	0.369	785	1297	0
23	1	7	2	190	3061	3122	0	0	6861	6898	0	0	0.446	0.420	3061	6861	0
24	1	8	2	200	2913	3095	0	0	5235	5567	0	0	0.556	0.514	2913	5235	0
25	1	9	2	210	2101	2178	0	0	3798	3991	0	0	0.553	0.488	2101	3798	0
26	1	10	2	210	2668	2701	0	0	3904	4131	0	0	0.683	0.576	2668	3904	0
27	1	11	2	210	14446	14090	0	0	22985	23311	0	0	0.628	0.602	14446	22985	0
28	1	12	2	200	5285	5471	0	0	7472	7468	0	0	0.707	0.717	5285	7472	0

Figure 11. Acquisition des données par les logiciels d'analyse d'images. (A) Extrait d'une image composite de puce à ADN *Myochip*. (B) Localisation et segmentation des *spots* : zoom sur un cadran de la puce. La grille de localisation et de segmentation est matérialisée en blanc. Les spots défectueux et les spots non identifiés (absents) sont respectivement marqués d'une croix et d'un trait vertical. Les *spots* atteignant la saturation sont blancs. (C) Exemples de méthodes de segmentation. Selon Scanalyze, le signal est estimé à l'intérieur du cercle rose et le bruit de fond à l'intérieur du carré bleu. D'après Genepix® le cercle rose définit les pixels appartenant au signal ; les carrés en pointillé rose matérialisent les zones d'estimation du bruit de fond. Les cercles concentriques vert définissent la zone de calcul du bruit de fond dans les logiciels Imagene® et QuantArray®. La « zone tampon » est définie par l'intervalle entre le cercle rose du spot, les carrés roses et le premier cercle vert ; elle limite l'introduction des valeurs aberrantes (artefacts) dans la mesure des signaux.

3. Acquisition des données d'expression par les logiciels d'analyse d'images

Suite à la lecture des puces à ADN par un scanner (*cf.* p. 10), les niveaux d'expression sont estimés grâce à des logiciels d'analyse d'images[25]. Ces logiciels extraient des informations qualitatives et semi-quantitatives pour chaque *spot* dans chacun des fluorochromes. Le traitement des images est un aspect clé de l'extraction des données de puces à ADN. L'interprétation biologique des données, comme le nombre de *gene reporters* détectés, dépend en partie de la qualité des logiciels d'analyse d'images (Le Meur, 2001 – Rapport de DEA bio-informatique).

Globalement, les logiciels d'analyse d'image sont basés sur le même principe et possèdent la même procédure de traitement qui se déroule en trois étapes (Fig. 11):

(i) **localisation** des *spots* sur la puce,

(ii) **segmentation** de l'image en pixels appartenant aux signaux et aux bruits de fond,

(iii) **extraction des données** qualitatives et semi-quantitatives.

Figure 11. **(Suite de la légende)** (D) Représentation graphique du résultat de la segmentation d'un spot. Graphique représentant la distribution des intensités des pixels appartenant au signal (jaune) et au bruit de fond (noir). Les niveaux d'intensités en Cy3 sont représentés en abscisse. Les niveaux d'intensités en Cy5 sont en ordonnées. Les pixels distribués verticalement sont saturés ; ils avoisinent des intensités proches de 65635. (E) Extrait d'un fichier issu le l'analyse qualitative et semi-quantitative d'une image de puce par le logiciel Genepix®. Le *spot* est identifié par un numéro dans la colonne S*pot*. Les colonnes *Block*, *Column* et *Row* définissent les coordonnées de chaque *spot* sur la puce. *F633* et *F532* sont respectivement les niveaux d'intensités des signaux en Cy5 et Cy3. *B633* et *F532* correspondent respectivement aux niveaux des bruits de fond en Cy5 et Cy3. F333Median-B633 et F532Median-B532 sont les intensités médianes des signaux corrigées de leur bruit de fond. Les pourcentages de pixels saturés par spots sont indiqués dans les colonnes *F633% Sat*. et *F532% Sat. Ratio of Medians* et *Rgn Ratio* sont deux modes d'estimation du rapport des signaux Cy3 et Cy5. Les *Flags* correspondent au mode d'identification des spots défectueux par le logiciel d'analyse d'image.

[25] http://genomicshome.com/

Tableau 4. Exemples d'algorithmes de segmentation implémentés dans les logiciels d'analyse d'images de puces à ADN.

Algorithme	**Géométrie des spots**	**Propriété de l'image**	**Inconvénients**
Cercle fixe (*fixed circle*) Ex : *Scanalyze (Eisen)*	Cercle de diamètre **fixe**	Segmentation basée le positionnement des *spots*.	Sensible aux - contaminations, - irrégularités de taille et de forme des spots Ajustement manuel coûteux en temps
Cercle adaptatif (*adaptative circle*) Ex : Genepix® ;Imagene®	Cercle au diamètre **auto ajusté**	Segmentation basée sur les niveaux d'intensité et le positionnement des *spots.*	Sensible aux - contaminations, - irrégularités de taille et de forme des spots
Test de *Mann-Whitney* ou Histogramme des intensités (histogram) Ex : QuantAray®,Imagene®	Non fixe (base circulaire)	Segmentation basée sur les niveaux d'intensité et le positionnement des *spots.* Définition d'un seuil d'intensité pour discriminer les pixels appartenant au signal ou au bruit de fond : - Test des rangs de Mann-Whitney - Quantiles de l'histogramme de distribution des intensités des zones du signal et du bruit de fond	Sensible aux - contaminations, - bruit de fond élevé Seuil fixe
Contour (*adaptative shape*) Ex : Genepix®	Contour exact du spot	Segmentation basée sur les niveaux d'intensité et le positionnement des *spots* Définition de proche en proche des pixels appartenant au signal ou au bruit de fond	Sensible aux - Contaminations, - Spots de petites tailles

La **localisation des spots ou « adressage » vise à préciser les coordonnées de chaque *spot* sur l'image à l'aide de grilles** (Fig. 11B). La structure d'une image de puce dépend de la configuration du robot de dépôts. Dans le cas d'un automate à aiguilles, chacune d'elles génère un cadran (bloc) de taille et de position connue qui se compose de *spots*. Le logiciel d'analyse d'image délimite les cadrans par une grille. Pour localiser un spot sur une image, c'est à dire pour faire correspondre un modèle idéal de puce avec une image scannée, de nombreux paramètres doivent être estimés et ajustés. Par exemple, les espaces entre les lignes et colonnes des cadrans, et entre les spots d'un même cadran, doivent être définis. Les possibles mouvements de translation des spots ou des cadrans, liés à la variation de position des aiguilles du robot de dépôt, doivent être détectés. La qualité d'un logiciel d'analyse d'image est notamment évaluée par sa capacité à repérer rapidement et automatiquement ces défauts, et à ajuster le positionnement des grilles sur les spots. L'efficacité des étapes ultérieures dépend de la précision des coordonnées des spots.

La segmentation est définie comme le processus de découpage de l'image en différentes régions, ayant chacune leurs propriétés physiques et géométriques. Pour un *spot* donné, **la segmentation permet de classer les pixels en tant que signal et bruit de fond** (Fig. 11C-D). Dans le processus de l'analyse d'image, la phase de segmentation est l'étape qui influe le plus sur les données. Les problèmes majeurs dans la conception d'un algorithme de segmentation sont l'irrégularité des *spots* et l'hétérogénéité du bruit de fond, respectivement liées, par exemple, à un défaut des aiguilles du robot de dépôt ou à l'auto-fluorescence des lames de verre (Yang *et al.*, 2001a).

ScanAlyze, premier logiciel académique pour l'analyse d'images des puces à ADN (Eisen et Brown, 1999), propose une segmentation en cercle fixe et un ajustement manuel de la définition de la zone d'un spot. Cette approche est relativement sensible aux valeurs aberrantes et à l'irrégularité des *spots*. De plus, la correction manuelle est particulièrement fastidieuse et augmente les erreurs liées à la subjectivité de l'expérimentateur. Le manque de souplesse de *ScanAlyze* a rapidement motivé le développement d'autres produits académiques (Adams et Bischof, 1994) et commerciaux (Genepix®, Imagene® ou QuantArray®) avec des méthodes de segmentation plus puissantes. Actuellement, les principales méthodes de segmentation se répartissent en quatre catégories, fonctions de la géométrie du *spot* et des propriétés de l'image (Tab. 4). Les méthodes les plus robustes utilisent les propriétés spatiales des images telles que le positionnement des *spots* et la distribution des intensités des pixels. Une zone dite « tampon » entre signal et bruit de fond permet de limiter l'effet des valeurs

aberrantes (Fig. 11C). Les approches par « histogramme » de Image® et «cercle adaptatif» de Genepix® sont parmi les plus exactes. Une bonne méthode d'estimation du signal est caractérisée par une absence de corrélation entre le niveau du signal et le bruit de fond environnant (Fig. 12). Lorsqu'ils sont corrélés, la mesure du signal risque d'inclure celle du bruit de fond. Ce point est particulièrement important pour les spots de faible intensité. La soustraction du bruit de fond a alors une grande influence sur les ratios.

L'extraction des données est qualitative et semi-quantitative (Fig. 11E). Le nombre de mesures par *spot* varie en fonction des logiciels d'analyse d'images. Les données qualitatives sont, par exemple, le diamètre, la surface ou encore le pourcentage de pixels saturés dans chaque *spot*. Les mesures semi-quantitatives sont, notamment, les intensités (moyenne ou médiane) des signaux et du bruit de fond, et le rapport (*ratio*) des intensités obtenues pour chaque fluorochrome. Ces mesures sont qualifiées de données « brutes » ou **données primaires** car elles nécessitent d'être traitées et validées avant toutes analyses. Ces données sont généralement sauvegardées dans des fichiers de format texte (*.txt,*.dat, *.gpr) pouvant être traité dans des tableurs types Excel® ou des logiciels de statistiques comme R (Ihaka et Gentleman, 1996).

En bref

Le traitement des images est un aspect clé de l'extraction des données de puces à ADN. Il existe de nombreux logiciels d'analyse d'images dont les qualités conditionnent la qualité des mesures et, par conséquent, l'interprétation des données.

Les logiciels d'analyse d'images extraient pour chaque *spot* des informations qualitatives et semi-quantitatives dites données « primaires ». Les données qualitatives permettent une validation des mesures. Les données semi-quantitatives offrent, au travers des ratios d'intensités, une quantification relative des niveaux d'expression entre les cibles.

Compte tenu de la quantité d'information générée et de l'aspect bruité de ces données, des traitements numériques automatisés sont nécessaires à la validation des mesures.

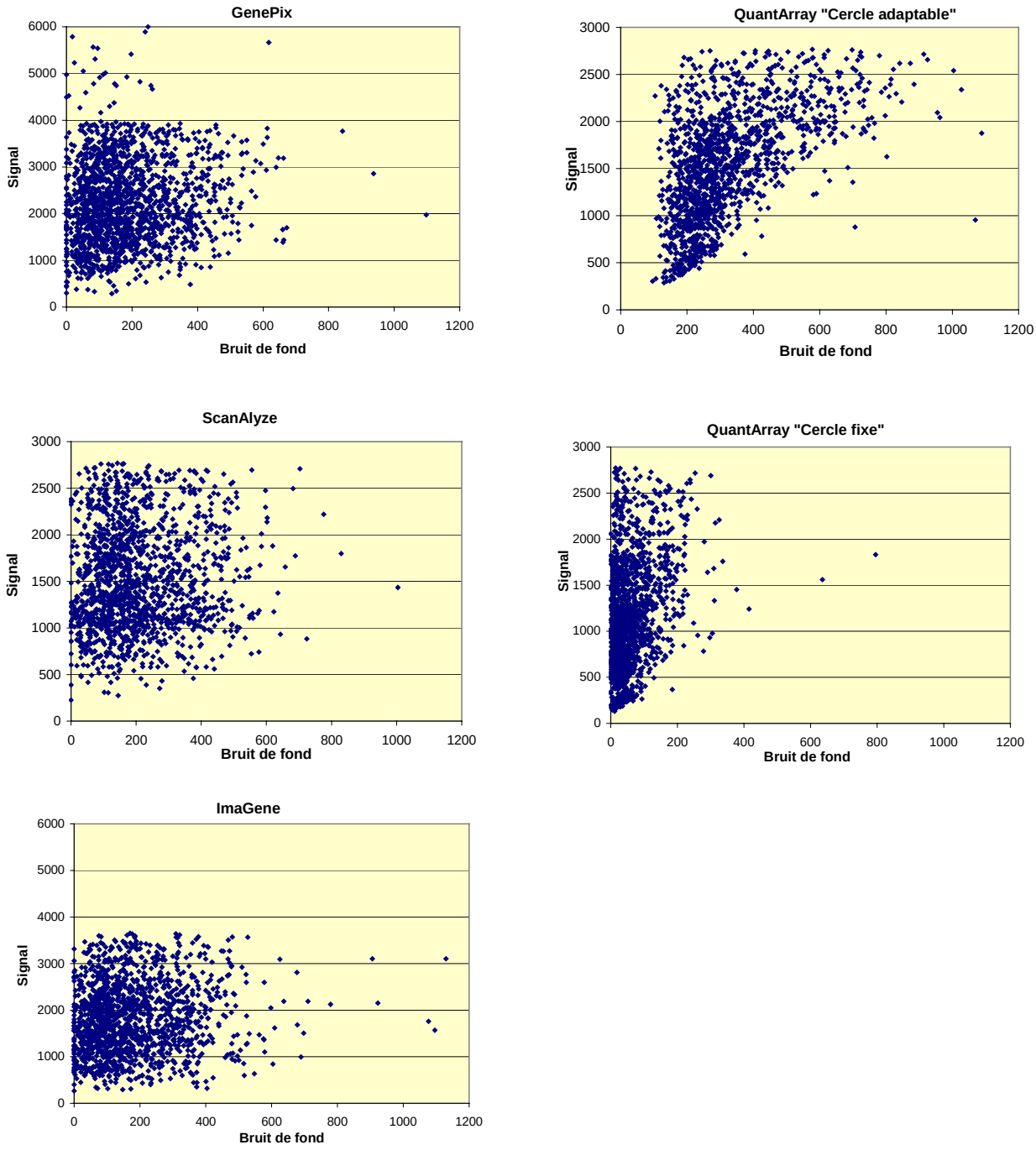


Figure 12. Importance de la méthode de segmentation dans les logiciels d'analyse d'images de puces à ADN. Représentation graphique de la corrélation entre l'intensité des spots en Cy3 et leur bruit de fond selon les logiciels d'analyses d'image ScanAlyze, Genepix® (Axon, Inc.), QuantArray®.(PerkinElmer, ,Inc) et ImaGene (BioDiscovery, Inc.). Les algorithmes de segmentation utilisés sont : « cercle adaptatif » pour Genepix®, Imagene® et QuantArray® et la méthode en cercle fixe pour ScanAlyze et QuantArray®. Aucune corrélation n'existe entre l'estimation des bruits de fond et des signaux associés pour les logiciels ScanAlyze, GenePix et ImaGene. Les valeurs de bruit de fond et de signal, selon QuantArray, tendent vers une corrélation positive.

II. MADSCAN : historique et développement des outils

Compte tenu de la densité des puces (~20 000 à 40 000 *reporters*) et du nombre de mesures par *reporter*, un traitement manuel des données primaires est rapidement devenu trop coûteux en temps, avec un risque d'erreur croissant lié à la subjectivité de l'expérimentateur. De nombreux algorithmes et outils informatiques pour le traitement des données primaires ont alors été développés. La complexité d'utilisation de certains de ces outils (manque d'interface graphique, définition de nombreux paramètres...) nous a motivé pour développer MADSCAN, ou ***Microarray Data Suite of Computed Analysis***. Ce logiciel, qui se veut simple d'utilisation, applique l'ensemble des traitements nécessaires à l'obtention de données « consolidées » d'expression à partir des données primaires issues des logiciels d'analyse d'images.

1. Ebauche de MADSCAN : macro Excel®

L'outil MADSCAN a tout d'abord été développé en Visual Basic pour Microsoft Excel®. Les premières fonctionnalités de cette macro complémentaire ont été la filtration et la normalisation linéaire des données primaires issues d'une puce à ADN. L'outil a ensuite rapidement évolué pour prendre en compte le développement des nouvelles méthodes de traitement des données mieux adaptées à la technologie des puces à ADN.

Les modes de normalisation ont rapidement changés et les approches non linéaires se sont avérées plus efficaces pour minimiser les biais expérimentaux (Tseng *et al.*, 2001; Yang *et al.*, 2002; Workman *et al.*, 2002; Faller *et al.*, 2003). Parmi les nouvelles techniques proposées, la librairie *sma*, développée en langage R par Yang *et al.* (2001), a beaucoup influencé le domaine du traitement des données primaires issues des expériences de puces à ADN. Ils ont proposé une normalisation non linéaire spatiale qui permet, entre autres, de minimiser l'effet des marqueurs fluorescents et l'effet des pointes (ou aiguilles) du robot de dépôts des sondes (*cf.* Tab. 3 p. 29). Parallèlement, Tseng *et al.* (2001) ont suggéré une recherche de gènes de références à partir desquels les coefficients de corrections non linéaires sont calculés. Cette méthode a également été développée en R sous la forme d'un *script*.

Nous avons choisi d'utiliser une combinaison des méthodes proposées par Yang *et al.* (2001) et Tseng *et al.* (2001) pour améliorer notre méthode de normalisation des données d'expression. Ce changement de techniques nous a conduit à utiliser R en complément de Microsoft Excel®. La liaison entre R et Microsoft Excel® a été réalisée grâce à un serveur de communication nommé RDCOMServer (Neuwirth et Baier, 2001).

Cette évolution a nettement amélioré les performances des traitements et la qualité des résultats. De plus, l'expérience a montré que cet outil était très utilisé. Les limites de puissance de calcul des stations de travail, le désir de proposer un outil portable et l'idée de développer de nouvelles fonctionnalités, comme le traitement par lot de puces, nous a incité à repenser l'outil. Nous avons choisi de transformer la macro complémentaire Excel® en une librairie R accessible *via* le Web.

2. Outils informatiques

La mise en place d'un service Web nécessite le développement d'un système client-serveur avec un poste serveur et des outils de liaison (logiciels et langages). Les paragraphes ci-dessous décrivent les choix que nous avons adoptés pour transformer MADSCAN en un service Web basé sur des formulaires HTML (Fig. 13).

2.1 Machine

L'ensemble des développements pour la gestion et le traitement des données issues des expériences de puces à ADN ont été réalisé sur un serveur dédié, nommé *cardioserve*. Cet ordinateur est une machine DELL PowerEdge 4600, dotée d'un processeur Xéon cadencé à 1,8 GHz et de 1Go de RAM. Le système d'exploitation de cette machine est actuellement Linux RedHat 9.0.

2.2 Serveur Web et interfaçage

Une application Web désigne un logiciel installé sur un serveur HTTP (***H**yper**T**ext **T**ransfer **P**rotocol*). Les postes clients se connectent sur cette application à l'aide d'un navigateur (Iexplorer, Netscape, Mozilla…). Le principe de fonctionnement repose sur la mécanique requête/réponse, le client interroge le serveur à l'aide d'une requête HTTP et le serveur retourne généralement une réponse sous la forme d'une page HTML (***H**yper**T**ext **M**arkup **L**anguage*).

a) Serveur Apache (v. 1.3.28)

Le **serveur Web** utilisé est **Apache** (Fig. 13). Cet outil compte parmi les serveurs Web *open-source* les plus utilisés pour l'interprétation des requêtes HTTP. Apache est indispensable pour l'interfaçage entre le serveur et l'utilisateur *via* les pages et formulaires HTML. Il permet également l'exécution de scripts développés côté serveur.

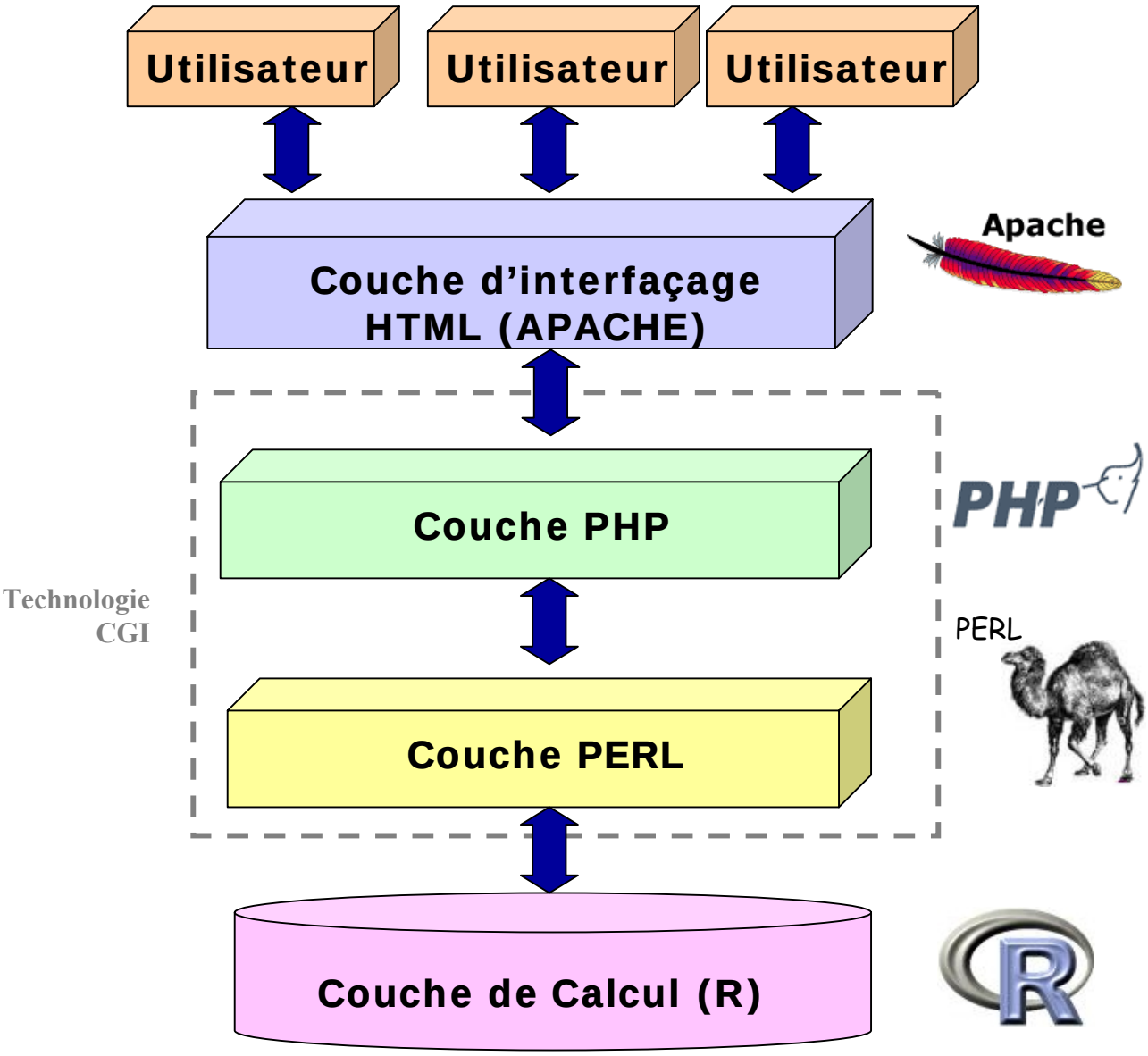


Figure 13. Outils informatiques pour la mise en place de MADSCAN sous la forme d'un service Web. Apache est le serveur Web qui permet l'interfaçage entre le serveur et l'utilisateur via les pages et formulaires HTML/PHP. La technologie CGI est utilisée pour interfacer les différentes applications. Les langages de script utilisés pour la technologie CGI sont PHP et PERL. Les scripts PHP sont utilisés pour la construction des pages de formulaire et optimiser le transfert des requêtes des utilisateurs aux scripts PERL. PERL permet l'exploitation des formulaires de MADSCAN, l'interaction avec les scripts R et l'affichage des résultats. R est le langage utilisé pour l'implémentation de l'ensemble des fonctions mathématiques développées dans MADSCAN.

b) Technologie CGI

La technologie CGI ou ***C**ommon **G**ateway **I**nterface* a été utilisée pour **interfacer les différentes applications** (Fig. 13). Cette norme définit l'interfaçage d'applications externes (par exemple des pages HTML) avec des serveurs d'information (comme une base de données) (Gundavaram 1996). En d'autres termes, cette méthode transfert les requêtes des internautes à des programmes installés sur les serveurs Web, puis des serveurs aux internautes. Ce modèle de programmation permet la génération de pages Internet dynamiques en réponse aux requêtes formulées par les utilisateurs. En effet, un document HTML est statique. Ce document est un fichier texte dont l'information ne change pas tant qu'il n'est pas édité. Grâce à la technologie CGI, l'information devient dynamique. Un programme peut être exécuté en temps réel.

CGI est souvent confondu avec PERL, qui est un langage de programmation, alors que CGI est une méthode. Le langage PERL (comme C++, Java, Python ou PHP) permet de créer une application de type CGI. L'utilisation de différents langages offre une beaucoup de souplesse. Il permet notamment l'adaptation de programmes déjà existants en un service Web. Enfin, le protocole d'échange entre le serveur Web et le programme qui affiche le contenu HTML est supporté par la majeure partie des systèmes d'exploitation existants.

2.3 Langages de script

Les langages de script ne nécessitent pas de compilations préalables, ils sont « interprétés » à la volée par un programme auxiliaire : l'interpréteur.

a) Langage R (v. 2.0.1)

Le langage R est utilisé pour **l'implémentation de l'ensemble des fonctions mathématiques** développées dans MADSCAN (Fig. 13). Ce langage a été choisi pour ces nombreux avantages (cf. p. 20-23):

(i) codes sources libres, et donc réutilisables,
(ii) codes interprétés, par conséquent rapide à déboguer,
(iii) possibilité de construire des librairies de fonctions,
(iv) documentation très riche.

De plus, les programmes R sont utilisables en tâche de fond. Ils peuvent être appelés par d'autres programmes tels que des scripts PERL sans que l'utilisateur ne s'en aperçoive.

b) Langage PERL (v 5.8.1)

PERL (***P**ractical **E**xtraction and **R**eport **L**anguage*) a principalement été choisi pour **l'exploitation des formulaires de MADSCAN, l'interaction avec les scripts R et l'affichage des résultats**. PERL est un langage de programmation dérivé des scripts *shell*. Il a été créé en 1986 par Larry WALL afin de mettre au point un système de « *News* » entre deux réseaux. PERL est un langage interprété caractérisé par un typage faible. Sa fonctionnalité principale est la manipulation de chaînes de caractères (d'où son utilisation en bio-informatique pour les analyses de séquences nucléiques et protéiques). Il est également pleinement adapté à la gestion des fichiers et des répertoires. Enfin, de nombreuses fonctionnalités, comme l'envoi de mails, peuvent être ajoutées aux scripts PERL grâce à différents modules disponibles notamment sur le site Web du CPAN[26] (***C**omprehensive **P**erl **A**rchive **N**etwork*).

c) Langage PHP (v. 4.3.3)

Ce langage a été choisi pour **construire les pages de formulaire et optimiser le transfert des requêtes des utilisateurs aux scripts PERL**. PHP, acronyme de « ***P**ersonal **H**ome **P**age* », est un langage interprété mis au point en 1994 par Rasmus LERDORF afin de détecter les visiteurs de sa page Internet personnelle. PHP est un langage exécuté côté serveur dont la syntaxe s'inspire des langages C et PERL. En effet, PHP possède un panel de fonctions très étendu allant de la simple génération de documents HTML à la production d'images GIF animées en passant par l'envoi automatique de courriers électroniques. En l'espace de quelques années, il est devenu le langage de référence des sites Internet à pages dynamiques.

3. Accès sur la Toile

MADSCAN est disponible sur le Web[27] depuis 2003 (Fig. 14). Il appartient à MADTOOLS (***M**icroArray **D**ata **TOOLS**),* un ensemble d'outils Web développés au sein du laboratoire pour gérer et analyser les données de puces à ADN.

[26] http://www.cpan.org/
[27] http://www.madtools.org

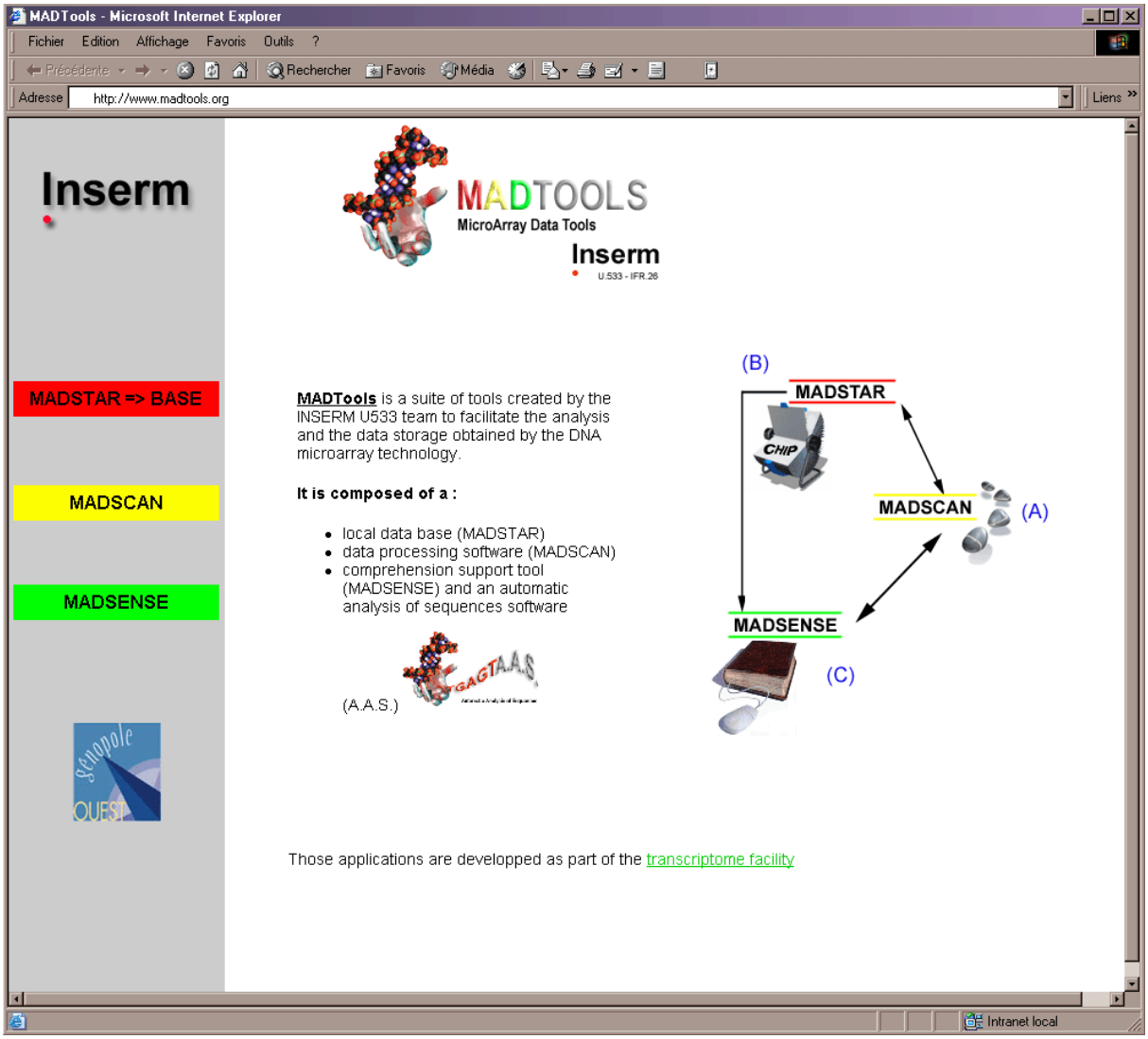


Figure 14. MADTOOLS: MicroArray Data TOOLS. (A). MADSTAR-BASE, Acronyme de *MicroArray Data STorage And Retrieval plus BASE*, est une base de données qui permet la sauvegarde et la gestion de l'ensemble des expériences (étapes expérimentales et résultats). (B). MADSCAN (*MicroArray Data Suite if Computed Analysis*) transforme les données brutes de puces à ADN (2 couleurs) en matrice de données consolidées (C) MADSENSE (*MicroArray Data SENSE*) annote les gènes du génome humain. Elle intègre les données biologiques et bibliographiques de divers banques de données publiques sous la forme d'une « carte de visite » pour chaque gène.

3.1 MADTOOLS

MADTOOLS se compose de 3 logiciels (Fig. 14) :

- MADSTAR/BASE, pour la gestion des données,
- MADSCAN, pour le traitement des données primaires,
- MADSENSE, pour l'annotation des gènes.

La sauvegarde et la gestion des données expérimentales ont tout d'abord été effectuées par MADSTAR (***M**icroArray **D**ata **ST**orage **A**nd **R**etrieval),* une base de données développée au laboratoire par Audrey Bihouée (Fig. 14B). Récemment, Audrey Bihouée, avec l'aide de Jonathan Brosseau (Brosseau, 2002), ont transféré l'ensemble des données de MADSTAR dans BASE[28] (Saal *et al.*, 2002) afin d'améliorer notre capacité de gestion et d'analyse des données. BASE offre, par exemple, une administration plus flexible des utilisateurs, assurant la confidentialité des différents travaux. Elle structure les données selon la norme internationale MIAME (Brazma *et al.*, 2001) et permet la prise en charge du format d'échange MAGE-ML[29] (Spellman *et al.*, 2002) pour une exportation facilité des données vers les banques de données publiques comme ArrayExpress[30]. De plus, BASE évolue rapidement grâce une importante communauté d'utilisateurs et de développeurs. Enfin, elle offre de nombreux outils de visualisation et permet le développement de modules (*plug-ins)* pour le traitement et l'analyse des données[31].

MADSENSE (Fig. 14C), acronyme de *MicroArray Data SENSE*, est un outil d'annotation des gènes développé par Raluca Teusan, au cours de son stage de DEA en bio-informatique et ensuite au sein du laboratoire. Pour un gène donné, cet outil intègre sous formes de « cartes de visite » les données biologiques et bibliographiques présentes dans les différentes bases et banques de données publiques telles SOURCE (Diehn *et al.*, 2003) ou Pubgene (Jenssen *et al.*, 2001). Les informations disponibles dans la « carte de visite » sont par exemple le symbole officiel du gène défini par le HGNC[32], ses alias, sa position chromosomique ou encore les maladies dans lesquelles il est possiblement impliqué. Les requêtes peuvent être faîtes gène par gène ou au moyen d'une liste de gènes.

[28] http://base.thep.lu.se/
[29] http://www.mged.org/Workgroups/MAGE/mage-ml.html
[30] http://www.ebi.ac.uk/arrayexpress/
[31] http://base.thep.lu.se/plugins/
[32] http://www.gene.ucl.ac.uk/nomenclature

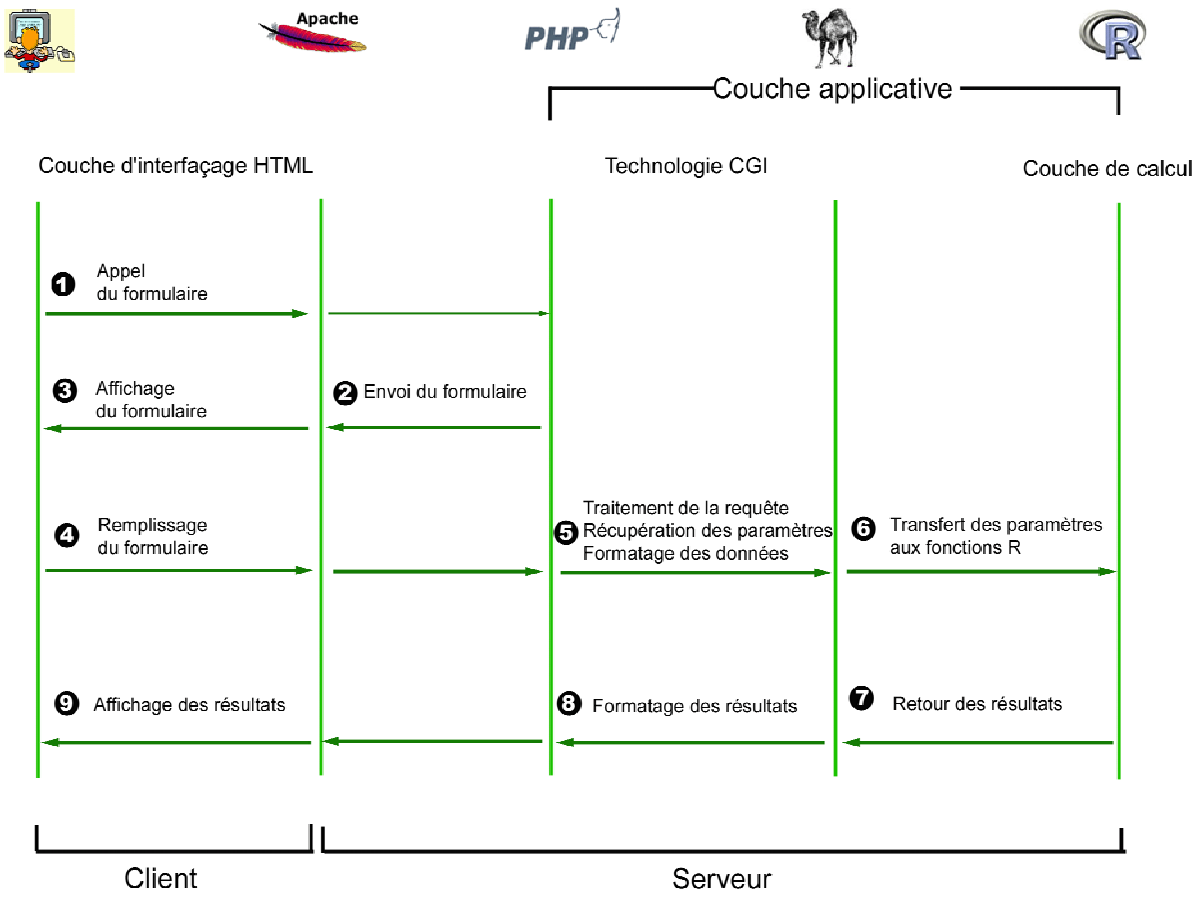


Figure 15. Modélisation dynamique de MADSCAN. (1) Connexion au site Web par un explorateur Internet et demande du formulaire d'analyse par l'utilisateur. (2) réception de la demande par le serveur Apache ; envoi et (3) affichage du formulaire d'analyse à l'utilisateur au niveau de son explorateur Internet. (4) Remplissage du formulaire par l'utilisateur et envoi aux programmes d'analyse. (5) Réception et formatage des données du formulaire par le programme CGI. (6) Transfert des paramètres au programme R de traitement des données. (7) Retour des résultats générés par le programme R au programme CGI. (8) Formatage et (9) affichage des résultats à l'utilisateur au niveau de son explorateur Internet.

3.2 MADSCAN : service Web

a) Librarie madscan

MADSCAN est avant tout une librairie de fonctions R nommée *madscan*. La librairie *madscan* compte environ 60 fonctions dont certaines ont été adaptées des travaux de Tseng *et al.* (2001) et de la librairie *sma* de Yang *et al.* (2002). Ces dernières sont utilisées, respectivement, pour le calcul des « gènes invariants » et le tracé des courbes de normalisation (*cf.* p.42-44, Annexe I).

L'ensemble des théories mathématiques utilisé pour le traitement des données est détaillé dans un manuel (en langue anglaise) disponible en ligne (*cf.* Annexe I). Ce manuel décrit les algorithmes utilisés, les transformations effectuées et les modes d'utilisation de MADSCAN.

b) Déroulement des événements

MADSCAN fonctionne sur le principe d'un service Web avec une interface client-serveur de type CGI (Fig. 15). L'utilisateur se connecte à MADSCAN *via* un explorateur Internet (Internet Explorer ou Mozilla) (Fig. 15-❶). Il demande l'accès à un formulaire d'analyse. Le serveur Web Apache retourne le formulaire (Fig 15-❷ ; Fig. 16A) dédié à l'analyse demandée et l'affiche (Fig. 15-❸). L'utilisateur renseigne le formulaire et dépose les données à traiter sous format compressé, *i.e.* *.zip (Fig. 15-❹ ; Fig. 16A). Les fichiers à analyser sont téléchargés sur le serveur (Fig. 15-❺). Les données du formulaire sont interprétées par un programme CGI écrit en PERL et R (Fig. 16B).

La première partie de ce programme, rédigée en PERL, formate les données extraites du formulaire. La seconde section du programme, écrite en R, récupère les paramètres (Fig. 15-❻ Fig. 16B) et construit le fichier *Launch.R* dont le contenu correspond à l'enchaînement des fonctions appelées dans *madscan* (Fig. 16B). R est invoqué en tâche de fond et le contenu du fichier *Launch.R* est exécuté. Au cours de l'analyse, un fichier *Result.R* répertorie les fonctions de *madscan* utilisées. En cas d'erreur d'exécution, les messages d'erreurs sont également suavegarder dans *Result.R* (Fig. 16B). *Result.R* reste du côté serveur (*i.e.* n'est pas retourné au client) et aide à déboguer les programmes.

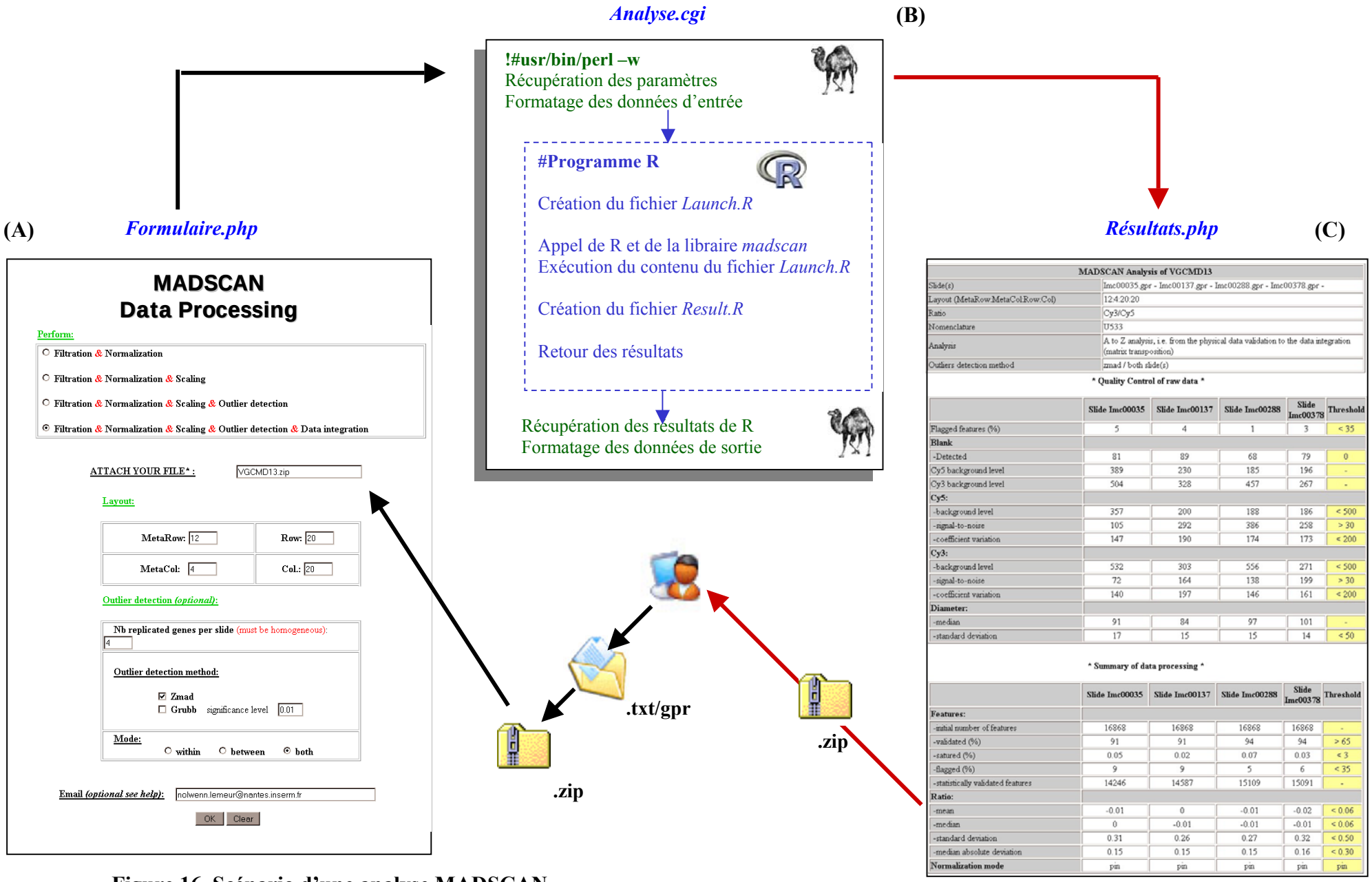


Figure 16. Scénario d'une analyse MADSCAN.

Les résultats obtenus par R sont sauvegardés dans des fichiers textes (Fig. 15-❼). Dans la dernière partie du programme CGI, certains de ces fichiers sont analysés par le script PERL (Fig. 15-❽) pour afficher à l'écran un résumé des résultats (Fig. 15-❾ ; Fig. 16C). Des représentations graphiques sont également créées et enregistrées sous forme d'images JPEG. Les fichiers textes, générés au cours de l'analyse, sont téléchargeables par les utilisateurs soit directement après l'affichage des résultats à l'écran soit après réception d'un courrier électronique notifiant la fin de l'analyse.

Le manuel (*cf.* Annexe I) présente sous la forme d'un exemple les champs des formulaires à remplir suivant l'analyse demandée. Il décrit également les fichiers résultats retournés. Enfin, des pages d'aide en ligne rappellent également la manière de remplir les formulaires.

Figure 16. Scénario d'une analyse MADSCAN.

(A) Formulaire d'analyse à remplir par l'utilisateur : spécifier l'analyse demandée (filtration + normalisation ou filtration + normalisation+ *scaling*, ...) ; insérer le dossier .zip à analyser, dossier contenant les différents fichiers texte (.txt) ou Genepix® (.gpr) des données d'expression primaires des lames à analyser ; définir la configuration des puces analysées (nombre de lignes et colonnes de cadrans – *MetaRow, MetaCol* –, nombre de lignes et colonnes par cadran - *Row, Col*); indiquer le nombre et le mode de détection des valeurs aberrantes ; indiquer votre adresse *email*.

(B) Récupération des données et paramètres d'analyse par le programme CGI : Formatage des données par un script PERL, transfert des paramètres aux fonctions R pour la création du fichier *Launch.R,* lancement de R et de la librairie *madscan* en tâche de fond pour l'exécution du contenu de *Launch.R;* création du fichier *Result.R, i.e.* compte rendu pour le programmateur du bon ou mauvais déroulement de l'analyse*;* retour des résultats sous forme de fichiers textes et formatage par le programme CGI/PERL de certaines données pour unaffichage à l'utilisateur.

(C) Affichage des résultats au niveau de l'explorateur Internet de l'utilisateur : Résumé des analyses demandées ; bilan de la qualité des données primaires, *i.e.* avant traitement, à savoir nombre de spots validés par le logiciel d'analyse d'image, niveau du rapport signal sur bruit... ; bilan de la qualité des données « consolidées », *i.e.* après traitement, à savoir nombre de spots validés et exploitables par la suite, mode de normalisation utilisée...Possibilité de récupérer l'ensemble des résultats (sous format .zip) via l'interface Web.

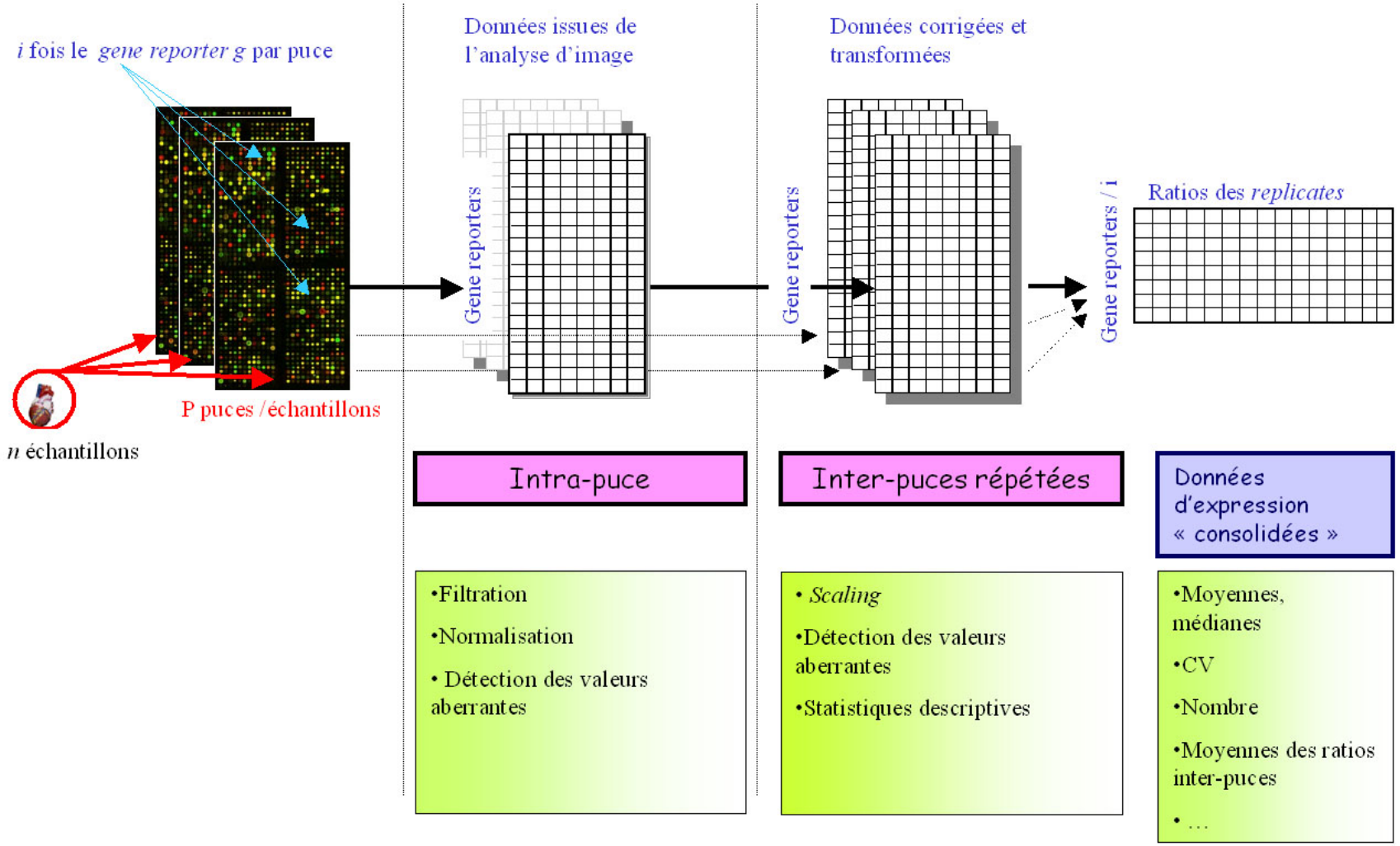


Figure 17. Suite des traitements numériques réalisés par MADSCAN, à l'intérieur et entre les puces répétées d'un jeu de données.

III. MADSCAN : traitement des données « primaires » de puces à ADN

1. Obtention de données « consolidées » dans les expériences de puces à ADN

Le plan de la procédure MADSCAN pour l'obtention des données « consolidées » est schématisé dans la figure 17.

1.1 Des images aux données d'expression

Une des difficultés de la technologie des puces à ADN est la quantité d'informations générée par les logiciels d'analyse d'images pour chaque *spot*. Choisir les données les plus informatives sur les mesures d'expression et leur qualité est donc indispensable. L'autre point critique est la suite des étapes mise en œuvre pour obtenir des données de qualité. A chaque étape des biais expérimentaux (*cf.* Tab. 3, p.20) peuvent entacher d'erreur la mesure finale. Les variations biologiques d'intérêt peuvent donc être masquées par des bruits techniques et biologiques. Aussi, outre la définition d'un plan expérimental adéquat (*replicates*, randomisation des facteurs de « nuisances »), il est nécessaire d'appliquer une procédure systématique de traitement et de transformation à ces données afin de minimiser (voire corriger) ces variations indésirables (Tseng *et al.*, 2001; Quackenbush, 2002).

a) Plan expérimental

Du point de vue expérimental, nous proposons de répéter les points de mesure à l'intérieur et entre les puces afin d'évaluer les bruits techniques et biologiques (Fig. 17). Pour estimer les biais techniques, nous suggérons un dessin de puce avec au minimum 3 fois le même *gene reporter*, déposé à différents endroits de la puce. Nous recommandons également de répéter les hybridations pour une même échantillon, soit au minimum 2 puces par échantillon. Enfin, pour apprécier les bruits biologiques, nous préconisons (au minimum) 2 extractions différentes (régionalement) pour un même tissu.

Le plan expérimental, et plus particulièrement le nombre de répétitions, détermine également la capacité de mettre en évidence les gènes différentiellement exprimés, *i.e.* la puissance des tests statistiques (*cf.* article p.45). En effet, nous avons montré que, grâce à un tel plan expérimental, il est possible de détecter statistiquement de faibles variations d'expression (~ 20%). De plus, la répétition des points de mesure à l'intérieur et entre des puces offre un meilleur contrôle du nombre de faux positifs (FP) et faux négatifs (FN).

b) Quelles données ?

Le premier choix est **le mode de calcul des intensités de chaque spot**. L'intensité correspond le plus à la moyenne ou à la médiane des intensités des pixels qui constituent le spot (soit pour une lecture à 10µm et des spots avec un diamètre de 100µm, ~200 pixels appartiennent au signal et 1200 appartiennent au bruit de fond local) (Fig. 18). MADSCAN utilise préférentiellement la médiane des intensités, étant moins sensible aux valeurs aberrantes. Pour un spot donné, ces intensités sont notées *Rmed* et *Gmed* pour, respectivement, la médiane des intensités des pixels en Cy5 (R pour *Red*) et en Cy3 (G pour *Green*).

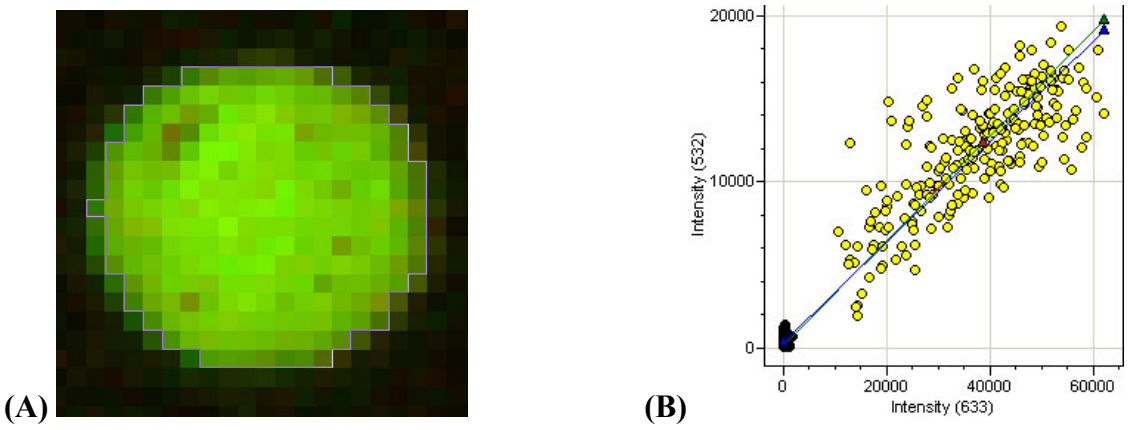


Figure 18. Segmentation et mode de calcul des niveaux d'intensités des *spots* (d'après le logiciel d'analyse d'images GenepixPro 5.0, Axon *Inc.*). **(A) Le mode de segmentation** appliqué est « *irregular feature* ». Le *spot* n'est pas nécessairement circulaire et l'algorithme de segmentation cherche à faire le contour exact du spot. **(B) Représentation graphique, pour chaque pixel du spot, des intensités en Cy3 *versus* les intensités en Cy5.** Quantification du niveau d'intensité des pixels appartenant au signal (jaune) et des pixels appartenant au bruit de fond (noir).

La première **correction** est **la prise en considération ou non du bruit de fond** des lames (Kooperberg *et al.*, 2002). Généralement, l'intensité du bruit de fond est soustraite de l'intensité du *spot*. L'estimation du niveau d'intensité du bruit de fond peut être faite localement (au niveau de chaque *spot*), par bloc ou en utilisant des spots « blancs » (Yang *et al.*, 2000; Draghici 2003). Par défaut, MADSCAN travaille avec le bruit de fond médian des *spots (Rbmed* et *Gbmed*, pour, respectivement, le bruit de fond médian en Cy5 et Cy3).

Enfin, les données d'expression retournées par MADSCAN sont présentées en **logarithme de base 2** sous la forme du **ratio des intensités** (**M**) et la **moyenne des intensités** (**A**) entre les 2 canaux. Le ratio en logarithme de base 2 est noté ***M***, pour *minus* (logCy3-logCy5), et la moyenne des intensités est nommée ***A*** pour *add* (1/2(logCy3+logCy5)) (Yang *et al.*, 2001b). Une transformation des ratios et des moyennes d'intensités en échelle logarithmique est effectuée systématiquement afin de travailler avec une distribution symétrique et quasi normale des valeurs. En effet, $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(1/2) = -1$, $\log_2(4) = 2$ et $\log_2(1/4) = -2$, *etc.* Le ratio offre ainsi une vision directe de la notion relative des niveaux d'expression ; à savoir la sur- ou sous-expression d'un gène dans une condition par rapport à une autre. La moyenne des intensités témoigne, quant à elle, du niveau moyen du signal. Par exemple, A=8 correspond à une intensité de 256 et A=16 à une intensité de 65536 soit la saturation.

c) Traitements des données primaires

Le traitement des données primaires (*preprocessing*) vise à minimiser l'effet des biais expérimentaux sur les mesures et écarter les données de mauvaise qualité. Les principales étapes de traitements couramment appliquées sont (Fig 17):

(i) filtration, afin d'écarter les *spots* défectueux (comètes, *spots* saturés),

(ii) normalisation,
(iii) *scaling,*
} pour minimiser les biais systématiques

(iv) détection des valeurs aberrantes, afin d'éliminer les mesures répétées non reproductibles.

MADSCAN propose l'ensemble de ces traitements en une seule procédure automatisée. Cette procédure peut être appliquée à une puce ou à plusieurs puces répétées (lot) en même temps.

L'étape de **filtration** des données primaires est complémentaire à la détection des spots non valides par les logiciels d'analyse d'images. La filtration des données primaires se base sur les critères de qualités physiques et géométriques des *spots* tels que le niveau de saturation du signal ou le rapport signal sur bruit (pour plus de détails sur les critères employés par MADSCAN, *cf.* Annexe I). Dans MADSCAN, un arbre de décision oriente la filtration des mesures et leur attribue un score de qualité.

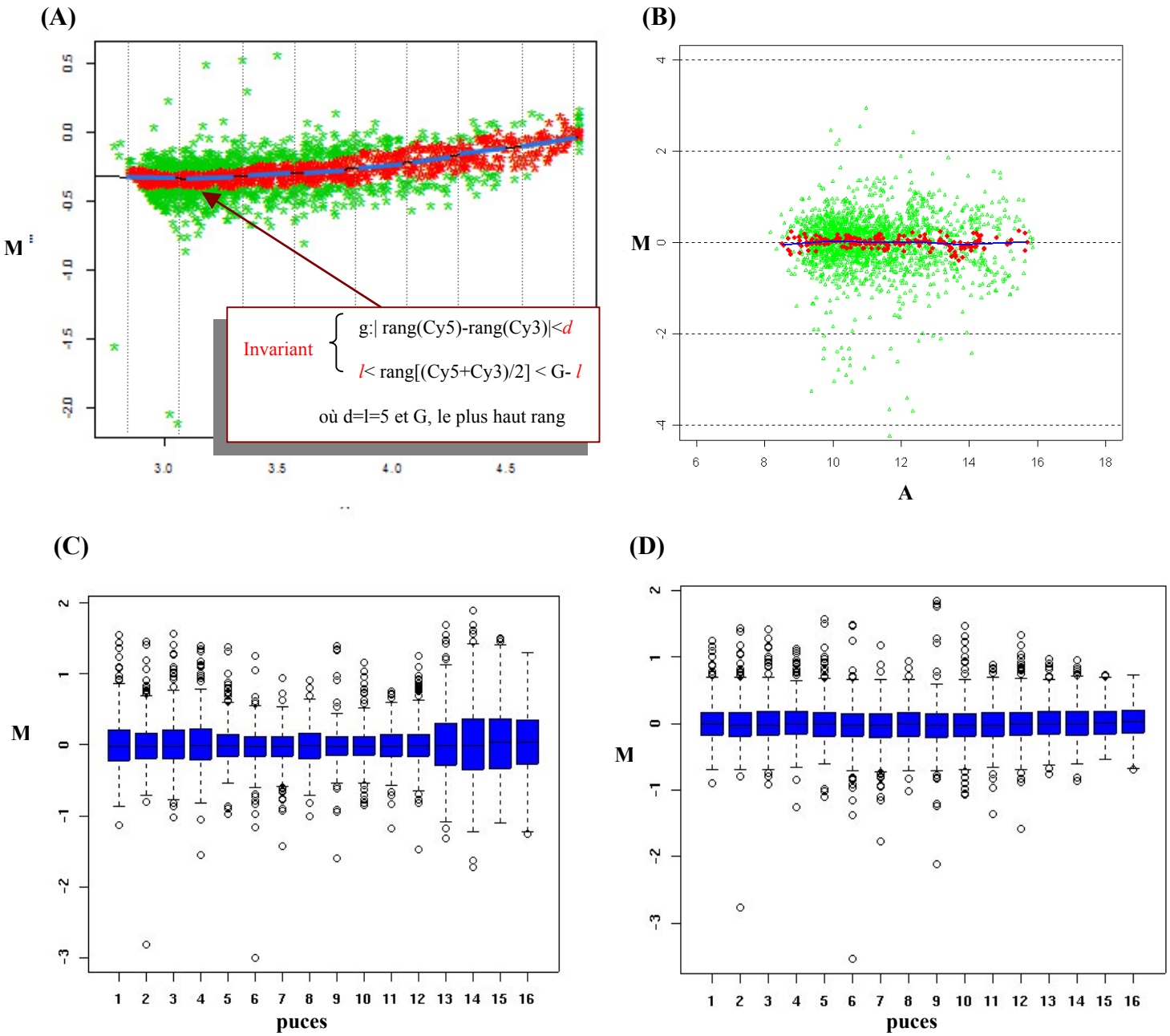


Figure 19. Normalisation *lowess fitness* et standardisation des données de puces à ADN avec MADSCAN. M et A représentent respectivement le ratio et la moyenne des intensités en log de base 2. (A-B) *MA plot* et description de la méthode de normalisation *lowess fitness* : (A) Estimation des *gene reporters* « invariants » (points rouges) selon l'algorithme de Tseng *et al.* (2002) (encadré). A partir du jeu de *gene reporters* « invariants », calcul des droites de régression locale (traits bleus) suivant un pas donné (lignes pointillées verticales). Lissage des droites de régression en une courbe de normalisation (courbe noire). (B) Normalisation de l'ensemble du nuage de points, les ratios M des intensités sont centrés sur 0. (C-D) Boîtes à moustache (*boxplot*) pour la description de la standardisation des puces à ADN aux même valeurs de dispersion. (C) La distribution des valeurs autour de la médiane (milieu de la boîte bleue) est différente d'une puce à l'autre. (D) Après *scaling,* les puces possèdent, approximativement, les même paramètres de dispersion (1er quartile, médiane, 3ème quartile).

La seconde étape du processus de traitement des données est la **normalisation intra-lame.** La normalisation des données vise à **minimiser (voire corriger) les biais techniques, systématiques ou liés au hasard**. Théoriquement, dans une expérience de puces à ADN, la majorité des *gene reporters* n'est pas différentiellement exprimée et la distribution des ratios est généralement centrée sur 0 (log2). Cependant, un déséquilibre des niveaux d'intensités entre les deux fluorochromes est souvent observé. Ils témoignent le plus souvent de biais techniques, tels que le biais systématique d'une incorporation hétérogène des fluorochromes ou l'usure inégales (au hasard) des aiguilles du robot de dépôt.

Les premières approches proposées pour normaliser les données primaires ont tout d'abord été l'utilisation de droites de régression linéaire ou exponentielle. Cependant, l'utilisation de droites de régression suppose que la population des données suive une loi normale et qu'elle soit linéaire sur la plage des intensités observées. Or, d'après la représentation M *versus* A, Dudoit *et al.*(2000) ont montrés que les spots de faible intensité présentent une variabilité plus importante que les spots de forte intensité (Fig. 19A). Les spots de faible intensité sont globalement plus sensibles aux biais expérimentaux. Suite à ces observations, des approches par **régressions non paramétriques** ont donc été proposées pour normaliser les données en fonction de leur niveau d'intensité (Tseng *et al.*, 2001; Yang *et al.*, 2001b; Workman *et al.*, 2002; Faller *et al.*, 2003). Parmi ces méthodes, l'algorithme de *lowess fitness* (***LO****cally* ***WE****ighted* ***S****catter plot* ***S****moothing*) (Cleveland, 1979) et ses variantes sont actuellement les plus employées. Leur principe est basé sur le découpage du nuage de points M *versus* A en fenêtres d'intensités de taille connue (généralement un pourcentage du nombre de données), suivi d'une somme de régressions locales pondérées (Fig. 19A). L'ensemble des régressions locales est ensuite lissé pour former la courbe d'ajustement des données. MADSCAN propose une version adaptée de l'algorithme de *lowess fitness* selon Yang *et al.* (2001b).

L'algorithme de *Lowess fitness* a été appliqué soit sur l'ensemble des *gene reporters* de la puce, soit sur des *gene reporters* sélectionnés *a priori* représentant des gènes dits de « ménages » (*housekeeping gene*) ou des gènes exogènes (Leung et Cavalieri, 2003). Cependant l'algorithme de *Lowess fitness* peut être sensible aux valeurs extrêmes de ratios et l'utilisation de l'ensemble des *gene reporters* de la puce pour normaliser n'est pas toujours efficace. De même, l'utilisation de jeux de *gene reporters* sélectionnés *a priori* c'est avéré inefficace car ces gènes effectivement stable dans certaines conditions peuvent variés dans d'autres. MADSCAN applique l'algorithme de *lowess fitness* non pas sur l'ensemble des *gene reporters* de la puce, ou un jeu de « gènes de ménage » mais à une sélection de **« gènes**

invariants » (Tseng *et al.*, 2001). Ces *gene reporters* sont sélectionnés *a posteriori* par une méthode non paramétrique de calcul des rangs des intensités. Les *gene reporters* sont dits « invariants » si et seulement si, d'une part la différence des rangs de leurs niveaux d'intensités (Cy3, Cy5) n'est pas significative (inférieure à un seuil *d*) et d'autre part que la moyenne de ces rangs soit comprise entre un seuil de signification (*l*) et le plus haut rang (Fig. 19A).

Enfin, Yang *et al.* (2002) ont également mis en évidence une corrélation entre la position des spots sur la lame (liée aux aiguilles du robot de dépôt) et leur niveau d'intensité. Une normalisation spatiale, par cadran (*print-tip group*), est alors nécessaire pour corriger cet « effet pointe ». MADSCAN propose trois modes de normalisation spatiale : « pointe par pointe », « proximale » et « globale ». Le choix de la méthode utilisée est fonction du pourcentage de « gènes invariants » estimés par cadran. Le minimum pour une analyse « pointe par pointe » est de 50% (pour plus de détails, *cf.* Annexe I).

Suite à la normalisation intra-lame, la distribution des ratios des intensités est centrée sur 0 (Fig. 19B). Toutefois, pour comparer des puces *replicates*, il est parfois nécessaire d'appliquer une **normalisation inter-lames** (***scaling***) ou standardisation afin de réduire l'écart de la variance des mesures entre les lames. Dans ce but, la distribution des ratios est standardisée, *i.e.* ramener aux même paramètres de dispersion (médiane, 1er et 3ème quartile…) (Fig. 19C-D). Comme pour la normalisation intra-lame, plusieurs techniques ont été développées (Yang *et al.*, 2002; Quackenbush, 2002). MADSCAN propose de réduire la distribution des mesures d'expression de chaque lame répétée à la même déviation absolue de la médiane ou MAD (***M**edian **A**bsolute **D**eviation*), cette mesure étant moins sensible aux valeurs atypiques que l'écart type.

La dernière correction est la recherche des **valeurs aberrantes** (*outliers*) parmi les mesures répétées d'un même *gene reporter*. Les valeurs aberrantes ont, par exemple, pour origine la faible spécificité des *gene reporters* et/ou un bruit de fond hétérogène à la surface des lames. Aussi, l'application de tests statistiques permet d'évaluer et valider le niveau de reproductibilité des mesures répétées. Toutefois, les expériences de puces à ADN disposent généralement d'un faible nombre de *replicates* ce qui limite l'application de la plupart des tests statistiques classiques. MADSCAN propose deux types de tests : un test *t* modifié par la médiane et le test de Grubb (Burke, 2001). Ces tests présentent l'avantage de pouvoir être appliqués à un petit nombre de répétitions (n=3) (pour plus de détails, *cf.* Annexe I).

MADSCAN permet également de tester la reproductibilité des mesures à l'intérieure des puces puis entre les puces répétées.

Finalement, une synthèse des niveaux d'expression par *gene reporter* à l'intérieur et entre les puces répétées offre une vision globale des résultats et de la qualité de l'expérience. MADSCAN propose l'intégration des données sous la forme d'une **matrice de données d'expression « consolidées »**. Pour chaque *gene reporter*, les moyennes des niveaux d'expression sont calculées par lames et entre les lames, associées à des indices de qualité des mesures, tels que le nombre de valeurs valides et le coefficient de variation.

1.2 Article

Ce travail a fait l'objet de la publication suivante :

Le Meur N, Lamirault G, Bihouee A, Steenman M, Bedrine-Ferran H, Teusan R, Ramstein G, Leger JJ. (2004). A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values: importance of replication. *Nucleic Acids Res*. **32**(18):5349-58.

Les données d'expression utilisées dans cet article sont disponibles sur le site Web de Gene Expression Omnibus[33] avec le numéro d'accession GSE1502.

[33] http://www.ncbi.nlm.nih.gov/geo/

Published online October 8, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 18 ***5349–5358***
doi:10.1093/nar/gkh870

A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values: importance of replication

Nolwenn Le Meur*, Guillaume Lamirault, Audrey Bihouée, Marja Steenman, Hélène Bédrine-Ferran[1], Raluca Teusan, Gérard Ramstein[2] and Jean J. Léger

Ouest genopole, Institut du Thorax, Institut National de la Santé et de la Recherche Médicale (UMR 533), Faculté de Médecine, 44035 Nantes, France, [1]Centre National à la Recherche Scientifique (UMR 6061), Rennes, France and [2]Laboratoire d'Informatique de Nantes Atlantique, 44322 Nantes, France

Received June 25, 2004; Revised August 17, 2004; Accepted September 17, 2004

ABSTRACT

We propose a freely accessible web-based pipeline, which processes raw microarray scan data to obtain experimentally consolidated gene expression values. The tool MADSCAN, which stands for MicroArray Data Suites of Computed ANalysis, makes a practical choice among the numerous methods available for filtering, normalizing and scaling of raw microarray expression data in a dynamic and automatic way. Different statistical methods have been adapted to extract reliable information from replicate gene spots as well as from replicate microarrays for each biological situation under study. A carefully constructed experimental design thus allows to detect outlying expression values and to identify statistically significant expression values, together with a list of quality controls with proposed threshold values. The integrated processing procedure described here, based on multiple measurements per gene, is decisive for reliably monitoring subtle gene expression changes typical for most biological events.

INTRODUCTION

During the last decade, cDNA microarray technology has been extensively applied to determine gene expression levels in diverse tissues, animals and diseases, at high throughput levels. As a result of the increasing knowledge of several genomes (especially the human genome), thousands of gene-fragments have been spotted or *in situ* synthesized to globally monitor various gene expression situations. Attention has been paid to mathematical (statistical) methods pertinent for the analysis, organization and handling of the enormous quantities of gene expression data [reviewed in (1)]. When comparing different situations like patients and controls, or when analyzing ontogenic or kinetical events, the challenge is to identify the combinatorial and hierarchical complexity of the gene expression profiles. Parallel to those efforts on interpretation of the data, other studies have aimed to identify the different physical and biological factors which have to be controlled to improve the reliability of massive gene expression studies (2–4). The multiple experimental steps involved in microarray procedures are sources of often badly controlled variation, which is superimposed on the biological variation under study. Experimental variation along the successive steps of preparation, purification and labeling of RNA samples, as well as the hybridization conditions, are inherent to all microarray experiments. Mechanical and optical distortions could locally or globally influence the raw expression values obtained after microarray image scanning. In addition, other factors like intrinsic heterogeneity, conditioning parameters and even erratic contamination of the biological samples may modify the gene expression results. To compensate (and/or better evaluate) the importance of these composite experimental and biological noises in microarray experiments, diverse numerical treatment procedures of the raw microarray scan image values and quality measures have been proposed. These procedures include filtering of bad spots following segmentation methods, normalizing between two channels (or signal scaling within monocolor microarrays), comparative scaling between different chips, and diverse statistical methods for selecting (ranking) differentially expressed genes (1,5). The most reliable way of evaluating the ratios between the different experimental noises and the biological signals is to produce replicate gene measurements within each microarray and to hybridize replicate microarrays with replicate targets obtained from the same biological samples. The metrological importance of such replications in microarray gene expression studies (6–9) casts doubts on the numerous microarray analyses performed with only singular gene spots and/or without sample replicates. The high cost associated with microarrays does not justify the metrological insufficiencies of any experiment. The accessibility to high throughput spotting robots deposing up to 25 000 spots per chip combined with a careful selection of a few thousands of theme-relevant genes now allows the use of such noise-informative chips and the design of corresponding reliable experiments.

*To whom correspondence should be addressed at INSERM U533, Faculté Médecine, BP 53508, 44035 Nantes cedex 1, France. Tel: +33 240412957; Fax: +33 240412950; Email: nolwenn.lemeur@nantes.inserm.fr

Here, we present a freely accessible web-based pipeline which processes raw microarray scan data to obtain experimentally consolidated gene expression values. The proposed module, MADSCAN, MicroArray Data Suites of Computed ANalysis (http://www.madtools.org), makes a practical choice among the numerous methods available for filtering, normalizing and scaling microarray data, in a dynamic and automatic way. Using a careful experimental design with replication information, we present different numerical and statistical methods, detection of outlying expression values and data integration with a list of quality controls, through proposed threshold values. A tutorial for MADSCAN is included on the website.

MATERIALS AND METHODS

Biological samples

Cardiac tissue was obtained from the left ventricle of explanted hearts from two male patients who underwent heart transplantation. One patient (experiment 1) was affected by idiopathic dilated cardiomyopathy; the other patient (experiment 2) was affected by valvular heart disease and coronary artery disease. Both samples were compared to a common reference that was obtained from pooled RNAs extracted from the left and right ventricles of explanted hearts from 47 end-stage heart failure patients (G. Lamirault, N. Le Meur, M.F. Le Cunff, C. Chevalier, I. Guisle, A. Bihouée, J.J. Léger and M. Steenman, manuscript in preparation).

RNA isolation and labeling

Total RNA was isolated using TRIZOL® Reagent (Life Technologies). Two parallel RNA extractions from two different samples (spatially separated) of the same tissue were performed. Poly A^+ RNA was isolated using the Oligotex mRNA kit (Qiagen) and quality was assessed using an Agilent 2100 bioanalyzer. Cy3- and Cy5-labeled cDNA was prepared using the CyScribe cDNA Post Labelling Kit (Amersham Pharmacia Biotech). Samples from the two patients under study were each labeled individually with Cy3. The reference pool was labeled with Cy5. No dye-swap was used.

Microarrays

Microarrays were prepared in-house using 50mer oligonucleotide probes (MWG Biotech). The probes were arrayed onto epoxy-silane-coated glass slides using the Lucidea printer from Amersham. The 4217 genes represented on the microarray were selected for involvement in skeletal muscle and/or cardiovascular normal and pathological functioning (10–13). Gene selection was based on (i) subtractive hybridization experiments, (ii) genome-wide microarray hybridizations, (iii) literature data. Each Cy3-labeled sample was mixed with equal amount of Cy5-labeled sample, pre-incubated with human Cot-I DNA (Gibco-BRL), yeast tRNA and poly A^+ RNA, and hybridized to the microarrays. Two independent hybridizations were performed for each RNA sample, leading to four hybridizations per patient. Hybridized arrays were scanned by fluorescence confocal microscopy on a ScanArray 4000XL (GSI Lumonics, Downers Grove, IL) at laser power ranging from 75 to 100% and photo-multiplier tube gain settings ranging from 65 to 100%. Measurements were obtained separately for each fluorochrome at 10 μm/pixel resolution.

Microarray data acquisition

Signal intensities were extracted with the GenePix Pro 5.0 image analysis software (Axon Instruments, USA). Segmentation of the spots was done using the adaptive approach. Segmentation criteria were optimized visually for each slide. Alternate standard deviation (SD2) was chosen to quantify background SD. This setting uses the median of the background pixels as an estimator of the center of the distribution. This method is less susceptible to skewing by very bright pixels. Our data processing procedure uses background corrected median intensities; the given ratio corresponds to the ratio of background corrected median intensities. For further quality controls (see the preprocessing step in Results and Discussion, and in the tutorial), analytic parameters provided by the GenePix Pro 5.0 image analysis software were used. Other image analysis software like Quantarray (PE. Packard Biochip technologies, USA), Imagene (BioDiscovery, USA) or ScanAlyze (http://rana.lbl.gov/, Stanford University, USA) also deliver the minimal set of parameters required to perform the MADSCAN procedure. For a comparison between different image analysis software, see http://cardioserve.nantes.inserm.fr/ptf-puce/publications.php. Analysis files issued from different types of image software can be reformatted following procedures noted in the MADSCAN tutorial.

Power study

A power study of a standard *t*-test was performed on the heart expression data set of experiment 1 (four replicate spots for each oligonucleotide and four replicate chips). Only genes with at least two valid *M*-values [$=\log_2$(signal ratio)] for each array were selected for power analysis. We thus selected 3804 genes with reliable *M*-values. The 'power *t*-test', which is implemented in the '*c*test' package of *R* (14) was applied in five replication conditions: 4, 6, 8, 12 or 16 replicate *M*-values. Parameters of the power test were defined as follows:

- For each gene the mean level of differential expression between the two fluorochromes (Δ) was defined as the arithmetic mean of the four arithmetic means of the 4 *M*-values in each of the quadruplicate chips.
- The parameter for data variability (SD) was arbitrarily set as identical for all genes. SD was calculated as the median of the 3804 SDs determined from the replicate *M*-values for each gene.
- Significance level (α) was a priori set to 0.05, but a Holm correction (15) was applied to α for each gene in order to account for multiple testing hypothesis. The 3804 genes under analysis are ranked according to their individual *P*-values, by application of a standard one-sample *t*-test. On the basis of the calculated rank of the corresponding gene, the basal α-value of 0.05 was then corrected individually.

Using the values fixed for SD and α for each gene, the individual power test was performed on the basis of one sample and a two-sided *t*-test. Power values of $(1 - \beta)$ were deduced for each gene in the five replication conditions analyzed. To evaluate the effect of between-gene differences in sampling

variation on the power test values, the power test was calculated with three different values of SD in two particular replication conditions (4 and 12 replicates). The three different SD values were defined as first quartile, median (as earlier) and third quartile of the 3804 SDs previously calculated. Other parameters of the power test were left unchanged. Power values ($1 - \beta$) were calculated for both replication conditions and the three variability levels.

Estimation of false positive and false negative rates

The false positive (FP) rate is the proportion of negative cases that were incorrectly classified as positive in the predicted condition compared to the experimental observation. The false negative (FN) rate is the proportion of positive cases that were incorrectly classified as negative in the predicted condition compared to the experimental observation. Genes differentially expressed between the heart expression data sets of experiments 1 and 2 were first identified by a SAM modified two-class *t*-test (16), using 16 (4 within- and 4 between-chip) replicates for each data set. The number of differentially expressed genes was then determined based on six different replication conditions: 4, 8 or 12 replicates with various proportions of within- and between-chip replicates (see Supplementary Material). Six different two-by-two confusion matrices (17) were built to determine the FP and FN rates in the six simulated replication conditions with regard to the experimental situation based on 16 replicates.

MADSCAN implementation

MADSCAN was written in R (14) and Perl. A user-friendly web-interface was implemented in PhP to allow easy access (http://www.madtools.org) and rapid handling of data on our local server (PowerEdge 4600, Dell, USA). Access is obtained through a password, given to any requester. The raw microarray data are uploaded as compressed tabulated text files. MADSCAN analysis can be done either step by step or from A to Z, i.e. one can either apply one test at a time or all steps in a single and complete procedure. The results can be downloaded from the web-interface, where a window of results displays a summary of the performed procedure. Alternatively, MADSCAN results can be recovered through an e-mail service.

RESULTS AND DISCUSSION

Outline of analysis procedure

Our goal was to provide a practical, accessible, integrated suite of different analytic procedures for the handling of raw data issued from two-fluorochrome (color) image scanning of microarray glass slides and to obtain consolidated expression values. According to the MIAME (Minimum Information About a Microarray Experiment) glossary, data processing means 'the set of steps taken to process the data, including the normalization strategy and the algorithm used to allow comparison of all data' (18). Draghiči (5) defined preprocessing as the initial step that extracts and enhances meaningful data characteristics from raw data files from scanned images. Preprocessing prepares the data for the application of successive procedures or analytical methods. Using tabulated text files of raw microarray image values issued from widely used scanners and related image analysis software, we have developed a four-step procedure to transform the raw data into consolidated, robust expression values: the first three steps concerned each individual chip, whereas step 4 integrated the expression values issued from replicate chips if available (Figure 1).

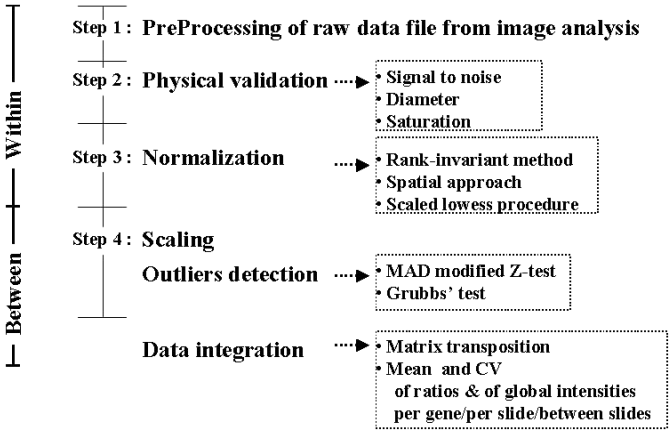


Figure 1. MADSCAN procedure steps within each chip and between replicate chips.

The proposed integrated tool has been constructed around a few now well-accepted main analytic steps to numerically handle microarray image values within one chip and between replicate chips. Most of the used algorithms or methods—such as the rank invariant method (19), lowess fitness normalization (20,21) and outlier detection (although never used for microarrays) (22,23)—have been defined and documented by others. The corresponding methods/algorithms have been reformatted in a plug-in architecture system to make the whole numerical procedure reliable and fluent. Algorithmic approaches chosen for each step and modifications or adaptive processes made along the procedure are described in the following subsections. The computational tool, MADSCAN, is freely accessible via a web server (http://www.madtools.org), where detailed information is available in a tutorial.

The importance of the experimental design

Before describing the different steps of the MADSCAN procedure, we addressed the important issue of the reproducibility of microarray experiments. We proposed a 'reference design' with an experimental procedure based on replicate spots within each microarray, and replicated microarrays for two spatially separated samples from each tissue, compared in a hybridization to a reference sample (Figure 2). The replicate spots are issued from different print-tips and are therefore printed in different array blocks. This procedure allows the evaluation of the importance of the biological noise—due to sample heterogeneity—and numerous experimental noises. The latter could arise from variations in the molecular biology procedures for the extraction and labeling of RNA samples (e.g. dye quality, or possibly dye-swapping), from physical distortions in glass slides and from the scanner (optical irregularities in the laser performances and in the excitation of the fluorochromes). To be able to take into account such

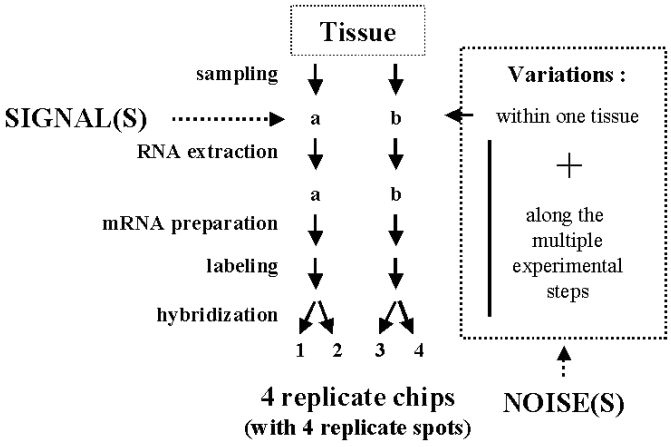


Figure 2. Experimental design. Two independent RNA samples (a and b) from the same tissue, replicated spots within one chip and replicated chips for one biological point are necessary to discriminate between the signal under study and those due to the inherent experimental noises.

composite noises from one chip, it is obvious that microarrays need to contain at least triplicate spots. This allows the statistical evaluation of the internal variability of the signals corresponding to each gene (oligonucleotides or PCR products) and the detection of outlying values within one chip. Furthermore, a minimum of four (two pairs) replicate chips is necessary to evaluate the variations between the two independent RNA samples issued from the same tissue (6,8). The within-chip replicates reveal only technical noises, whereas the between-chip replicates give information on both technical and biological noises. Microarray experiments as designed in Figure 2 actually allow simultaneous measurements of the different technical noises, together with that of the biological signal under investigation. In addition, randomized print-tip usage allows a non-uniform distribution of the replicate spots throughout the array. Together with a randomization of the numerous experimental procedures and the use of replicates, this is crucial to obtain statistically significant data.

Preprocessing of raw data files from image scan analysis

Whatever the type of scanner or related software used, MADSCAN starts with tabulated text files composed of at least eight columns per hybridized microarray. These columns contain information on block designation, gene name, gene ID or annotation, measured intensities in both channels, local (or equivalent) background intensity values for both channels and image analysis software flags for a first determination of spot quality (diameter deviation of the spot, location). Nomenclature and gene annotation have to be carefully formatted. Replicate spots (if present) must be precisely annotated to be identified as such during the data processing procedure.

Physical validation or quality filtration

The overall quality of the raw image data (before any filtering) is calculated for each print-tip group (block) of spots according to the median values and the variation coefficients of signal and background intensities. Spot diameters and their SD are also determined. In spite of the importance of assessing spot quality, relatively few image analysis software packages provide systematic quality filtration (5). MADSCAN offers physical validation and quality filtration step by step, following a decision tree with a scoring procedure based on successive exclusion thresholds. Each feature is thus tested against a series of quality criteria (image analysis flags, signal-to-noise level, diameter variation and saturation level). Five different arbitrary scores can be attributed according to the spot quality. Score 0 is used for flawed spots whereas most of the good spots have score 2. Scores 3 and 4 are attributed to spots that are partially saturated for one of the channels. For those spots, the expression ratio is calculated from the regression ratios between the intensities of each pixel composing the spot. Score 5 is attributed to features partially saturated in both channels and their expression ratio is calculated as the ratio of their percentages of saturation. Fully saturated spots in both channels are flagged because no reliable information on the pixel values and distribution is available (see tutorial for details). Chips made in-house contain 15 000 to 20 000 spots. Using our conditions for hybridization and image scanning, 5–8% of the spots are flagged (=score 0) whereas 92–95% of the spots pass the quality control criteria. The percentage of partially saturated spots (score 3–5) is generally relatively low (0.05–0.1% of the spots).

Within-chip normalization step

Normalization issues have been addressed early in the development of microarray data treatment (19–21,24). It is considered an essential step to minimize experimental systematic and random biases, arising from technical variations inherent to the high throughput and complex experiments. The main aspects of any normalization process are whether or not to select a set of reporter (invariant) genes as a reference for the normalization process and whether or not to consider spatial and intensity value-dependent biases. Since most microarrays contain several thousands of spots, and since hybridization values are mostly distributed in an equilibrated (pseudo-gaussian) way in experiments comparing test and control tissues, we chose to adapt the rank invariant method developed by Tseng *et al*. (19) in our procedure. A set of invariant spots or non-differentially expressed genes (if no replicates were spotted) is a posteriori selected from all validated expression ratios for each chip. The rank of Cy3 and Cy5 intensities of each gene on the chip is computed separately. If the ranks of the two intensities for a given gene differ less than a fixed threshold and the rank of their averaged intensity is not among the highest or lowest ranks, this gene is classified as a non-differentially expressed gene. Figure 3A shows an *M–A* plot [$M = \log_2$(signal intensity ratio) and $A = \log_2$(averaged signal intensities) (20)], with a selection of such invariant spots, following the application of the rank invariant method. The invariant spots in Figure 3A are sandwiched between the differentially expressed genes. As has already been described (20), the distribution of expression ratios is intensity dependent and therefore a non-linear normalization method must be used. The *lowess fitness* method, using the set of identified invariant genes, has been incorporated in our MADSCAN procedure. To assess the efficiency (robustness) of coupling

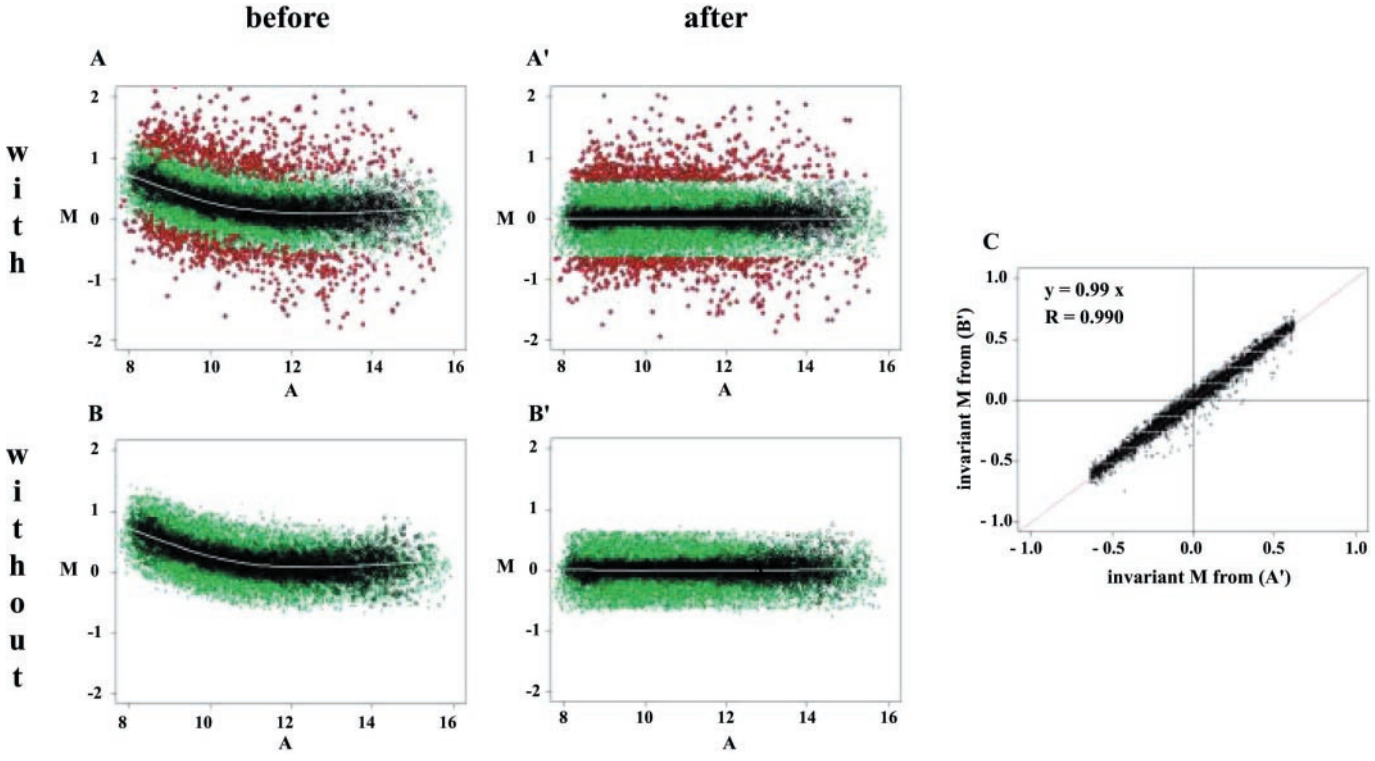


Figure 3. *M–A* plots before (**A** and **B**) and after (**A′** and **B′**) global *lowess* normalization, using rank invariant spots. The spots that are potentially differential in graphs (A) and (A′), [M >/< Median(M) ± 2.5 * MAD(M)], were eliminated for the determination of invariant spots used for further data normalization in graphs (B) and (B′). 'With' and 'without' refers to the presence or absence, respectively, of potentially differentially expressed genes. The presented expression values were from experiment 1. (**C**) Represents the correlation between the 85% invariant genes, common to the gene populations in graphs A′ and B′.

both rank invariant and *lowess fitness* methods, we removed all identified putative differentially expressed spots (genes) from the original raw expression file [M >/< Median(M) ± 2.5 * MAD(M) (median absolute deviation of M-values)]. We then applied both complementary methods for normalization on the reduced raw expression files. Figures 3B and 3B′ show the new set of invariant genes and the raw and normalized expression data, respectively. Eighty-five percent of the invariant genes selected before and after the file reduction are identical. A very strong correlation coefficient of 0.99 is observed between both sets of independently normalized expression values (Figure 3C). This is obviously due to the high number of invariant spots (genes) present (∼50% of the total amount of spots).

As described by Yang *et al*. (21), the use of a spatial approach is also crucial. The signal as well as the background intensity is often heterogeneous within a slide. This is due to the unavoidable spot dispersion over a relatively large surface, the use of several spotting pins and possible geometrical variations within glass slides. An additional refinement of the normalization procedure thus has to be applied to chips containing more than a few thousand spots. A normalization procedure per zone, usually print-tip group, allows to correct spatially dependent dye biases and probe delivery variations between the different pins as well as other geometrical and optical defects. Practically, in MADSCAN, the normalization is first attempted pin-by-pin (print-tip group), then by proximal or global approach, depending on the number of invariant spots present within individual blocks, contiguous blocks or the whole chip, respectively. We found that at least 50% of invariant genes among all genes under analysis are needed to obtain a robust normalization curve. To illustrate the used procedure and the importance of spatial normalization, *lowess* normalization procedures were applied based on invariant spots selected from either individual blocks (individual print-tip) or proximal blocks or all blocks in a 48-block chip with 420 spots per block. Five individual lowess fitness curves arbitrarily chosen among the 48 different ones obtained in each spatial condition are graphically represented in Figure 4. It is easily seen that the best superposition of the five curves is obtained when the rank invariant method was applied pin-by-pin rather than using proximal blocks or all blocks.

Scaling and outlier detection

In a metrologically controlled experiment—as described in Figure 2—the presence of replicated features within each slide and of replicated slides for each biological sample allows a statistical validation of the expression results after the three first steps of the procedure (Figure 1). First, scaling procedures have to be applied to bring the variances of filtered and normalized expression values between replicated chips at the same variation level (5,20). Outlying values within the series of expression values obtained for each gene from several spots can then be identified. Because of the low number of replicates in microarray experiments, we propose to apply modified statistical tests. A z-score modified by MAD is used to detect outliers within and between slides. In the MADSCAN

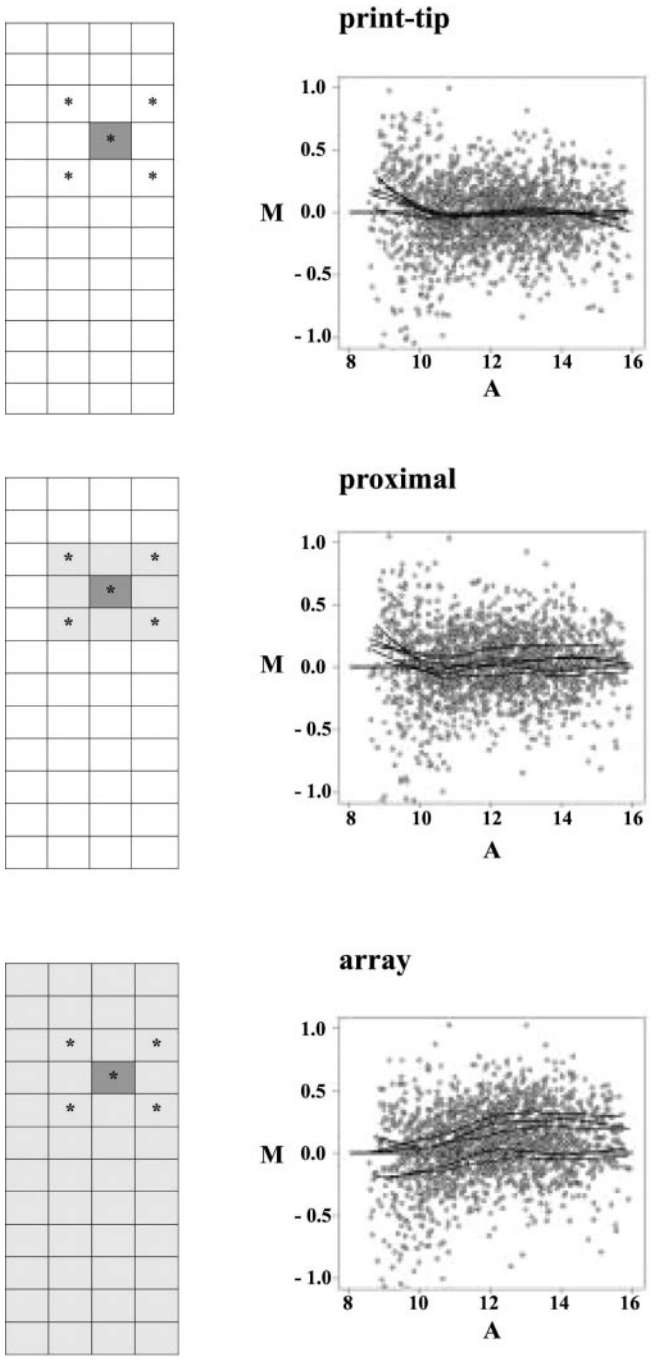


Figure 4. Comparisons between global, proximal and local normalization procedures. Five individual lowess fitness curves corresponding to five arbitrarily chosen blocks (asterisks) are represented according to each of the three spatial normalization modes. Light gray blocks represent an example of blocks chosen for selection of invariant genes, to normalize the raw *M*-values in the dark gray block, in each of the three modes. Invariant genes within the dark gray block are part of the invariant population used in each mode. The superposition of the five selected curves shows how uncontrolled local variations may influence the final expression values. The expression values presented here were from experiment 2.

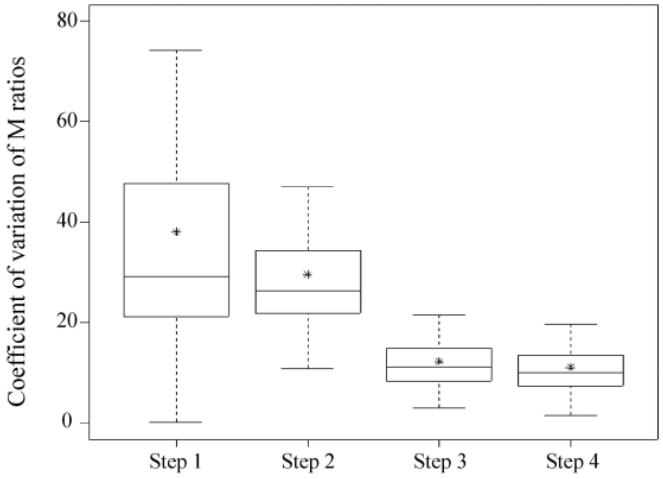


Figure 5. Decrease of the coefficients of variation of expression ratios, along the different MADSCAN analysis procedure steps. The expression values were from experiment 2, using 2 × 2 replicate chips with four replicate spots for each gene. The 'box' in a 'box and whisker' plot shows the median of the values as a line, the mean as an asterisk and the first (25th percentile) and third quartile (75th percentile) of the expression values distribution as the lower and upper parts of the box, respectively. The 'whiskers' shown above and below the boxes represent the largest and smallest observed values, respectively, that are less than 1.5 box lengths (interquartile range) from the end of the box. When the box is in the middle of the whiskers, the data are probably more evenly distributed (steps 3 and 4). Steps 1 to 4 are as in Figure 1.

procedure, we have implemented both the MAD and the ESD (Extreme Studentized Deviate or Grubb's test) procedures (22,23). The procedure for detecting outliers requests a minimum of three replicated values. The replicates may be within chips and/or between chips. Outlier detection can be applied iteratively with both tests, until no more outlier is detected.

Crucial steps in microarray data treatment

The presence of replicate spots for each gene on each individual chip and on replicate chips allows the calculation of the within-chip as well as the between-chip coefficients of variation (CV) of the expression ratios (=2^M), respectively, at each of the four steps described in Figure 1. Figure 5 shows the variations of the CV calculated from the medians of expression ratios for each gene, in a typical experiment involving 2 × 2 replicate chips with four replicate spots for each gene (Figure 2). Step 3, corresponding to the within-chip normalization procedure, is clearly the most decisive step for reducing the absolute value and the variation of the CV. The CV distributions around their median values are approximatively gaussian, even though they are obviously higher for low intensity values (25). First and third quartile values in each of the four CV distributions are central visual elements for evaluating and controlling the quality of the expression values obtained for each individual chip and for replicate chips in the MADSCAN procedures. In contrast, step 4 does not significantly alter the CV values and their relative variations. This has to be related to the very small number of outlying values usually detected for each gene. However, this does not mean that outlier detection and elimination do not play a role in the CV calculations.

Spot replicates and the detection of subtle expression changes

The robustness of the proposed 'reference design' with within- and between-chip replicates is illustrated by means of (i) a

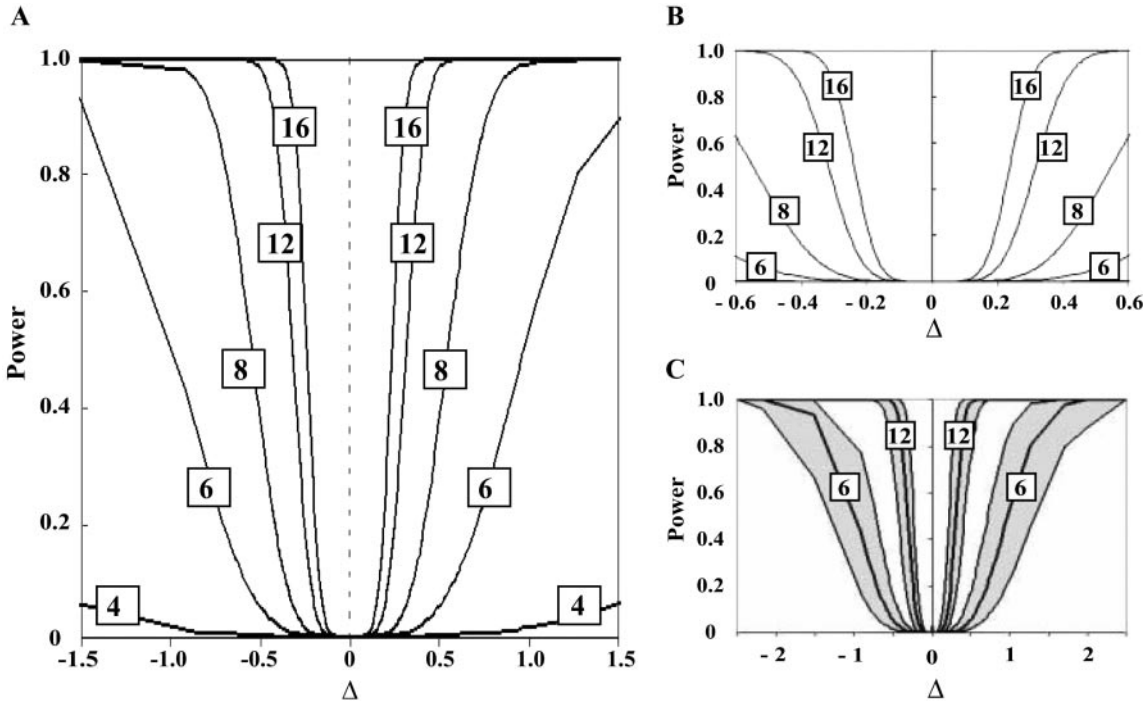


Figure 6. Validation experiment and power analysis, using replicate spots and replicate chips. The set of expression values for power calculations were from experiment 2. (**A**) Power values (1 − β) calculated in five replication conditions (4, 6, 8, 12 or 16 replicate *M*-values) were plotted against Δ, the mean level of expression values between the two fluorochromes, which is calculated as the arithmetic mean of the four arithmetic means of the 4 *M*-values in each of the quadruplicate chips, for each gene. (**B**) The same results as in (A) for four replication conditions, but zoomed to a smaller *x*-axis (Δ values ranging from −0.6 to 0.6), to underline the capacity of 12 and 16 replicates to detect small gene expression changes. (**C**) The gray zones around the power values were defined from the power values calculated from the first and third quartiles of all the SD values of the *M*-values in the 6 and 12 replication conditions. The same results as in (A), but zoomed out to a larger *x*-axis (Δ values ranging from −2.5 to 2.5).

power analysis (8) performed using 4, 6, 8, 12 or 16 replicate *M*-values and (ii) an estimation of FP and FN rates at different replication conditions (4, 8 or 12 replicates) as compared to the 16 replications experimentally used.

Power study. Power values (1 − β) in each replication condition are plotted against the mean level of differential expression (Δ), which is defined as the arithmetic mean of the four arithmetic means of the 4 *M*-values in each of the quadruplicate chips, for each of the 3804 analyzed genes (Figure 6). Δ represents the most probable (informative) value for the expression ratio for each gene, since it results from the maximum number (in this analysis: 16) of experimental determinations (see Materials and Methods for the definition of the parameters used in the power *t*-test). Figures 6A and 6B show that two-digit replicates (in this analysis: 12 and 16 replicates) allow to detect stable changes in the expression ratios as low as 15% (roughly a variation of 0.20 in *M*) with a probability value lower than 0.05. The methodological sensitivity to detect limited variations in gene expression dramatically decreases when the expression ratios are determined on <6 replicate values. The grayed area between the corresponding power values calculated for the first and the third quartiles (Figure 6C) represent the variations of the SD values of the expression ratios for each gene, deduced from 6 and 12 replicate spots. Only genes with relatively high differential expression levels ($M > \pm 1.5$ at least) show sufficient reproducibility when only six replicates have been used. The present observations on the capacity of replicates to detect limited gene expression changes using DNA chips are in concordance with other studies (6–9). Replicate gene spots, as well as replicate chips, are crucial for reliable monitoring of subtle gene expression changes typical for most biological events. Only large expression changes can be obtained reproducibly from microarray studies performed with chips containing no, or a very limited number of within- and/or between-chip, replicates.

False positive and false negative rates. The gain from replications can also be calculated from paired sets of FP and FN rates, determined from differentially expressed gene collections with variable (<16) numbers of mixed within- and between-chip replications. Significantly low FP rates were obtained only with repeated hybridizations (chips) (Figure 7). In parallel, the number of within-chip replicates decreased the FN values. The concomitant use of additional within- and between-chip replicates allowed obtaining balanced values of both FP and FN rates. The simulation of 4 slides with 3 replicates per chip generated a tolerable FP rate of 7% and an FN rate of 14%. This replication pattern, which allowed the evaluation of both technical and biological variations, seems feasible with regards to labor intensity and cost of chips. Both FP and FN rate analysis and power analysis led to coherent conclusions on the importance of replications. This clearly defines the limitations in the use of genome-wide microarrays, which contain many genes but almost no replicates. Any additional experimental variability inherent to other chip

designs—particularly to all 'even designs' including the use of dye-swaps (26)—could be evaluated in the same way.

Data integration and MADSCAN use

MADSCAN offers multiple processing steps such as filtration, normalization and outlier detection. Raw scan data can be fully analyzed chip by chip or in chip batches. All procedures can be applied independently, i.e. step by step or they can be run in a single and complete procedure, according to the experimental design. Only filtration and normalization procedures are applied to single chips without replicate spots or experiments without replicate chips. All expression data for one experiment are resumed in a consolidated matrix, which allows further comparisons with other data sets in a complete experiment. MADSCAN creates an end file that contains for each gene; its name, the within-slide median of expression ratios in $\log_2(M)$, the within-slide coefficient of variation (CV-M), the between-slide median expression ratio and its CV, and the same data for the geometric mean of the intensities. Intermediate files are accessible for any step. Figure 8A shows the start menu of web-accessible MADSCAN. Figure 8B displays a typical summary of data processing for an example of four replicate chips. A few empirically defined inclusion threshold values for the quality of the chip(s), some statistical parameters on expression ratios and the spatial mode used for normalization are also shown. Detailed information on the quality control parameters is available online (http://www.madtools.org). MADSCAN and its 'A to Z' approach were principally developed to handle replicate chip experiments with a 'reference design'. Other experimental designs, such as 'time series', 'dye-swap' or 'loop'designs, can be analyzed in batches until after the normalization step. The final steps of the current online MADSCAN version, i.e. outlier detection and data consolidation, can easily be performed step by step by reformatting the normalized data files (for details, see tutorial).

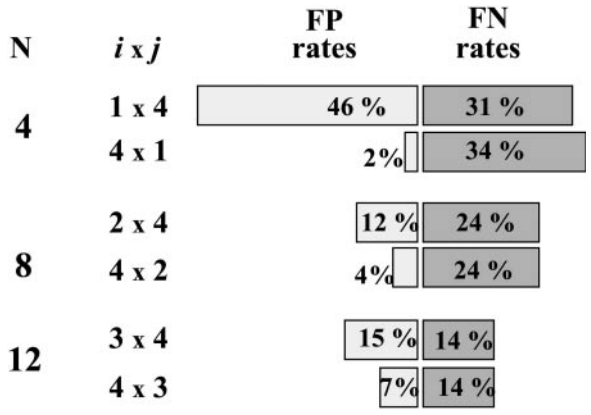
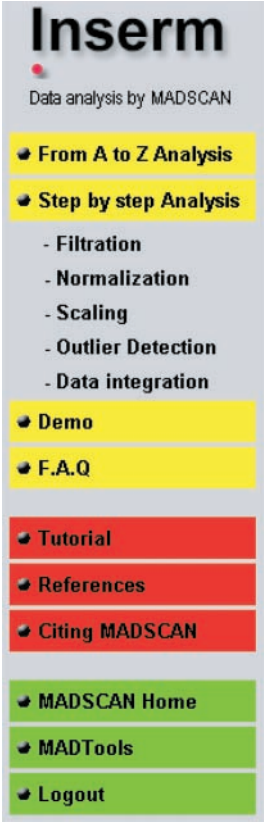


Figure 7. Gain from replications. FP and FN rates determined in six simulated replication conditions with regards to the experimental situation based on 16 replicates (*N* replicates obtained by *i* chips with *j* repeated spots).

A



B

	Slide Imc00035	Slide Imc00137	Slide Imc00288	Slide Imc00378	Threshold
Features:					
Initial number of features	16868	16868	16868	16868	-
Validated	92 %	92 %	95 %	95 %	> 65
Saturated	0.1 %	0 %	0.1 %	0 %	< 3
Flagged	8 %	8 %	5 %	5 %	< 35
Statistically validated features	14357	14636	15255	15228	-
Ratio:					
Mean	0	0.01	0.02	0	< 0.06
Median	0	0.01	0.02	0	< 0.06
Standard Deviation	0.31	0.27	0.28	0.32	< 0.50
Median Absolute Deviation	0.16	0.16	0.16	0.16	< 0.30
Normalization mode	Proximal	Proximal	Proximal	Proximal	Pin

Figure 8. Web-accessible starting menu (**A**) and data summary page (**B**) in the MADSCAN module. For details, see pp. 22 and 27 in the tutorial. Note the definition of threshold values concerning the quality of the chip(s) and related expression measurements.

CONCLUSION

The challenge of determining thousands of values of gene expression levels in a parallel but unique way using DNA microarrays forces the biologist of today to reliably manage and analyze a deluge of biological data. During the last few years, many alternative algorithms, based on relatively sophisticated and diverse mathematical methods, have been proposed and validated to successfully transform the image scan raw data into consolidated gene expression data. Based on a careful and pragmatic selection among the numerous methods and software available for filtering, normalizing and scaling the raw microarray data, the web-accessible MADSCAN resource presented here offers a dynamic and automatic procedure to obtain a set of reliable gene expression values. The incorporation of methods for within- and between-chip scaling and outlier detection, together with the online access to quality control parameters, complements the proposed plug-in architecture resource in an original way. A careful experimental design—including multiple measurements for each gene under each biological condition—is clearly central to the evaluation of most experimental noises inherently present in high throughput measurements. The significance and quality of any further biological interpretation—gene clustering, coexpression, etc.—are directly dependent on the reliability and significance of the set of consolidated gene expression values derived from image scan values. Obtaining such an initial set of metrologically relevant chip data is the exclusive scope of the MADSCAN procedure.

More or less sophisticated computational tools with various methods for microarray data processing are offered today in many commercially available and/or academic web-accessible software (for a list, see http://ihome.cuhk.edu.hk/~b400559/arraysoft.html or http://genopole.toulouse.inra.fr/bioinfo/microarray/index.php?page=logiciels). Among the available software, the steps corresponding to the initial treatment of raw scan data are either limited to some basic and inadequate transformation algorithms (like a linear normalization based on a few house-keeping genes), or numerous sophisticated, interconnected or independent, algorithmic modules are proposed. In all cases, the biologist has to adjust a series of 'default' parameters, more or less adapted to their own experimental design and the variables measured (27). Some knowledge and even understanding of the details of the algorithms/languages used are necessary to fully appreciate how such changes in the parameters do affect the expression results. To avoid those types of difficulties, we propose MADSCAN. MADSCAN, which has been successfully tested by diverse users on >2000 chips, containing 500 to 24 000 spots, represents an intelligent and powerful tool for the many biologists using DNA chips (12–13,28). The MADSCAN procedure is now plugged into BASE (BioArray Software Environment) (29). Therefore, information on raw image data and their transformation into consolidated expression values is accessible to the entire scientific community, in agreement with the most recent recommendations of the MGED consortium (18).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Catherine Chevalier, Isabelle Guisle and Martine Le Cunff for assistance with facilities and experimentation. We are grateful to Geoffroy Golfier, Jean Mosser and Marie-Claude Potier for helpful discussions on methods. We are also grateful to the company Perkin Elmer for their support in the early phase of our study and exchanges concerning the interpretation of microarray data. This work was supported by the 'Institut National de la Santé et de la Recherche Médicale', 'le Centre National à la Recherche Scientifique', 'l'Association Française contre les Myopathies', 'le Conseil Régional des Pays de la Loire' and 'Ouest Génopole'.

REFERENCES

1. Chipping forecast (2002) *Nature Genet.*, **32** (Suppl.), 461–552.
2. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
3. Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H. and Herzel,H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.
4. Yang,Y.H., Buckley,M.J., Dudoit,S. and Speed,T. (2000) Comparison of methods for image analysis on cDNA microarray data. Technical Report 107, Department of Statistics, University of California, Berkley, CA.
5. Draghiči,S. (2003). *Data Analysis Tools for DNA Microarrays. 1st edn.* Chapman & Hall, Boca Raton, FL.
6. Lee,M.L., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.
7. Hwang,D., Schmitt,W.A. and Stephanopoulos,G. (2002) Determination of minimal sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–1193.
8. Pan,W., Lin,J. and Le,C.T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.*, **3**, research 0022.
9. Pavlidis,P., Li,Q. and NobleW.S. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.
10. Tkatchenko,A.V., Le Cam,G., Leger,J.J. and Dechesne,C.A. (2000) Large-scale analysis of differential gene expression in the hindlimb muscles and diaphragm of mdx mouse. *Biochim. Biophys. Acta*, **1500**, 17–30.
11. Cros,N., Tkatchenko,A.V., Pisani,D.F., Leclerc,L., Leger,J.J., Marini,J.F. and Dechesne,C.A. (2001) Analysis of altered gene expression in rat soleus muscle atrophied by disuse. *J. Cell. Biochem.*, **83**, 508–51.
12. Steenman,M., Chen,Y.W., Le Cunff,M., Lamirault,G., Varro,A., Hoffman,E. and Leger,J.J. (2003) Transcriptomal analysis of failing and nonfailing human hearts. *Physiol. Genomics*, **12**, 97–112.
13. Steenman,M., Lamirault,G., Le Meur,N., Le Cunff,M., Escande,D. and Léger,J.J. (2004) Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays. *Eur. J. Heart Fail.*, in press, doi:10.1016/j.ejheart.2004.05.008.
14. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
15. Holm,S. (1979) Simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
16. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5120.
17. Kohavi,R. and Provost,F. (1988) Glossary of terms. *Machine Learning*, **30**, 271–274.
18. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C., Gaasterland,T., Glenisson,P., Holstege,F.C., Kim,I.F., Markowitz,V., Matese,J.C., Parkinson,H., Robinson,A., Sarkans,U., Schulze-Kremer,S., Stewart,J., Taylor,R., Vilo,J. and Vingron,M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genet.*, **29**, 365–371.

19. Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*., **29**, 2549–2557.
20. Yang,Y.H., Dudoit,S., Luu,P. and Speed,T. (2001) Normalization for cDNA microarray data. *Brief. Bioinformatics*, **2**, 341–349.
21. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.
22. Burke,S. (2001) Missing values, outliers, robust statistics & non-parametric methods. Statistics and data analysis. *Statistics and Data Analysis. LC.GC. Europe Online Supplement*, **59**, 19–24.
23. Müller,J.W. (2000) Possible advantages of a robust evaluation of comparisons. *J. Res. Natl Inst. Stand. Technol*., **4**, 551–555.
24. Chen,Y., Dougherty,E. and Bittner,M. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt*., **2**, 364–374.
25. Golfier,G., Tran,D.M., Dauphinot,L., Graison,E., Rossier,J. and Potier,M.C. (2004) VARAN: a web server for VARiability ANalysis of DNA microarray experiments. *Bioinformatics*, **20**, 1641–1643.
26. Kerr,M.K. (2003) Experimental design to make the most of microarray studies. *Methods Mol. Biol*., **224**, 137–147.
27. Bottomley,S. (2004) Bioinformatics: smartest software is still just a tool. *Nature*, **429**, 241.
28. Bédrine-Ferran,H., Le Meur,N., Gicquel,I., Le Cunff,M., Soriano,N., Guisle,I., Mottier,S., Monnier A., Teusan,R., Fergelot,P., Le Gall,J.Y., Leger,J.J. and Mosser,J.. (2004) Transcriptome variations in human Caco-2 cells: a model for enterocyte differentiation and its link to iron absorption. *Genomics*, **83**, 747–950.
29. Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,A. and Peterson,C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol*., **3**, SOFTWARE0003.

Estimation of false positive and false negative rates. SAM (7) modified two-class t-tests were performed to detect differentially expressed genes between data sets extracted from the heart expression data of experiments 1 and 2 using different mixtures of within and between replication conditions:

- 1 chip with 4 replicates,
- 2 chips with 4 replicates per chip,
- 3 chips with 4 replicates per chip,
- 4 chips with 1 replicate per chip,
- 4 chips with 2 replicates per chip,
- 4 chips with 3 replicates per chip.

The replication conditions were tested against the experimental condition of 4 chips with 4 replicates per chip. Two by two confusion matrices (Table 1) containing information about the experimental and the predicted classifications in each of the different replication conditions were used to estimate the false positive and false negative rates.

Table 1. Two by two confusion matrix for false positive and false negative rates calculation.

		Predicted	
		Negative	Positive
Experimental	Negative	a	b
	Positive	c	d

Where *a* is the number of correct predictions that an instance is negative,
b is the umber of incorrect predictions that an instance is positive,
c is the number of incorrect predictions that an instance is negative and
d is the number of correct predictions that an instance is positive.

The false positive rate (FP) is defined as the proportion of negatives cases that were incorrectly classified as positive in the predicted condition compared to the actual observation (Eq. 1.). The false negative rate (FN) is defined as the proportion of positive cases that were incorrectly classified as negative in the predicted condition compared to the actual observation (Eq. 2.).

$$FP \equiv \frac{b}{b+a} \qquad \text{Eq. 1.}$$

$$FN \equiv \frac{c}{c+d} \qquad \text{Eq. 2.}$$

2. Résumés des articles publiés ou soumis utilisant MADSCAN pour le traitement des données primaires

MADSCAN a été utilisé pour le traitement des données primaires d'un certain nombre d'études réalisées au sein du laboratoire ou dans le cadre de collaborations. Pour certains de ces travaux, j'ai également participé à l'analyse des données (*i.e.* au-delà des étapes de traitement des données par MADSCAN). Quelques uns de ces travaux ont fait l'objet d'une publication tandis que d'autres sont actuellement soumis à différents journaux. Dans ce paragraphe sont présentés : la première page des articles publiés, les résumés des articles soumis et un résumé en français des hypothèses et des résultats de ces différentes études.

Articles publiés

Steenman, M., Lamirault, G., **Le Meur, N.**, Le Cunff, M., Escande, D., and Leger, J. J. (2005). **Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays**. Eur.J.Heart Fail., **7**: 157-165.

Bedrine-Ferran, H., **Le Meur, N.**, Gicquel, I. *et al.* (2004). **Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption**. Genomics, **83**: 772-789.

Le Bouter, S., El Harchi, A., Marionneau, C. *et al.* (2004). **Long-term amiodarone administration remodels expression of ion channel transcripts in the mouse heart.** Circulation, **110**: 3028-3035.

Articles soumis

Lamirault, G.; **Le Meur, N.**, Chevalier, C.; Le Cunff, M.F., Guisle, I., Bihouée, A., Baron, O., Trochu, J.N., Léger, J.J. and Steenman, M. **The clinical profile of heart transplant candidates is associated with progressive transcriptomal cardiac remodeling.**

Troadec, M.-B, Glaise, D., Lamirault, G., Le Cunff, M., Guérin, É., **Le Meur, N.**, Zindy, P.J., Leroyer, P., Guisle, I., Duval, H., Gripon, P., Théret N., Guillouzo, C., Brissot, P., Léger, J.J. and Loréal, O. **Gene expression modulation associated with the differentiation process of the liver bipotent HepaRG cell line : Implications for the understanding of iron metabolism.**

McIlroy, D., Tanguy-Royer, S., **Le Meur, N.**, Guisle, I., Royer, J-P., Léger, J.J., Meflah, K. and Marc Grégoire, M. **Profiling dendritic cell maturation with dedicated micro-arrays.** Journal of Leukocyte biology *(accepté)*

ELSEVIER

The European Journal of Heart Failure 7 (2005) 157–165

The European Journal of Heart Failure

www.elsevier.com/locate/heafai

Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays

Marja Steenman*, Guillaume Lamirault, Nolwenn Le Meur, Martine Le Cunff, Denis Escande, Jean J. Léger

Laboratoire de Physiopathologie et de Pharmacologie Cellulaires et Moléculaires, INSERM U533, Faculté de Médecine, 1 Rue Gaston Veil, BP 53508, 44035 Nantes, Cedex 1, France

Received 31 March 2004; received in revised form 19 April 2004; accepted 17 May 2004
Available online 30 July 2004

Abstract

Aims: This study aimed to investigate whether a molecular profiling approach should be pursued for the classification of heart failure patients. *Methods and results:* Applying a subtraction strategy we created a cDNA library consisting of cardiac- and heart failure-relevant clones that were used to construct dedicated cDNA microarrays. We measured relative expression levels of the corresponding genes in left ventricle tissue from 17 patients (15 failing hearts and 2 nonfailing hearts). Significance analysis of microarrays was used to select 159 genes that distinguished between all patients. Two-way hierarchical clustering of the 17 patients and the 159 selected genes led to the identification of three major subgroups of patients, each with a specific molecular portrait. The two nonfailing hearts clustered closely together. Interestingly, our classification of patients based on their molecular portraits did not correspond to an identified etiological classification. Remarkably, patients with the highest medical urgency status (United Network for Organ Sharing, Status 1A) clustered together. *Conclusion:* With this pilot feasibility study we demonstrated a novel classification of end-stage heart failure patients, which encourages further development of this approach in prospective studies on heart failure patients at earlier stages of the disease.

Keywords: Classification; Idiopathic dilated cardiomyopathy; Coronary artery disease; Gene expression profiling; Cluster analysis

1. Introduction

Molecular expression profiling studies conducted in human breast cancer [1], prostate cancer [2], and embryonal tumours of the central nervous system [3] have led to the ultimate utilization of cDNA microarrays to predict clinical outcome based on a tumour's expression profile. Heart failure has recently been compared to cancer [4], in that it involves the same biological principles of cell growth, death, and survival. We thus wondered whether molecular expression profiling could also be used to classify failing hearts, i.e. whether failing hearts show distinct molecular profiles irrespective of their aetiology.

Since human cardiac tissue is evidently less accessible than tumour material, expression profiling has been performed to a lesser extent in human cardiac disease. The first cardiac expression profiles were based on in silico analyses of expressed sequence tags (ESTs) obtained from cardiac cDNA libraries, leading to catalogues of genes expressed in normal or hypertrophied hearts [5–7]. These studies were followed by microarray analyses identifying genes with aberrant expression levels in failing hearts [8–13]. More recently, attempts were conducted to classify small groups of patients with end-stage heart failure based on their expression profile. One study described the classification of seven failing and five nonfailing hearts using the expression ratios of *all* clones on their array [11]. In this study, most of the failing hearts clustered together, which raised a concern as to the usefulness of classification based on molecular portraits. A second study classified eight failing and eight nonfailing hearts, based on the expression ratios of those genes that were differentially expressed between the *group* of failing and the *group* of nonfailing hearts [12]. Their results showed that two patients with an aetiology distinct from the rest (alcoholic and familial cardiomyopathy) clustered away from the other failing hearts.

* Corresponding author. Tel.: +33-240412844; fax: +33-240412950.
E-mail address: marja.steenman@nantes.inserm.fr (M. Steenman).

1388-9842/$ - see front matter © 2004 European Society of Cardiology. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ejheart.2004.05.008

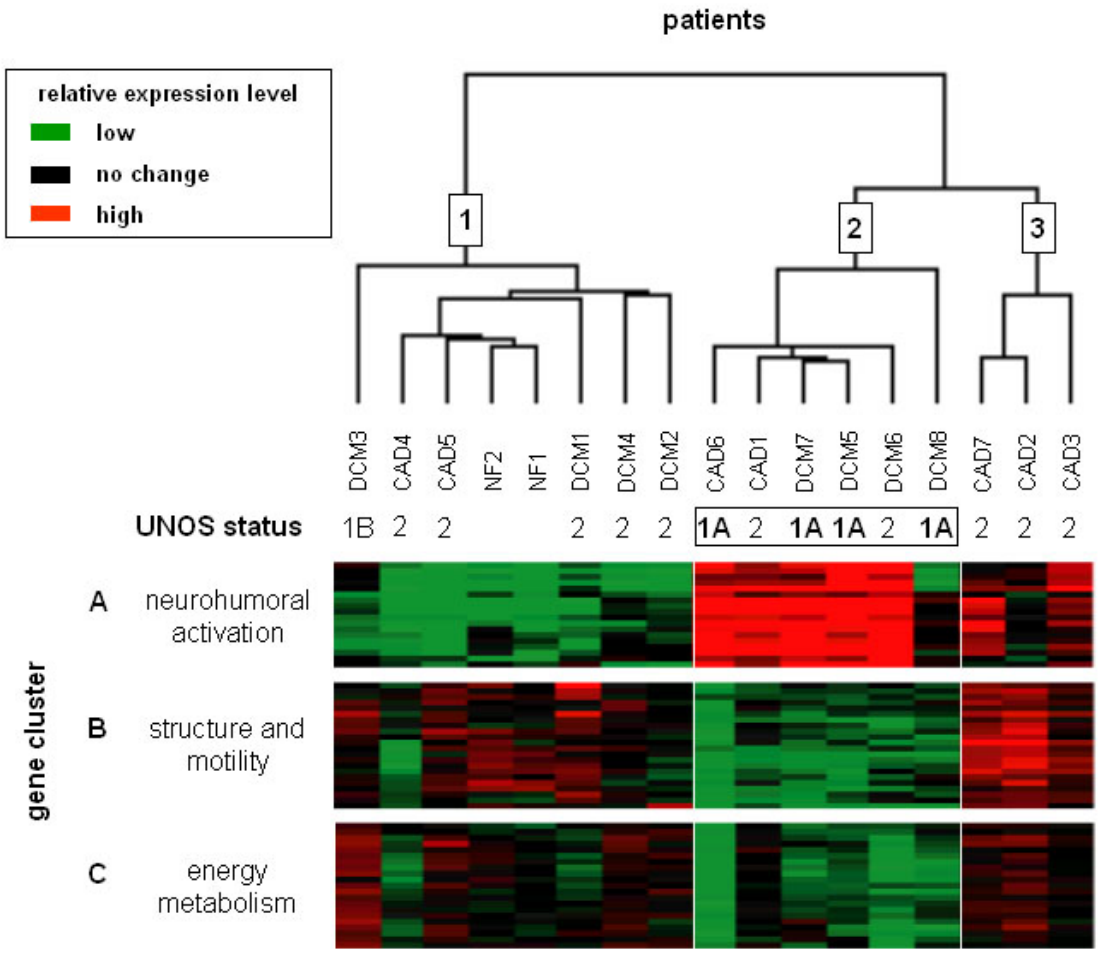


Figure 20. Classification des patients insuffisants cardiaques à partir des profils d'expression génique (d'après Steenman *et al.* (2005)). Pour chaque *gene reporter* (ligne) et chaque patient (colonne), la valeur d'expression est ajustée de telle sorte que l'expression médiane de chaque *gene reporter* pour les 17 patients correspond à « no change » (noir). Les malades sont classés en 3 groupes (« 1 », « 2 », « 3 »). Les profils d'expression des groupes de *gene reporters* A, B et C sont fortement corrélés avec la classification des malades en 3 groupes. Pour chacun des groupes, une fonction moléculaire prépondérante est mise en évidence. Le groupe A est associé aux fonctions d'activation neuro-humorale, le groupe B aux fonctions de structure et de contraction musculaire et le groupe C aux fonctions du métabolisme énergétique. DCM : cardiomyopathie dilatée ; CAD cardiopathie ischémique ; NF : (*non failing*) absence d'insuffisance cardiaque.

2.1 Portaits moléculaires de patients en insuffisance cardiaque : étude pilote

Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays. (Steenman *et al.*, 2005)

Par analogie aux travaux menés dans les cancers, nous avons réalisé une étude pilote pour évaluer la faisabilité d'une classification des patients en insuffisance cardiaque en fonction de leur portrait moléculaire. Dans ce but, nous avons étudié le transcriptome du ventricule gauche de 17 patients dont 15 en insuffisance cardiaque.

Pour cette étude, une puce à ADN dédiée à l'étude du transcriptome cardiaque a été dessinée. Les clones déposés sur la puce ont été sélectionnés par SSH et une analyse approfondie de la littérature. Cette puce contient 440 *gene reporters*, représentant 440 gènes exprimés dans le cœur. Afin d'évaluer la reproductibilité des mesures, ces *gene reporters* ont été déposés au minimum 2 fois sur la puce. De même, pour chaque échantillon (patient), l'ARN a été hybridé sur 3 puces. Les 17 échantillons ont été comparés entre eux grâce à un pool de référence composé de l'ARN de 8 de ces patients.

Après l'obtention, par MADSCAN, d'une matrice de données consolidées pour chaque échantillon et la mise en évidence de 159 gènes significativement différentiels par SAM (*cf.* p.64-65) (Tusher *et al.*, 2001), les patients ont été regroupés selon une classification hiérarchique ascendante (*cf.* p. 72-73). Les résultats ont montré que les regroupements basés sur les profils moléculaires ne correspondaient pas à l'étiologie des patients (Fig. 20). En revanche, l'un des groupes obtenus contenait une majorité de patients particulièrement atteints par la maladie, *i.e.* appartenant à la classe d'urgence médicale la plus élevée soit UNOS 1A (***U**nited **N**etwork for **O**rgan **S**haring*). Malgré certaines limites (petit nombre d'échantillons, phase terminale de l'insuffisance cardiaque), cette étude a permis de regrouper des patients en insuffisance cardiaque sur la base de leur portrait moléculaire. Elle offre également des résultats encourageants quant à l'amélioration de la définition des classifications cliniques. Il semble donc raisonnable de penser qu'un projet à plus grande échelle, avec le recueil de biopsies cardiaques, pourrait aider au pronostic et au diagnostique thérapeutique des patients insuffisants cardiaques.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 83 (2004) 772–789
GENOMICS
www.elsevier.com/locate/ygeno

Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption

Hélène Bédrine-Ferran,[a] Nolwenn Le Meur,[b] Isabelle Gicquel,[a] Martine Le Cunff,[b] Nicolas Soriano,[a] Isabelle Guisle,[b] Stéphanie Mottier,[a] Annabelle Monnier,[a] Raluca Teusan,[b] Patricia Fergelot,[a] Jean-Yves Le Gall,[a] Jean Léger,[b] and Jean Mosser[a,*]

[a] *UMR-6061 CNRS, IFR 97, Faculté de Médecine, CS 34317-35043 Rennes, France*
[b] *U533 INSERM, 44035 Nantes, France*

Received 25 September 2003; accepted 19 November 2003

Abstract

Complete clinical expression of the HFE1 hemochromatosis is very likely modulated by genes linked to duodenal iron absorption, whose level is conditioned by unknown processes taking place during enterocyte differentiation. We carried out a transcriptomic study on CaCo-2 cells used as a model of enterocyte differentiation in vitro. Of the 720 genes on the microarrays, 80, 50, and 56 were significantly down-regulated, up-regulated, and invariant during differentiation. With regard to iron metabolism, we showed that *HEPH*, *SLC11A2*, *SLC11A3*, and *TF* are significantly up-regulated, while *ATP7B* and *SLC39A1* (and *SFT*) are down-regulated and *ACO1*, *dCYTb*, *FECH*, and *FTH1* show constant expression. Ontological annotations highlight the decrease in the expression of cell cycle and DNA metabolism associated genes as well as transcription, protein metabolism, signal transduction, and nucleocytoplasmic transport associated genes, whereas there are increases in the expression of genes linked to cell adhesion, lipid and xenobiotic metabolism, iron transport and homeostasis, and immune response.

Keywords: Iron overload; Enterocytes; Microarray analysis of gene expression; CaCo-2 cells; Cell differentiation; Ontological annotations

Introduction

Primary iron overload or hemochromatosis (HHC) is a frequently encountered human disease characterized by excessive iron absorption. This disorder appears to be genetically heterogeneous. Although *HFE1* [1] accounts for most of the cases, it has recently been demonstrated that other genes are involved in either hemochromatosis (for those coding TfR2, Ireg-1, H-Ferritin, and Hepcidin, respectively designated *HFE3* [2], *HFE4* [3], *HFE5* [4] and *HFE6* [5]) or its juvenile form for the *HFE2* locus [6]. Not all of these genes account for all of the primary iron overload disorders, and others are still to be identified. Despite this genetic heterogeneity, more than 90% of patients from Northern Europe are homozygous for the C282Y mutation in *HFE1* [7]. Nevertheless, the incomplete penetrance of this mutation [8] strongly suggests that its clinical expression is modulated by other factors, some of which are genetic in nature. Among these factors, those regulating intestinal iron absorption are certainly determining since hemochromatosis is due to excessive iron absorption [9].

Intestinal iron absorption is certainly a critical step for the regulation of the total amount of iron in the organism. It takes place almost exclusively in the duodenum, where apical iron uptake and basolateral transfer are performed by SLC11A2 (solute carrier family 11 member 2, also known as iron uptake transporter DMT1/DCT1/Nramp2) and SLC11A3 (solute carrier family 11 member 3, also known as iron export transporter IREG1/ferroportin1/MTP1). In the duodenum, it is known that the HFE protein can form a complex with the transferrin receptor (TfR) [10] and that these two proteins are expressed in crypt cells, exclusively for HFE and predominantly for TfR [10]. Since crypt cells are mostly undifferentiated, this localization

* Corresponding author. UMR6061 CNRS, Faculté de Médecine, 2 av du Prof L. Bernard, 35043 Rennes Cedex, France. Fax: +33-2-23-23-44-78.
E-mail address: Jean.Mosser@univ-rennes1.fr (J. Mosser).

0888-7543/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2003.11.014

2.2 Etude du transcriptome des cellules CaCo-2 au cours de leur différenciation et en relation avec leur capital ferrique -

Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption (Bedrine-Ferran *et al.*, 2004).
Collaboration : UMR6061, Jean Mosser, Rennes

Les travaux de Bédrine-Ferran *et al* (2004) portent sur la régulation de l'absorption duodénale du fer et son dysfonctionnement dans l'hémochromatose génétique. Cette maladie héréditaire est l'une des plus fréquente en Europe du Nord avec une prévalence d'environ 3-5 individus sur 1000. Elle se caractérise par une réduction de la capacité de stockage du fer dans le système réticulo-endothélial et une hyper-absorption du fer au niveau de l'intestin grêle. Cette surcharge peut être à l'origine de pathologies plus sévères telles que le diabète, les cardiomyopathies ou la cirrhose du foie. L'absorption duodénale du fer est en partie régulée par des mécanismes intervenant au cours de la différenciation des entérocytes. Ces mécanismes semblent également moduler la pénétrance clinique de l'hémochromatose génétique. Afin d'étudier la mise en place de ces mécanismes, nous avons analysé le transcriptome de cellules CaCo-2, modèle *in vitro* des cellules de l'intestin grêle, lors de leur différenciation.

Ce travail présente une exploitation originale des résultats générés par MADSCAN. L'outil a été utilisé non seulement pour traiter et transformer les données mais aussi pour mettre en évidence des gènes « invariants » sur la base de l'algorithme de Tseng *et al.* (2001). L'annotation de ces gènes par MADSENSE (Teusan, 2002) et PubGene (Jenssen *et al.*, 2001) a notamment permis de montrer que les phénomènes d'apoptose et les mécanismes de réponse au stress restent constants durant la différenciation. Nous avons également montrés que le nombre de gènes sur-exprimés croît de façon constante au cours de la différenciation. Ainsi, des gènes directement impliqués dans le métabolisme du fer sont davantage exprimés dans les cellules différenciées.

Long-Term Amiodarone Administration Remodels Expression of Ion Channel Transcripts in the Mouse Heart

Sabrina Le Bouter, MSc; Aziza El Harchi, MSc; Céline Marionneau, MSc; Chloé Bellocq, MSc; Arnaud Chambellan, MD; Toon van Veen, PhD; Christophe Boixel, PhD; Bruno Gavillet, MSc; Hugues Abriel, MD, PhD; Khai Le Quang, MD; Jean-Christophe Chevalier, MD; Gilles Lande, MD; Jean J. Léger, PhD; Flavien Charpentier, PhD; Denis Escande, MD, PhD; Sophie Demolombe, PhD

Background—The basis for the unique effectiveness of long-term amiodarone treatment on cardiac arrhythmias is incompletely understood. The present study investigated the pharmacogenomic profile of amiodarone on genes encoding ion-channel subunits.

Methods and Results—Adult male mice were treated for 6 weeks with vehicle or oral amiodarone at 30, 90, or 180 mg $\cdot$ kg^{-1} $\cdot$ d^{-1}. Plasma and myocardial levels of amiodarone and *N*-desethylamiodarone increased dose-dependently, reaching therapeutic ranges observed in human. Plasma triiodothyronine levels decreased, whereas reverse triiodothyronine levels increased in amiodarone-treated animals. In ECG recordings, amiodarone dose-dependently prolonged the RR, PR, QRS, and corrected QT intervals. Specific microarrays containing probes for the complete ion-channel repertoire (IonChips) and real-time reverse transcription–polymerase chain reaction experiments demonstrated that amiodarone induced a dose-dependent remodeling in multiple ion-channel subunits. Genes encoding Na^+ (SCN4A, SCN5A, SCN1B), connexin (GJA1), Ca^{2+} (CaCNA1C), and K^+ channels (KCNA5, KCNB1, KCND2) were downregulated. In patch-clamp experiments, lower expression of K^+ and Na^+ channel genes was associated with decreased $I_{to,f}$, $I_{K,slow}$, and I_{Na} currents. Inversely, other K^+ channel α- and β-subunits, such as KCNA4, KCNK1, KCNAB1, and KCNE3, were upregulated.

Conclusions—Long-term amiodarone treatment induces a dose-dependent remodeling of ion-channel expression that is correlated with the cardiac electrophysiologic effects of the drug. This profile cannot be attributed solely to the amiodarone-induced cardiac hypothyroidism syndrome. Thus, in addition to the direct effect of the drug on membrane proteins, part of the therapeutic action of long-term amiodarone treatment is likely related to its effect on ion-channel transcripts. **(*Circulation*. 2004;110:3028-3035.)**

Key Words: antiarrhythmic agents ■ ion channels ■ molecular biology ■ electrophysiology

Amiodarone, a widely used antiarrhythmic drug, has remarkable efficacy for the treatment of ventricular tachyarrhythmias and atrial fibrillation. However, the basis for its effectiveness is still poorly understood. The pharmacologic profile of this drug is complex, and much remains to be clarified about both short- and long-term actions. Amiodarone has been referred to as a class III antiarrhythmic agent,[1] but it also possesses electrophysiologic characteristics of class I and IV agents and minor class II effects.[2] The drug is also known to modify thyroid function extensively because of its iodinated nature.[3]

The question arose as to whether the long-term effects of amiodarone might stem from its molecular interaction with thyroid hormone receptors or other mechanisms. In particular, it has been hypothesized that the effects of amiodarone could depend on modulation of transcript expression in addition to its direct effect on cell membrane channels.[4] Genomic techniques now bring gene expression studies to a genome scale, allowing investigators to examine simultaneous changes in the expression of the complete gene repertoire. We have developed a specific cDNA microarray (IonChip) containing probes for virtually all known mouse and human ion-channel genes (α- and β-subunits).[5] With this new tool, our previous investigation explored ion-channel remodeling as produced by altered thyroid status in the mouse heart.[5] We now evaluate the effects of long-term amiodarone

Received June 15, 2004; revision received August 4, 2004; accepted August 19, 2004.

From l'Institut du thorax—INSERM U533 (S.L.B., A.E.H., C.M., C.B., A.C., K.L.Q., J.C.C., G.L., J.J.L., F.C., D.E., S.D.), Faculté de Médecine, Nantes, France; the Department of Medical Physiology (T.V.V.), University Medical Center, Utrecht, The Netherlands; and Institut de Pharmacologie (C.B., B.G, H.A.), Lausanne, Switzerland.

The first 2 authors contributed equally to this work.

An online-only Data Supplement is available at http://www.circulationaha.org

Correspondence to Sophie Demolombe, INSERM U533, Faculté de Médecine, 1 rue G. Veil, 44035 Nantes Cedex, France. E-mail sophie.demolombe@nantes.inserm.fr

Circulation **is available at http://www.circulationaha.org** **DOI: 10.1161/01.CIR.0000147187.78162.AC**

2.3 Effet pharmacologique de l'amiodarone sur le remodelage de l'expression des canaux ioniques du cœur de la souris.

Long-term amiodarone administration remodels expression of ion channel transcripts in the mouse heart (Le Bouter *et al.*, 2004).

L'amiodarone est un médicament couramment employé dans le traitement des arythmies cardiaques telles que les tachyarythmies ventriculaires et les fibrillations auriculaires. Dans le but de mieux comprendre les mécanismes de son efficacité, Le Bouter *et al.* (2004) ont étudié l'effet à long terme d'une dose d'amiodarone sur le cœur de souris.

Pour cette étude, une puce à ADN dédiée à l'étude du transcriptome des canaux ioniques (*IonChip*) a été dessinée (Le Bouter *et al.*, 2003). Cette puce contient près de 208 *gene reporters*, représentant la majorité des canaux ioniques et des connexines actuellement connus chez la souris. Les échantillons étudiés ont été divisés en 4 groupes. Trois groupes de souris « test » ont reçu différentes concentrations d'amiodarone durant 6 semaines. Un dernier groupe de souris « contrôle » (*Sham*) a été utilisé comme référence. L'ensemble des puces réalisées a été traité et validé par MADSCAN.

Les résultats obtenus par l'analyse des *IonChips* mettent en évidence un effet dose dépend de l'amiodarone sur le remodelage de l'expression des canaux ioniques du cœur. Cette étude a améliorée notre connaissance sur les mécanismes d'action de l'amiodarone et démontre ainsi l'efficacité des puces à ADN dédiées dans l'analyse de l'effet des traitements thérapeutiques sur le trancriptome.

2.4 Le profil d'expression de patients en attente de transplantation cardiaque est associé avec un remodelage progressif du transcriptome cardiaque. (Article soumis)

The clinical profile of heart transplant candidates is associated with progressive transcriptomal cardiac remodeling. Lamirault, G., Le Meur, N., Chevalier, C. Le Cunff, MF.; Guisle, I., Bihouée, A., Baron, O., Trochu, JN., Léger, JJ. and Steenman, M.

Abstract

Gene expression profiling has led to successful classification of patients and prediction of clinical outcome in many diseases. We previously demonstrated that end-stage heart failure patients can be distinguished based on their cardiac transcriptomal profile. In the here presented study, we investigated whether the clinical profile of heart transplant candidates was associated with a specific cardiac transcriptomal profile. We selected 44 end-stage heart failure patients who underwent a heart transplantation or a total artificial heart placement. Patients were divided into 3 groups (deteriorating, intermediate, and stable clinical profile). Gene expression profiles were obtained for both left and right ventricles of the 44 explanted hearts using an in-house made microarray representing 4217 cardiac-relevant genes. Data were filtered and analyzed using MADTOOLS, a suite of software developed in our laboratory (www.madtools.org). Significance Analysis of Microarrays was used to select 165 and 170 genes that distinguished between deteriorating and stable clinical profiles in left and right ventricles respectively. For each cardiac chamber, a patient classification was determined. Principal component analysis (PCA) was conducted on the selected genes (used as variables). The first two components of PCA analysis were then used to classify all 44 patients. The obtained gene expression-based classifications revealed a clear distinction between the three clinical profiles associated with a progressive gene expression remodeling. Interestingly, a strong correlation between left and right ventricle-based classifications was found. Genes involved in the extracellular matrix, muscle contraction, and the cytoskeleton were highly represented in this remodeling process. We demonstrated that clinical profiles of end-stage heart failure patients are associated with a specific gene expression profile. These results reinforce the hypothesis that, in the future, gene expression profiles may be used to improve the prognostic evaluation of heart failure patients.

*

* *

Ce travail fait suite à l'étude pilote, menée sur 15 patients en insuffisances cardiaques, qui nous a permis de montrer la faisabilité et l'intérêt d'une classification des patients en insuffisance cardiaque en fonction de leur portrait moléculaire (Steenman *et al.*, 2005).

Cette nouvelle étude est réalisée sur 44 patients en insuffisance cardiaque terminale ayant subis une transplantation cardiaque ou la mise en place d'un cœur artificiel. Des portraits moléculaires ont été obtenus pour les ventricules gauches et droits des 44 cœurs explantés, utilisant les *MyoChips*. Ces puces dédiées, développées au sein du laboratoire, se composent de 4217 *gene reporters* spécifiques du muscle cardiaque. L'ensemble des étapes expérimentales et les résultats obtenus ont été sauvegardés dans MADSTAR puis transférés dans BASE grâce à Audrey Bihouée. Les données primaires ont été traitées par MADSCAN (Le Meur *et al.*, 2004) et l'annotation des *gene reporters* a été réalisée par l'outil MADSENSE de Raluca Teusan.

Une comparaison entre les profils d'expression de 12 patients aux statuts « critiques » (récent épisode aigu de la maladie) et 13 patients aux statuts « stables » (pas d'épisode aigu de la maladie depuis 3 mois) a permis de mettre en évidence 165 et 170 *gene reporters* différentiellement exprimés, respectivement dans les ventricules gauches et droits (Tusher *et al.*, 2001).

Une analyse en composante principale (ACP) sur 25 patients a été appliquée à l'ensemble des gènes sélectionnés dans les ventricules gauches (165). Les deux premières composantes ont été utilisées pour classer l'ensemble des 44 patients. La classification obtenue met en évidence une séparation en 3 profils cliniques, caractérisés par un remodelage progressif de l'expression des gènes. De plus, une forte corrélation a été observée entre les classifications basées sur les profils d'expression des ventricules droits et gauches. Les gènes principalement impliqués dans ce phénomène de remodelage interviennent au niveau des mécanismes de la matrice extra-cellulaire, de la contraction musculaire et du cytosquelette.

Cette étude confirme nos premiers résultats à savoir l'intérêt de l'utilisation de portraits moléculaires pour améliorer le pronostic des patients en insuffisance cardiaque. De plus, ces résultats démontrent que les profils cliniques (plus ou moins graves) sont associés à de profils d'expression spécifiques.

2.5 Profils d'expression des cellules dendritiques en maturation par des puces ADN dédiées. (Article accepté)

Collaboration INSERM U419, Dorian McIlroy, Nantes

Profiling dendritic cell maturation with dedicated micro-arrays

Dorian McIlroy, Séverine Tanguy-Royer, Nolwenn Le Meur, Isabelle Guisle, Pierre-Joseph Royer , Jean Léger, Khaled Meflah, and Marc Grégoire.

Abstract

Dendritic cell (DC) maturation is the process by which immature DC in the periphery differentiate into fully competent antigen presenting cells that initiate the T-cell response. However, DC respond to many distinct maturation stimuli, and different types of mature DC induce qualitatively different T-cell responses. Since DC maturation involves the co-ordinated regulation of hundreds of genes, comprehensive assessment of DC maturation status would ideally involve monitoring the expression of all of these transcripts. However, whole-genome micro-arrays are not well suited for routine phenotyping of DC, since the vast majority of genes represented on such chips are not relevant to DC biology, and their cost limits their use for most laboratories. We therefore developed a DC-dedicated micro-array, or "DC Chip" incorporating probes for 121 genes upregulated during DC maturation, 93 genes downregulated during maturation, 14 DC-specific genes and 90 other genes with known or probable immune functions. These micro-arrays were used to study the kinetics of DC maturation and the differences in maturation profiles between five healthy donors after stimulation with TNFα + polyI:C. Results obtained with the DC Chip were consistent with flow cytometry, ELISA and real-time PCR, as well as previously published data.. Furthermore, the co-ordinated regulation of a cluster of genes (IDO, KYNU, KMO, WARS and HAAO) involved in tryptophan metabolism was observed. These data demonstrate the utility of the DC Chip for monitoring the molecular processes involved in the orientation of the immune response by DC.

*

* *

La maturation des cellules dendritiques (DC –*Dendritic Cell*) est le mécanisme par lequel les DC immatures périphériques se différencient en cellules présentatrices d'antigènes qui initient la réponse immunitaire des cellules T. Toutefois, la réponse des cellules dendritiques aux différents stimuli de la maturation ainsi que les différents types de DC matures peuvent induire des réponses de cellules T qualitativement très différentes.

Dans le but d'étudier la diversité et le mode de régulation des gènes impliqués dans ces mécanismes, une puce à ADN dédiée à l'analyse du transcriptome des cellules dendritiques, baptisée « DC-chip », a été développée. Cette puce se compose de 300 gènes reporters, dont 121 et 93 *gene reporters* respectivement sur- et sous-exprimés pendant la phase de maturation des DC, 14 gènes spécifiques des DC et 90 probablement impliqués dans les mécanismes de la fonction immunitaire. Cette puce à ADN a été utilisée pour une étude pilote sur la cinétique des mécanismes de la maturation chez 5 donneurs sains, après stimulation aux TNFβ et polyIC des DC immatures. L'ensemble des données primaires ont été traitées et validées par MADSCAN. Les résultats obtenus ont été confirmés par cytométrie de flux, ELISA et RT-PCR, ainsi que par les données de la littérature.

2.6 Remodelage de l'expression des gènes associée au processus de différenciation de la lignée cellulaire hépatique bipotente HepaRG. (Article soumis)

Collaboration : INSERM U522, Olivier Loréal, Rennes.

Gene expression modulation associated with the differentiation process of the liver bipotent HepaRG cell line : Implications for the understanding of iron metabolism

Troadec M.-B, Glaise D., Lamirault G., Le Cunff M., Guérin É., Le Meur N., Zindy P.-J., Leroyer P., Guisle I., Duval H., Gripon P., Théret N., Guillouzo C., Brissot P., Léger J.J. and Loréal, O.

Abstract

An association between hepatocyte iron storage capacity and differentiation has been suggested. In order to characterize biological processes involved, we studied gene expression modulation by a transcriptomic approach in the human HepaRG cell line which undergoes high hepatocyte differentiation. Four successive differentiation stages of this cell line were studied: proliferation, confluence, superconfluence and high differentiation. From proliferation to confluence, culture did not show iron-citrate loading. In contrast, from superconfluence, iron exposure led to iron loading. This iron storage capacity occurred when genes implicated in cell motility and biosynthesis were underexpressed, and those involved in lipid metabolism and immune response, signing hepatocyte differentiation, were overexpressed. Our results demonstrate that hepatocyte iron storage capacity is associated with both a decreased expression of genes involved in cell motility and with the appearance of

differentiated hepatocyte functions. They provide new insights in the understanding of iron and hepatocyte differentiation relationship during iron-related hepatic diseases.

*
* *

Les hépatocytes jouent un rôle majeur dans le métabolisme du fer, notamment par leur capacité à capter et stocker le fer plasmatique circulant en excès. Chez l'homme, la présence d'une surcharge en fer du foie, comme dans le cas d'une hémochromatose génétique, facilite le développement des cancers du foie. De plus, dans les cellules d'un hépatocarcinome, cellules proliférantes et dédifférenciées, aucun dépôt de fer n'est observé même si la partie non-tumorale de ce foie est surchargée en fer. Tout ceci suggère que le fer jouerait un rôle dans la carcinogenèse hépatique et que le métabolisme du fer serait modulé par l'état de prolifération/différenciation des hépatocytes.

La lignée cellulaire HepaRG est, à ce jour, la seule lignée d'origine hépatocytaire, présentant une capacité de différenciation en hépatocyte. Elle est a été obtenue à partir d'un hépatocarcinome et est caractérisée par sa grande capacité de différenciation en deux types cellulaires : des cellules de type hépatocytaire et des cellules de type biliaire. Les hépatocytes étant un des sites majeurs de stockage du fer, cette lignée a donc été choisie comme modèle cellulaire pour étudier la capacité de stockage du fer en relation avec la différenciation hépatocytaire, en suivant la modulation de l'expression des gènes.

Afin d'étudier le transcriptome des cellules HepaRG, une puce à ADN dédiée à l'analyse du transcriptome du foie humain a été dessinée. Le plan expérimental mis en place est une cinétique en 4 phases, correspondant aux principales étapes de la différenciation de la lignée HepaRG : prolifération, confluence, super-confluence et différenciation. La validation et les traitements des données issues des puces à ADN ont été réalisés par MADSCAN.

Les résultats montrent que la lignée ne se surcharge pas en présence de fer dans le milieu de culture tant qu'elle reste proliférante, *i.e.* jusqu'à la confluence. Nous avons observé que la lignée HepaRG est capable de se surcharger en fer lorsque la culture devient superconfluente, avec l'apparition de cellules de morphologie et phénotype hépatocytaire comme biliaire. La surcharge en fer est visible uniquement dans les hépatocytes, et est renforcée avec leur différenciation. En l'absence d'addition de fer, cette capacité de surcharge en fer est, d'une part, associée à une répression de l'expression des gènes impliquées dans la

prolifération, la migration, la motilité cellulaire et métabolisme du fer tels que le récepteur de la transferrine 1 et les ferritines L. D'autre part, elle est associée à une sur-expression de gènes de différenciation comme ceux impliqués dans la réponse au stress, par exemple des gènes des cytochromes P450, ou lié au métabolisme du fer, tels la transferrine, l'hémopexine et l'haptoglobine.

Ces résultats démontrent que la capacité de stockage du fer dans les hépatocytes est associée à la diminution de l'expression des gènes impliqués dans les mécanismes de motilité cellulaire et l'apparition des fonctions des hépatocytes différenciés.

3. Etat actuel et évolution de MADSCAN

3.1 Fréquentation

MADSCAN est accessible à la communauté scientifique depuis mars 2003. Depuis cette date, plus de 3000 connections ont été effectuées et 3300 analyses réalisées. La figure 21 présente la fréquentation pour l'année 2004. MADSCAN est utilisée majoritairement en France mais commence à être employé dans la majorité des pays européens, aux USA et en Asie (Inde, Corée du Sud).

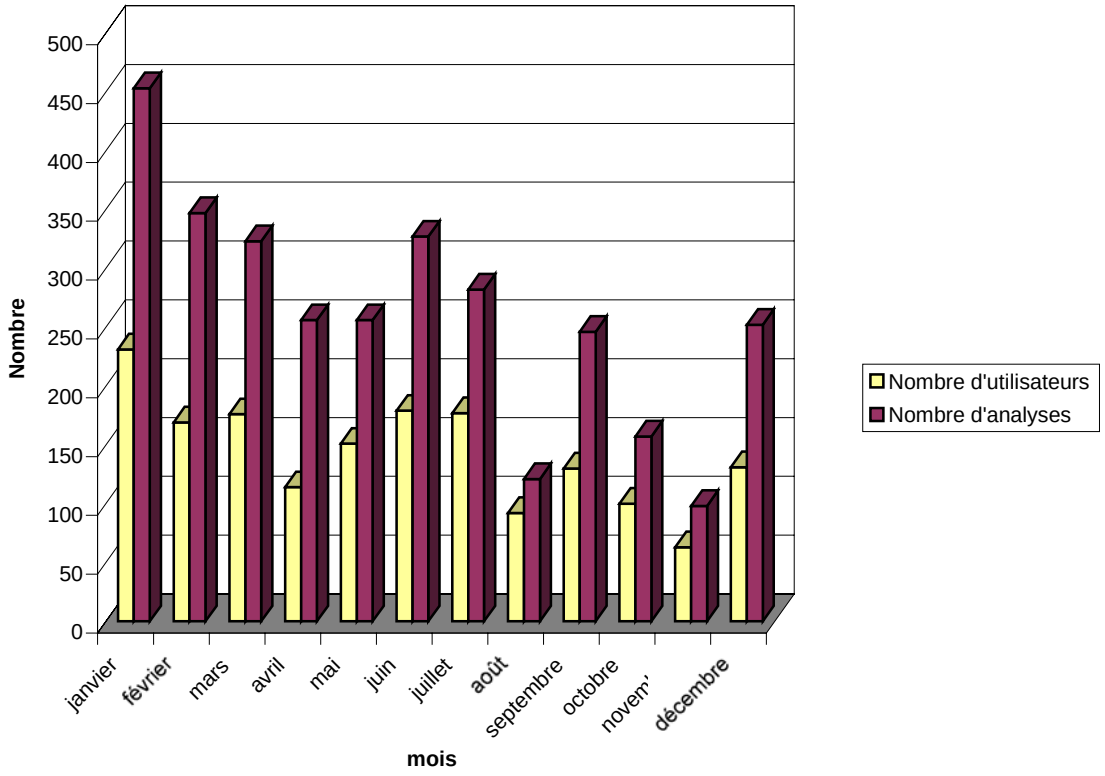


Figure 21. Fréquentation de MADSCAN pour l'année 2004. (source : AddFreeStats http://www3.addfreestats.com/index.html). Le nombre d'analyse étant supérieur au nombre d'utilisateurs, un utilisateur réalise une ou plusieurs analyses.

3.2 Evolution récente

Un progrès récent est l'**interfaçage de MADSCAN avec la base de données BASE** (Saal *et al.*, 2002) (*cf.* p.39-40), réalisé au sein du laboratoire par Audrey Bihouée. En effet, BASE permet d'intégrer des modules d'analyses, ou *plug-ins,* dans son architecture. Cinq *plug-ins* sont disponibles en standard avec la version de distribution de BASE et environ une

vingtaine sont téléchargeables sur le site Web du projet. Toutefois, ces *plug-ins* effectuent généralement un traitement partiel des données tandis que MADSCAN propose une analyse intégrée.

L'interfaçage de MADSCAN avec la base de données BASE offre un gain de temps, en limitant l'intervention de l'utilisateur. Les traitements sont directement réalisés sur les données primaires stockées dans les tables de la base de données. De même, les résultats y sont automatiquement sauvegardés. De plus, BASE dispose de nombreux outils graphiques qui complètent les représentations graphiques proposées par MADSCAN.

MADSCAN, version *plug-in,* a été proposé à la communauté BASE et sera intégré à un article dédié aux plug-ins développés pour BASE. De ce fait, les codes sources de MADSCAN (version *plug-in* BASE) sont disponibles sous la licence GPL. MADSCAN peut dès lors être intégré à d'autres applications pour le traitement et l'analyse des données de puces à ADN telle que M@IA (Badlou, 2004).

3.3 Optimisation et futurs développements

La première amélioration concerne les **codes sources de MADSCAN**. Ils peuvent certainement être optimisés, notamment en terme de temps de calcul pour certains des algorithmes comme le choix du mode de normalisation spatial. Les **outils de visualisation**, *i.e.* la représentation graphiques des résultats, doivent également être améliorés. Des graphiques interactifs seraient particulièrement intéressants.

De même, de **nouvelles fonctionnalités** peuvent être ajoutées. Une version du module *limma* (Smyth, 2004) pour la **mise en évidence de gènes différentiels** (*cf.* p. 70-82), disponible sur le site de BioConductor, a déjà été implémentée. Les algorithmes de MADSCAN pourraient également être généralisés à l'analyse d'**autres types de bio-puces** : puces à ADN mono-couleurs (type Affymetrix®, Applied Biosystems® ou Codelink® (Amersham, *Inc.*)), chIP-Chip, puces à protéines…Nous envisageons dans un premier temps de développer une interface pour certaines fonctions de la librairie *affy (Bolstad et al., 2003; Irizarry et al., 2003),* disponible sur le site du projet BioConductor.

Enfin, un rendu des **résultats sous un format d'échange de type XML** (*eXtensible Markup Language*) permettrait un transfert plus direct des données « consolidées » vers d'autres applications comme des bases de données ou d'autres outils d'analyse.

Chapitre III.

Analyse Bibliographique : Extraction de Connaissances à partir des Données d'Expression de Puces à ADN

Chapitre III. Analyse bibliographique : Extraction de connaissances à partir des données d'expression de puces à ADN

La finalité des expériences de puces à ADN est la compréhension des mécanismes de régulation de l'expression des gènes. Dans ce but, les expériences sont construites pour répondre, entre autres, aux questions suivantes : Quels sont les gènes différentiellement exprimés entre les échantillons étudiés ou les traitements testés ? Dans le cas de différences significatives, qu'est ce qui distingue les échantillons ? Existe-il des regroupements selon les échantillons, les gènes ? Pouvons nous prédire des regroupements ? Quelles sont les catégories fonctionnelles, les mécanismes impliqués ? Autant de questions qui nécessitent l'utilisation de différentes méthodes d'analyse et techniques de fouilles de données (*data mining*). Dans ce chapitre seront présentées, de façon non exhaustive, les techniques les plus employées **pour la mise en évidence des gènes différentiels, les méthodes de regroupements et de classification, l'annotation des gènes et les approches de biologie intégrative.**

Nota bene : pour plus de clarté, le terme *gene reporter,* utilisé jusqu'à présent pour qualifier les fragments de gènes étudiés (sondes), est remplacé par le terme gène

I. Mise en évidence des gènes différentiellement exprimés

L'un des objectifs d'une expérience menée au moyen des puces à ADN est de mettre en évidence des gènes différentiellement exprimés, par exemple, entre 2 groupes d'individus. Intuitivement, mesurer cette différence revient à estimer l'amplitude des variations (*fold change*) entre les 2 groupes (différence ou rapport) et à évaluer si ce résultat est significativement supérieur à un seuil fixé (Schena *et al.*, 1996; DeRisi *et al.*, 1997). Cependant, cette approche est insuffisante car elle ne prend pas en compte les erreurs sur la mesure. En effet, les différences observées correspondent non seulement aux différences biologiques mais aussi aux erreurs de mesure, dues au hasard ou inhérentes à l'expérience (Nadon et Shoemaker, 2002).

Tableau 5. Abrégé pour les tests d'hypothèses

Ho (hypothèse nulle) : il n'existe pas de différence significative au risque α de se tromper
H1 (hypothèse alternative) : il existe une différence significative

α Erreur de type I ou risque de première espèce.
Probabilité d'obtenir des faux positifs, *i.e.* dire qu'un gène est différentiellement exprimé alors qu'il n'est pas différentiellement exprimé.

β Erreur de type II ou risque de seconde espèce.
Probabilité d'obtenir des faux négatifs, *i.e.* dire qu'un gène n'est pas différentiellement exprimé alors qu'il l'est.

Confiance (*confidence*) **1- α**
Probabilité de ne pas obtenir un résultat « faux positif », *i.e.* probabilité de conclure qu'un gène n'est pas différentiellement exprimé lorsqu'il ne l'est pas en réalité.

Puissance (*power*) **1- β**
Probabilité de ne pas obtenir un résultat « faux négatif », i.e. probabilité de conclure qu'un gène est différentielle ment exprimé lorsqu'il l'est véritablement.

***p*-value**
Probabilité *p* d'obtenir une valeur de ratio plus surprenante que celle observée, si l'hypothèse nulle est vraie. Plus la *p*-value est petite, moins les variations observée sont dues au hasard.

***q*-value**
Probabilité *q* d'obtenir des faux positifs quand l'hypothèse nulle est vraie. Moins la q value est grande, moins le risque d'obtenir des faux positifs est élevés.

Résultats possibles d'un test d'hypothèses.

Hypothèses	Acceptée	Rejetée
H_0	1-α	α
H_1	β	1-β

La difficulté de faire la distinction entre les variations biologiques et les variations expérimentales a deux origines. D'une part, les variations biologiques et expérimentales peuvent être amalgamées (*confounded*). D'autre part, les matrices de données d'expression sont, le plus souvent, dissymétriques ; le nombre de variables (gènes) excédant toujours celui des d'échantillons. Certains auteurs parlent même de la « malédiction de la dimension » (« *Curse of dimensionality »)* (Draghici 2003; Butcher *et al.*, 2004). Aussi, de nombreuses approches statistiques spécifiques aux matrices de données d'expression ont été développées pour prendre en compte et/ou évaluer les différents types d'erreur sur les mesures et répondre au problème de dimension des matrices de données.

1. Concepts statistiques

En premier lieu, pour une meilleure compréhension des techniques de mise en évidence des gènes différentiellement exprimés, il nous paraît nécessaire de définir le concept d'inférence statistique et les tests d'hypothèses (Tab. 5).

1.1 Inférence statistique

Le concept d'inférence statistique est la base des tests statistiques classiques et des méthodes scientifiques (Stekel 2003). L'inférence statistique est le fait d'**extrapoler** (généraliser) **les résultats obtenus pour un échantillon d'individus à l'ensemble de la population dont il est issu**. En effet, l'objectif de nos études est de décrire le comportement des gènes dans une population. Or, pratiquement nous travaillons sur des échantillons de taille réduite que nous espérons représentatif de la population.

1.2 Tests d'hypothèses et *p-value*

D'un point de vue statistique, pour mettre en évidence des gènes différentiellement exprimés, il faut réaliser un test d'hypothèses (synthèse par Nadon et Shoemaker (2002)). Un test d'hypothèses est la vérification de la validité de l'hypothèse nulle sur les données observées grâce à un modèle probabiliste (Tab.5).

L'hypothèse nulle est qu'il n'existe pas de différence (pas d'effet biologique) significative au risque α de se tromper. Dans ce cas, les variations observées sont liées à l'erreur de mesure. Soit la comparaison de 100 gènes dans 2 populations différentes et le seuil de signification α=5%, l'hypothèse H_0 est que l'expression de ces gènes ne soient pas significativement différents avec une **confiance 1- α** = 95%. La probabilité de dire que les

gènes sont différents (rejeter H_0) alors qu'ils ne le sont pas est α, *i.e.* sur les 100 gènes testés 5 peuvent être des faux positifs. La capacité de mettre en évidence un gène différentiellement exprimé est la **puissance du test ou 1-β**.

Pour mettre en évidence des gènes différentiels, nous calculons, sous l'hypothèse nulle, la **probabilité *p* d'obtenir une valeur de ratio plus « extrême » que celle observée**. Si cette probabilité, encore appelée *p-value*, est très petite, l'événement est « surprenant » et Ho est rejetée. Par exemple, si la *p-value*=0,01 cela signifie que nous avons 1% de chance d'observer ce phénomène (un gène différentiel) au hasard. Le niveau d'expression d'un gène est déclaré significativement différentiel, lorsque la *p-value* est inférieure à un risque *α* fixé. Le **risque *α*** est également qualifié d'**erreur de type I** (ou risque de première espèce) et représente la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie (Tab. 5). Ce type d'erreur correspond au fait de dire que des gènes sont différentiellement exprimés alors qu'ils ne le sont pas. Ce sont des **faux positifs**.

L'un des **enjeux des tests statistiques** est de **minimiser le nombre de faux positifs**, *i.e.* l'erreur de type I. Cet objectif est particulièrement vrai dans l'analyse des matrices de données d'expression. En effet, chaque gène fait l'objet d'un test statistique avec un risque *α* d'avoir des faux positifs. Or, à chaque test réalisé le risque *α* se multiplie et le risque d'obtenir des faux positifs augmente (Tab. 6). Une attention toute particulière doit donc être apportée au contrôle de ce type d'erreur (Reiner *et al.*, 2003). Il existe deux grandes catégories de contrôle de l'erreur de type I : les approches FWER (*Family Wise Error Rate*) et FDR (*False Discovery Rate*). Le contrôle FWER mesure la probabilité *p* de faire une ou plusieurs erreurs de type I parmi l'ensemble des hypothèses testées ou *false positive rate*. L'approche FDR estime la **proportion *q* des erreurs de type I parmi les hypothèses rejetées** (*i.e.* les gènes estimés différentiellement exprimés). Par analogie avec la *p-value*, le résultat du contrôle FDR est parfois appelé *q-value* (Storey et Tibshirani, 2003a).

L'autre objectif des tests statistiques est de **minimiser le nombre de faux négatifs**, *i.e.* conserver une puissance statistique suffisante. La puissance d'un test statistique (*1-β*) étant sa capacité à détecter un gène différentiel et qu'il le soit vraiment (Tab. 5). Les statisticiens qualifient les faux positifs d'**erreur de type II**. L'erreur de type II ne peut pas être contrôlée explicitement mais doit être contrôlée implicitement *via* le plan expérimental. Plus précisement, la **puissance d'une expérience dépend du nombre de *replicates* biologiques** utilisés. Plus le nombre de replicates biologiques est impotant, plus la puissance de l'expérience augmente.

Tableau 6. Augmentation du nombre d'erreurs de type I avec le nombre de variables (d'après Drăghici , 2003). Si aucune correction n'est appliquée pour prendre en compte les comparaisons multiples, le nombre de faux positifs augmente avec le nombre de gènes testés. Ainsi, pour 10 gènes analysés, au risque α=0.05, moins d'un gène sera identifié comme différentiel alors qu'il ne l'est pas ; tandis que pour 10.000 gènes testés, 500 risquent d'être faussement qualifiés de différentiels.

Nombre de gènes	**Seuil de signification par gène**			
	0.01	0.05	0.1	0.15
10	<1	<1	1	1.5
20	<1	1	2	3
50	<1	2.5	5	7.5
100	1	5	10	15
500	5	25	50	75
1000	10	50	100	150
5000	50	250	500	750
10000	100	500	1000	1500

Finalement, pour tout test d'hypothèses, il faut faire un choix pour équilibrer le nombre de faux positifs par rapport au nombre de faux négatifs, *i.e.* entre le seuil de signification et la puissance du test. Un seuil de signification élevé augmente la confiance dans les résultats du test statistique mais réduit la puissance. Inversement, un seuil de signification peu strict diminue la confiance mais augmente la puissance. En réalité, il y n'a pas de choix meilleur qu'un autre, tout dépend de la question biologique posée. Par exemple, si l'objectif de l'expérience est de mettre en évidence de nouvelles cibles thérapeutiques, et qu'une grande somme d'argent est mise en jeu pour chaque cible, il peut être préférable de minimiser le nombre de faux positifs. Un faux positifs aura pour conséquence une perte d'argent considérable. En revanche, si les puces à ADN sont utilisées pour le diagnostic d'une maladie mortelle (*e.g.* tumeur maligne), il est plus important de ne pas avoir de faux négatifs. Un faux négatif est un patient qui mourra alors qu'il aurait pu être soigné si la maladie avait été mise en évidence à temps.

1.3 Correction pour tests multiples

Afin de prendre en considération le risque d'augmenter le nombre de faux positifs par la multiplication des comparaisons, les résultats des tests statistiques doivent être corrigés. Le mode de contrôle de l'erreur de type I détermine le type de correction à appliquer. Les corrections de Bonferroni, Šidák, Holm ou Hochberg ajustent les *p-values* pour

le contrôle FWER, tandis que les tests de Benjamini-Hochberg et Benjamini-Yekutieli ajustent les *p-values* pour le contrôle FDR (synthèses par Dudoit *et al.* (2002) et Reiner *et al.* (2003)). De plus, parmi ces méthodes de contrôle de l'erreur de type I, il existe trois modes d'ajustements : *single step*, *step down* et *step up* (Callow *et al.*, 2000). Le tableau 7 résume les résultats d'une étude comparative, réalisée au sein du laboratoire (Mainard, 2003), sur les différentes méthodes et modes d'ajustements les plus employés pour corriger les résultats des tests statistiques.

Les méthodes *single-step* sont des ajustements de *p-values* qui ne prennent pas en compte l'ordre des résultats des tests statistiques non corrigés. Chaque hypothèse est évaluée à une valeur critique, indépendante des résultats des autres tests statistiques. La correction la plus connue est le contrôle du FWER selon Bonferroni. Le test statistique ainsi corrigé est très conservateur : il rejette peu d'hypothèses nulles, et par conséquent peu de gènes seront estimés comme différentiellement exprimés.

Les ajustements *step-down* et *step-up*, sont des améliorations des méthodes *single-step*. La méthode *step-down* fait correspondre à chaque hypothèse (test) successivement, le test statistique le plus significatif, *i.e.* la *p-value* la plus petite. A l'inverse, la méthode *step-up* fait correspondre successivement à chaque hypothèse, le test statistique le moins significatif, *i.e.* la p-value la plus importante. Les tests de Holm et Hochberg sont respectivement des corrections FWER *step-down* et *step-up*. Moins strictes que les approches *single-step*, ces méthodes détectent toutefois moins de gènes différentiels lorsque le nombre de gènes testés augmente. De plus, dans les expériences de puces à ADN, les gènes testés (par conséquent les tests d'hypothèses) ne sont pas indépendants. Or, ces approches ignorent la dépendance des données (Holm), voire nécessitent leur indépendance (Hochberg) (Dudoit *et al.*, 2002; Reiner *et al.*, 2003). Pour répondre à ce problème, Dudoit *et al.* (2002) proposent une méthode non paramétrique basée sur des permutations. Cette approche permet de prendre en compte la dépendance entre les données mais reste très conservatrice et demande un grand nombre d'échantillons ainsi qu'une grande puissance de calcul.

Si dans certain cas le contrôle de FWER est utile, les corrections apportées entraînent généralement une perte de la puissance des tests statistiques, *i.e.* la capacité de mettre en évidence les gènes différentiellement exprimés. Par conséquent, il est sans doute plus intéressant de contrôler le taux de faux positifs parmi les hypothèses rejetées (FDR) que de minimiser le taux de faux positifs parmi l'ensemble des tests effectués (FWER). En effet, un certain nombre de faux positifs est acceptable tant que celui-ci reste relativement faible par

rapport aux nombres hypothèses rejetées, *i.e.* 5% de gènes faussement détectés différentiels parmi 100 gènes différentiels est préférable à 5% parmi 20.

Le contrôle de l'erreur de type I par FDR a été suggéré par Benjamini et Hochberg en 1995. Leur méthode est globalement plus puissante et moins conservatrice que les approches de contrôle de l'erreur de type I par FWER. Cependant, elle nécessite l'indépendance des données. De nombreuses adaptations ont été proposées et comparées aux résultats du contrôle par FWER (Tusher *et al.*, 2001; Dudoit *et al.*, 2002; Reiner *et al.*, 2003; Mainard, 2003). Les méthodes FDR par permutation semblent être les plus pertinentes (Reiner *et al.*, 2003). Elles permettent également un meilleur équilibre entre le nombre de faux positifs et les faux négatifs (Storey et Tibshirani, 2003a).

Tableau 7. Méthodes d'ajustement de l'erreur de type I des tests statistiques dans le cas de comparaisons multiples (d'après Mainard, 2003)**.**

Contrôle	Ajustement	Méthode	Contrôle[1]	Relations entre les données[2]
FWER	Bonferroni	Single step	5	*
	Šidák SS	Single step	4	-
	Šidák SD	Step down		-
	Holm	Step down	3	*
	Hochberg	Step up	3	I
	Westfall & Young (MinP)	Step down	4	D
	Westfall & Young (MaxP)	Step up	4	D
FDR	Benjamini-Hocheberg	Step up	2	I
	Benjamini-Yekutieli	Step up	3	-
	Efron	Single step	1	-
	Tusher	Single step	3	D

1) « Force » du contrôle de l'erreur de type I :

Echelle de 1 à 5 allant de l'ajustement le moins conservateur au plus conservateur.

2) Relations entre les jeux de données :

- : inconnu

D : prise en compte de la dépendance entre les données

I : nécessite l'indépendance la dépendance entre les données

* : ignore la dépendance entre les données

En bref

L'hypothèse nulle (H_0) est qu'il n'existe pas de différence (pas d'effet biologique) significative entre 2 ou plusieurs groupes au risque α (erreur de type I) de se tromper. Le résultat du test d'hypothèse est la probabilité p (*p-value*) d'obtenir une valeur plus « surprenante » que celle observée. Moins la *p-value* est grande, moins le phénomène observé (un gène différentiel) est lié au hasard.

Les résultats des tests statistiques multiples doivent être corrigés pour minimiser le nombre de faux positifs. Les modes de contrôle de l'erreur de type I (α) sont :

- FWER (*Family Wise Error Rate*) ou mesure de la probabilité p de faire une ou plusieurs erreurs de type I parmi l'ensemble des hypothèses testées (gènes).
- FDR (*False Discovery Rate*) ou estimation de la proportion q des erreurs de type I parmi les hypothèses rejetées, *i.e.* les gènes estimés comme différentiellement exprimés.

Les méthodes FDR sont généralement plus puissantes et moins conservatrices que les approches FWER.

2. Tests statistiques

Comme souligné précédemment (*cf.* p.59), la mise en évidence des gènes « différentiels » par la simple analyse descriptive des amplitudes de variations d'expression (*fold change)* est insuffisante. Des approches statistiques sont nécessaires afin d'estimer et de distinguer la variabilité intra et inter-groupes. De nombreux tests statistiques ont ainsi été proposés allant du test *t* de Welch (lorsque les groupes ont des variances inégales) aux approches bayésiennes (Efron *et al.*, 2001; Lönnstedt et Speed, 2002) en passant par les analyses de variance (Kerr *et al.*, 2000).

L'application des tests dépend de plusieurs paramètres. Tout d'abord, il faut savoir si les données analysées sont indépendantes (Golub *et al.*, 1999), appariées (Perou *et al.*, 2000) ou multi-variées (Khan *et al.*, 2001). Ensuite, le mode de distribution des données doit être évalué : distribution gaussienne ou pas. En effet, les tests paramétriques tels que les tests *t* supposent une distribution normale des données. A l'inverse, les tests non paramétriques sont moins sensibles au mode de distribution des données et aux valeurs atypiques. Enfin, la variance intra et inter-groupe doit être estimée. Les tests paramétriques sont moins sensibles à

un écart à la normalité qu'à une mauvaise estimation de l'homogénéité des variances. Il est donc admis, sous condition d'un bon estimateur de la variance, de réaliser un test paramétrique même si le mode de distribution des données s'écarte légèrement de la normalité.

2.1 Tests paramétriques classiques ou tests *t*

Le test de Student ou test *t* est la méthode la plus couramment utilisée pour évaluer si la différence observée entre 2 échantillons est significative. Deux versions du test *t* existent suivant l'indépendance (*unpaired*) ou l'appariement (*paired*) des échantillons.

Le test *t* pour données appariées (*one sample t-test*) s'applique, par exemple, à l'**analyse des prélèvements d'un même patient avant et après un traitement thérapeutique**. Comme dans les travaux de Le Bouter *et al.* (2004), supposons une population d'individus atteints d'arythmie cardiaque dont des échantillons sont analysés avant et après traitement à l'amiodarone. Si M est le ratio (en log2) des niveaux d'expression d'un patient avant et après traitement (Eq.1), M_{moy} est la moyenne des ratios des niveaux d'expression de la population et σ_M est l'écart type des ratios M. L'équation 2 présente le calcul de la statistique *t*.

$$M \equiv \log_2(Cy5_{Avant}) - \log_2(Cy3_{Après}) \qquad \text{Equation 1}$$

$$t \equiv \frac{M_{moy.}}{\sigma_M / \sqrt{N}} \qquad \text{Equation 2.}$$

où N est le nombre d'individus testés

Une *p-value* est calculée à partir de la valeur *t* calculée en la comparant à la distribution du test t de Student pour un degré de liberté *N-1*. Le degré de liberté (*ddl*) est le nombre de variables indépendantes dans l'analyse ; dans ce cas le nombre d'individus moins un.

Le test *t* pour données non appariées ou « test *t* 2 classes » permet la comparaison de deux populations différentes. Par exemple, une population d'adolescents atteints de scoliose peut être comparée à une population de patients plus âgés pour mettre entre évidence l'évolution de la pathologie au niveau transcriptomal. Pratiquement, le « test *t* 2 classes » est

très similaire au test *t* pour des données appariées. La principale différence est l'estimation des variances des populations. Lorsque les variances des deux populations sont identiques, la moyenne des variances peut être utilisée. Dans les expériences de puces à ADN, elles sont généralement différentes et doivent être calculées pour chacune des populations (test *t* de Welch).

D'après l'équation 2, le *t* calculé peut être anormalement élevé (et la *p-value* correspondante anormalement petite) si le M_{moy} est très grand ou si σ_M est très petit. Bien qu'un test *t* puisse donner des résultats satisfaisants pour un nombre d'échantillons *N* important, les résultats sont peu fiables pour un petit nombre de mesures répétées. De plus, l'application du test *t* requiert un jeu de données distribuées selon la loi normale. Dans le cas de données appariées cela signifie que les ratios des niveaux d'expression « avant » et « après traitement », de chaque gène testé, doivent être distribués normalement. Pour le « test *t* 2 classes », les deux jeux de données pour un même gène doivent suivre une loi normale. Aussi, l'utilisation d'un test *t «* classique » est rarement applicable aux données de puces à ADN compte tenu des contraintes d'application des tests paramétriques, auxquelles s'ajoutent les problèmes de dimensions des matrices (Pan, 2002).

2.2 Tests non paramétriques

Les tests non paramétriques sont généralement robustes face aux données bruitées comme dans le cas des données de puces à ADN. Deux grandes catégories de tests non paramétriques existent :

(i) tests dits « classiques », peu différents des tests paramétriques,

(ii) tests par ré-échantillonnages ou permutations aléatoires, plus récents.

a) Tests non paramétriques « classiques »

Parmi les tests non paramétriques « classiques », l'équivalent du test *t* sur des données appariées est **le test des signes de Wilcoxon**. Ce test ordonne les ratios en fonction de leur amplitude de variation et attribue un rang à chacune des valeurs. La somme des rangs pour les valeurs positives des ratios est ensuite calculée et comparée à la table de Wilcoxon pour en déduire la *p-value*. Le **test de Mann Withney** (ou test de la somme des rangs de Mann Withney) est, quant à lui, l'équivalent non paramétrique du test *t* sur les données non appariées. Les données des deux groupes sont combinées en un seul classement et classées dans un ordre croissant. Les rangs d'un groupe sont comparés à la distance des rangs de

l'autre groupe. Cette valeur est alors comparée à la table de Mann-Withney pour obtenir une *p-value*.

Ces tests possèdent l'avantage d'être applicables à des données dont la distribution ne suit pas la loi normale. Toutefois, le test de Wilcoxon requiert une distribution symétrique des données. De plus, ces tests sont globalement moins puissants que leur équivalent paramétrique ou que les méthodes par ré-échantillonnage aléatoire (Thomas *et al.*, 2001).

b) Analyses par ré-échantillonnage aléatoire

L'objectif des analyses par ré-échantillonnage aléatoire (*bootstrapping*) est de comparer un jeu de données observées expérimentalement à des jeux de données empiriques (générées) afin de **déterminer si la distribution des données observées est liée au hasard**. L'outil le plus utilisé actuellement pour ce type d'analyse est SAM, ou ***S**ignificance **A**nalysis of **M**icroarrays* de Tusher *et al.* (2001).

Afin d'illustrer le principe des analyses par permutation, la figure 22 présente l'exemple d'une analyse par SAM d'un jeu données non appariées. Les données expérimentales sont extraites de l'analyse de 2 groupes de patients atteints de pathologies cardiaques différentes, notés AF et AC pour fibrillation auriculaire et valvulopathie (Steenman *et al.*, soumis).

Les jeux de données empiriques sont créés à partir des données d'origine par permutations aléatoires des individus entre les différents groupes (Fig. 22A-B). La base de comparaison des distributions est généralement le test *t* car il est associé à l'amplitude des variations (*fold change*) et au nombre d'individus de l'expérience. Les statistiques *t* des données réelles (observées) sont comparées à la distribution des statistiques *t* des données permutées au hasard (attendues) (Fig. 22C). La *p-value* est calculée à partir de la distribution des statistiques *t* calculées sur les données permutées. Si la distribution des statistiques *t* des données réelles est identique à celle des données attendues, la différence entre les jeu de données est du au hasard. Inversement, un écart entre les distributions témoigne d'une différence significative entre les groupes (Fig. 22C).

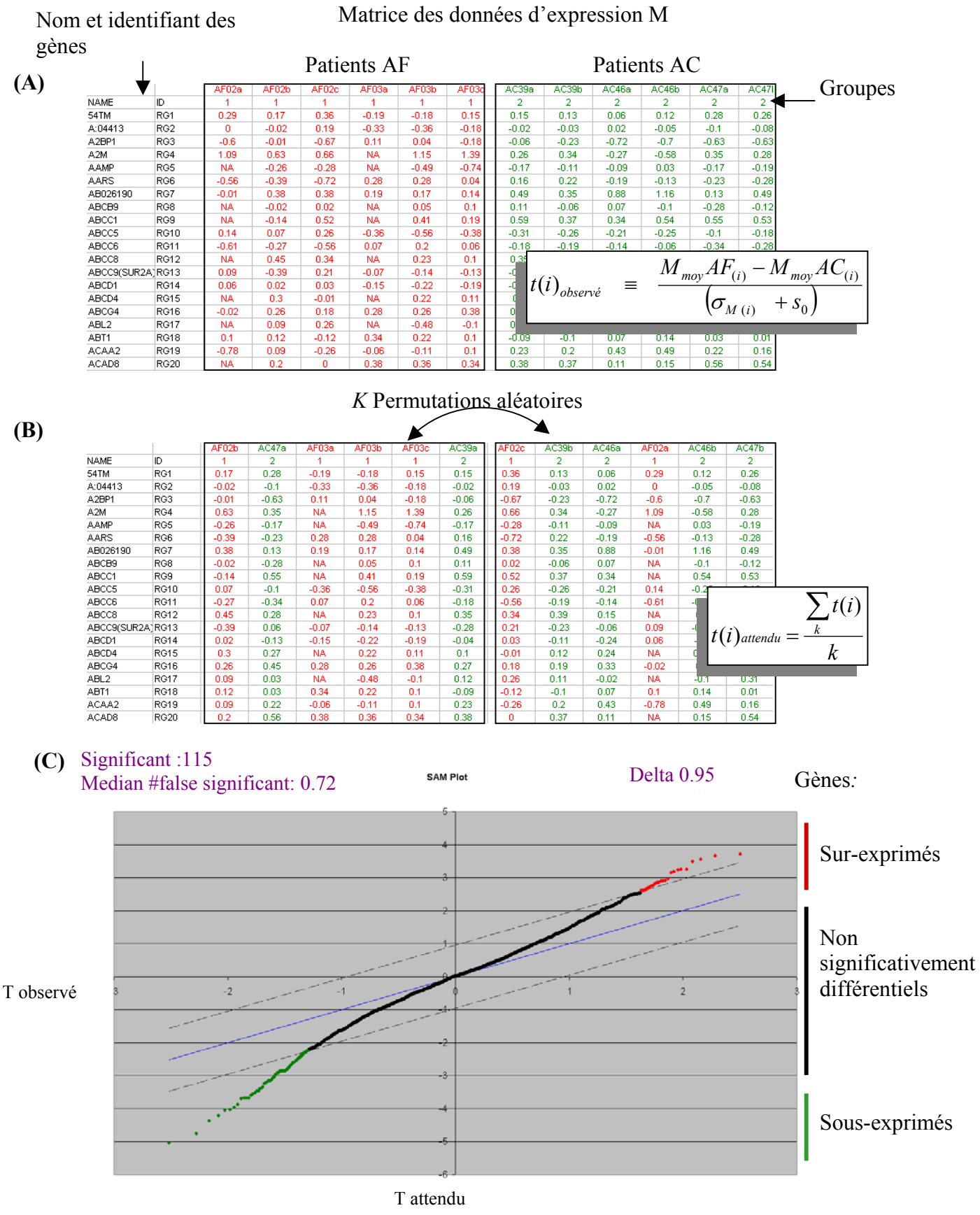


Figure 22. Recherche de gènes différentiellement exprimés, exemple de l'outil SAM pour Excel® (Tusher *et al.*, 2001)

Figure 22. Recherche de gènes différentiellement exprimés grâce à SAM pour Excel® (Tusher *et al.*, 2001).
(A) Matrice de données d'expression à analyser : chaque ligne représente les données pour un gènes *i* à savoir son nom (*NAME*), son identifiant (*ID*) et ses valeurs d'expression dans les différents échantillons (en colonne). Chaque groupe à analyser se constitue de 6 individus. Certaines données d'expression sont manquantes et identifiées par « *NA* ». Une valeur $t(i)_{observé}$ est calculée pour chaque gènes *i* observé. $M_{moy}.AF_{(i)}$ *et* $M_{moy}.AC_{(i)}$ sont respectivement les ratios moyens des niveaux d'expression du gène *i* chez les individus AF et AC. $\sigma_{M\,(i)}$ est l'écart type moyen de gène *i.* s_0 est la constante de correction du test *t.*
(B) Permutations aléatoires des colonnes au sein du jeu de données : chaque colonne est permutée *k* fois au hasard. Pour chaque configuration, une valeur $t(i)_{empirique}$ est estimée. La somme des $t(i)_{empirique}$ sur le nombre *k* de permutations appliquées est le $t(i)_{attendu}$ qui sera comparé au $t(i)_{observé}$ de la population.
(C) Q-Qplot et mise en évidence des gènes d'intérêt : chaque point représente les $t(i)_{observé}$ *versus* $t(i)_{attendu}$ d'un *gene reporter i*. Les points à l'extérieur de l'intervalle *delta* (droites en pointillé), soit ici 115 valeurs, sont estimés comme significativement différentiellement exprimés pour un nombre de faux positifs tolérés, ici moins de 1%. Les gènes représentés par des points rouges sont sur-exprimés dans le groupe 1 par rapport au groupe 2 et, inversement, les gènes représentés par des points verts sont sous-exprimés. Les points noirs indiquent les gènes qui ne présentent pas de différence significative entre les 2 groupes.

Afin de répondre au problème de dimension des matrices d'expression, Tusher *et al.* (2001) proposent un test *t* modifié et une estimation du taux de faux positifs (FDR) par permutation. Le test *t* utilisé est modifié de telle sorte que le dénominateur ne devienne par trop petit pour les faibles valeurs de $\sigma_{M,}$ réduisant ainsi le taux de faux positifs (Eq. 3). En effet, de nombreuses études ont montrées qu'un test *t* modifié diminue le taux de faux positifs et faux négatifs par rapport à un test *t* standard (Tusher *et al.*, 2001; Efron *et al.*, 2001; Lönnstedt et Speed, 2002; Smyth, G. K., Yang, Y. H., et Speed, T., 2003; Broberg, 2003).

$$t(i) \equiv \frac{M_{moy}AF_{(i)} - M_{moy}AC_{(i)}}{\left(\sigma_{M(i)} + s_0\right)} \qquad \text{Equation 3.}$$

où $\sigma_{M(i)} \equiv \sqrt{\frac{s^2_{AF_{(i)}}}{n_{AF}} + \frac{s^2_{AC_{(i)}}}{n_{AC}}}$

- t(i) est la valeur du test t calculé pour le gène *i,*
- $S_{AF(i)}$ et $S_{AC(i)}$ sont respectivement les écarts types des échantillons AF et AC pour le gène *i*,
- $M_{moy}.AF_{(i)}$ *et* $M_{moy}.AC_{(i)}$ sont respectivement les ratios moyens des niveaux d'expression du gène *i* chez les individus AF et AC,
- s_0 est la constante de correction du test *t*

L'estimation de s_0 peut être basée sur différentes approches. Tusher *et al.* (2001) calcule s_0 afin de minimiser le coefficient de variation des valeurs absolues du test *t*. Efron *et al.* (2001) utilisent, quant à eux, le 90ème percentile de la distribution des écarts types des échantillons.

Les analyses par ré-échantillonage font encore l'objet de nombreuses recherches (Tusher et al., 2001; Kerr et Churchill, 2001a; Dudoit et al., 2002; Datta et al., 2004). Ces techniques sont applicables à l'ensemble des dessins expérimentaux, *i.e.* données appariées ou non et analyses multi-variées. Tout comme les tests non paramétriques « classiques », elles ne nécessitent pas une distribution normale des données. Elles sont par conséquent robustes aux valeurs atypiques et aux données bruitées. Toutefois, ces approches, requièrent un certain nombre de répétitions et nécessitent de la ressource informatique.

Tableau 8. Inférence bayésienne

Le théorème de Bayes vise à calculer les probabilités *a posteriori* d'un évènement A (*posterior*), en fonction des probabilités *a priori* de cet évènement (*prior*). L'évènement A peut être une hypothèse, un modèle ou un paramètre numérique (comme la variance d'un échantillon).

$$P(A|B) \equiv P(A) \times \frac{P(B|A)}{P(B)}$$

Où P(A|B), la probabilité *a posteriori* (*posterior*), probabilité que l'évènement A soit vérifié connaissant les données B.

P(A), probabilité *a priori* (*prior*)

P(B|A)/P(B), vraisemblance (*likelihood)*

Ex :

- Grâce à des connaissances épidémiologiques, un médecin a une idée de la **probabilité de la maladie « M » dans la population, soit p (M)** cette probabilité dite probabilité *a priori.*

- Grâce à ses connaissances en pathologie, il connaît (approximativement) la fréquence d'apparition de chaque symptôme dans la maladie M, soit S_1 la fréquence d'un symptôme donné dans la maladie M, *i.e.* **p(S_1/M) la probabilité d'avoir le symptôme S_1 quand le sujet a la maladie M**.

- Le médecin connaît également la **fréquence du symptôme S_1 dans les autres maladies** possibles, soit **p(S_1)**

L'évaluation de la probabilité *a posteriori* de la maladie M, sachant que le patient a le symptôme S1, est fonction de la probabilité *a priori* de M :

$$P(M|S_1) \equiv P(M) \times \frac{P(S_1|M)}{P(S_1)}$$

2.3 Approche bayésienne

a) Inférence bayésienne

Une alternative à l'inférence statistique classique est l'inférence bayésienne. Le théorème de Bayes vise à calculer les probabilités *a posteriori* d'un événement en fonction des probabilités *a priori* de cet évènement (Tab. 8). *A priori* et *a posteriori* s'entendent par rapport à la connaissance d'une information. Plus précisément, le principe de base du théorème de Bayes est de calculer la probabilité *a posteriori* d'un évènement A, sachant qu'un évènement B s'est produit, en fonction sa probabilité.

b) Mise en évidence de gènes différentiellement exprimés par une approche bayésienne

Dans le théorème de Bayes, l'évènement A, dont nous cherchons la probabilité *a posteriori* (Tab. 8), peut être une hypothèse, un modèle ou un paramètre numérique telle la variance d'un échantillon. Ainsi, afin de palier aux limites des tests *t* classiques (normalité, nombre d'échantillons importants) dans la recherche de gènes différentiellement exprimés, certains auteurs proposent une estimation de la variance des échantillons par le théorème de Bayes (Baldi et Long, 2001; Lönnstedt et Speed, 2002; Smyth, 2004).

Baldi et Long (2001) présentent un test *t* modifié où la variance est estimée grâce au théorème de Bayes. Leur approche est implémentée dans un logiciel nommé Cyber-T disponible sous la forme d'un service Web[34].

Lönnstedt et Speed (2002) suggèrent un modèle mixte combinant une approche linéaire à une estimation de la variance par une méthode bayésienne paramétrique. Ils calculent ainsi la statistique B ou *log posterior odds ratios* qui correspond au ratio entre la probabilité qu'un gène donné soit différentiellement exprimé sur la probabilité que ce gène ne soit pas différentiellement exprimé. Une variante de cette solution, adaptée à la problématique des puces à ADN, a été développée par Smyth (2004) sous la forme d'une librairie R nommée *limma* (disponible sur le site du projet Bioconductor[35]).

Les méthodes d'estimation de la variance selon Lönnstedt et Speed (2002) et Smyth (2004) sont de bonnes alternatives à la pondération des variances. Tout d'abord, elles ne sont pas basées sur un modèle mathématique ni associées à un mode de distribution. Elles sont par

[34] http://visitor.ics.uci.edu/genex/cybert/
[35] http://bioconductor.org

conséquent plus puissantes et moins conservatrices que les méthodes par pondérations des variances selon Efron *et al.* (2001) ou Tusher *et al.* (2001). De plus, la méthode proposée par Smyth (2004) peut être généralisée à la comparaison de données multi-factorielles (*e.g.* effet de la maladie croisée avec l'effet de la souche de souris et son âge). Enfin, la principale différence entre les approches de Tusher *et al.* (2001) et Smyth (2004) est le mode d'estimation du nombre de faux positifs. Tusher *et al.* (2001) estime le nombre de faux positif en calculant la *q-value* directement à partir des permutations tandis que Smyth (2004) la calcule à partir de la *p-value* (Storey et Tibshirani, 2003b).

2.4 ANOVA

Dans une expérience de puces à ADN, lorsque de multiples facteurs (age, sexe..) sont à analyser, le test *t* et ses variantes ne suffisent généralement pas à l'interprétation. Un modèle plus complexe doit être construit et une analyse de variance, ou ANOVA (***AN****alyse* ***Of*** ***VA****riance*), peut être utilisée pour mettre en évidence l'impact de chaque facteur. En effet, une ANOVA permet d'évaluer si les moyennes de un ou plusieurs groupes d'échantillons sont significativement différentes et si un ou plusieurs facteurs affectent les mesures (synthèse par Draghici (2003)).

Comme pour les approches par test *t* classiques, l'analyse par ANOVA sous-entend une distribution normale des données. Aussi, de nombreuses techniques (paramétriques ou non) basées sur l'ANOVA ont été développées pour analyser les matrices de données d'expression (Kerr *et al.*, 2000; Ideker *et al.*, 2000; Thomas *et al.*, 2001; Draghici *et al.*, 2003b; Smyth, 2004; Datta *et al.*, 2004). Kerr *et al.* (2000) utilise une ANOVA pour modéliser les données et un ré-échantillonnage pour estimer la *p-value*. Une approche similaire est proposée par Park *et al.* (Park *et al.*, 2003) pour étudier les gènes différentiellement exprimés au cours d'une cinétique. Dans SAM, Tusher *et al.* (2001) proposent également une ANOVA à 1 facteur ou *One-Way ANOVA* pour comparer plusieurs groupes d'échantillons simultanément. Ideker *et al.* (2000) suggère le calcul des paramètres de l'ANOVA par la méthode du maximum de vraisemblance ou *maximun likelihood*. Enfin, Symth (2004) présente l'analyse de données multi-factorielles par son approche bayésienne.

L'ANOVA et ses variantes sont des techniques d'analyse puissantes qui permettent de mettre en évidences les différences entre plusieurs groupes de facteurs. Toutefois, ces approches requièrent un plan expérimental bien construit (Kerr et Churchill, 2001b), un grand

nombre de mesures (Pavlidis, 2003; Draghici *et al.*, 2003b) et une certaine expertise pour la construction du modèle. Enfin, une ANOVA est applicable si les facteurs dont dépend le niveau d'expression des gènes sont des variables discrètes telles le sexe, le type de maladie. Dans le cas de variables continues, comme une dose médicamenteuse, et si les valeurs d'expression répondent de manière linéaire à ce facteur, des modèles linéaires généralisés doivent être appliqués (Stekel 2003). Ces modèles sont une généralisation des ANOVA et des régressions linéaires.

3. Outils

La plupart des approches présentées ci-dessus proposent les algorithmes et/ou les logiciels associés. Parmi les plus employés dans le domaine académique, nous avons cité MAANOVA de Kerr *et al.* (2000), SAM de Tusher *et al.* (2001) et *limma* de Symth (2004). SAM est actuellement l'outil le plus utilisé. D'abord implémenté sous la forme d'une macro complémentaire Excel®, l'algorithme est désormais disponible dans le module *siggene* du projet BioConductor. Le module *limma*, également distribué par BioConductor, peut être utilisé en ligne de commande ou *via* une interface graphique (Wettenhall et Smyth, 2004; Smyth, 2004). Dans cette dernière version, l'analyse va du traitement des données primaires de l'ensemble des puces d'une expérience à la recherche des gènes différentiellement exprimés. Enfin, MAANOVA est un ensemble de fonctions pour l'analyse de variance appliquée au puces à ADN (Kerr *et al.*, 2000). Premièrement développé sous Matlab®, l'outil MAANOVA est désormais une application Java™ couplée à un module R.

De nombreux outils commerciaux ont également été spécifiquement développés pour la problématique « puce à ADN » tels ArrayStat® d'Imaging Research Inc., S^{+}ArrayAnalyzer™ d'Insigthful ou encore le module SAS® Microarray. Ils sont généralement plus conviviaux que les outils académiques. Toutefois, ils reprennent le plus souvent l'essentiel des fonctions disponibles dans les projets R et Bioconductor. Une comparaison de ces différents outils (académiques et commerciaux) est disponible sur le site Web *Functional Genomincs* de Y.F. Leung[36].

[36] http://genomicshome.com

Tableau 9. Avantages (vert) et inconvénients (rouge) des différentes approches statistiques pour la mise en évidence des gènes différentiellement exprimés.

Tests *t* « classiques »	Tests non paramétriques « classiques »	Tests avec ré-échantillonnage	Analyses bayésiennes	ANOVA
▪ Simple ▪ Puissant ▪ Implémenté dans de nombre logiciels ▪ Nécessite une distribution normale des données ▪ Très sensible aux valeurs aberrantes	▪ Simple ▪ Robuste ▪ Implémenté dans de nombre logiciels ▪ Peu puissant	▪ Robuste ▪ Puissant ▪ Nécessite une certaine expertise de l'utilisateur ▪ Technique toujours en développement	▪ Puissant ▪ Robuste ▪ Analyses multi-factorielle ▪ Nécessite une certaine expertise de l'utilisateur ▪ Technique toujours en développement	▪ Puissant ▪ Robuste ▪ Analyses multi-factorielle ▪ Nécessite une distribution normale des données ▪ Nécessite une certaine expertise de l'utilisateur

En bref

Des approches statistiques sont nécessaires à la mise en évidence des gènes différentiellement exprimés. Compte tenu de l'aspect bruité des données et des problèmes de dimensions des matrices, les tests statistiques « classiques » ne sont pas adaptés à l'analyse des données de puces à ADN (Tab. 9):

(i) il est nécessaire de corriger le résultat des tests statistiques pour tenir compte des comparaisons multiples,

(ii) les tests non paramétriques sont plus robustes que les approches paramétriques face aux bruits expérimentaux. Les méthodes par permutations sont, actuellement, les approches les plus puissantes.

L'analyse de données multi-factorielles peut nécessiter des approches bayésiennes, des analyses de variances (ANOVA) ou l'application de modèles linéaires généralisés.

II. Méthodes de classification des données d'expression

Suite, ou parallèlement, à la mise en évidence des gènes différentiellement exprimés dans une matrice de données d'expression, il est intéressant de **rechercher et visualiser les éventuels regroupements de gènes et/ou d'échantillons d'après leur profil d'expression.**

Une matrice de données d'expression se compose de *n* gènes et *m* échantillons. Elle représente ainsi un espace de données à *n* points et *m* dimensions (ou inversement). Ces **matrices peuvent être étudiées selon les lignes (gènes) et/ou selon les colonnes (échantillons).** Les algorithmes de classification se sont montrés particulièrement efficaces pour envisager ces deux approches. Ils permettent notamment de répondre aux questions suivantes : Existe-il des regroupements selon les portraits moléculaires des échantillons, selon les profils d'expression des gènes ? Qu'est ce qui distingue ces échantillons, ces gènes ? Pouvons nous prédire des regroupements, des classifications ?

Un grand nombre d'algorithmes de classification existe et de nouvelles approches sont proposées chaque mois. Par conséquent, il est difficile d'être exhaustif et seules quelques approches, parmi les plus communément utilisées pour l'analyse des données d'expression, seront décrites dans les paragraphes suivants.

1. Définitions

Les algorithmes de classification sont définis comme des méthodes de répartition d'un ensemble d'objets (points ou vecteurs) en plusieurs sous-ensembles, sur la base de leurs similarités ou dissimilarités (Gilbert *et al.*, 2000). Le but est de construire des groupes qui minimisent la variabilité intra-groupe tout en maximisant les distances inter-groupes. Plus précisément, ils visent à. trouver l'ensemble des groupes (gènes ou échantillons) dont les membres sont très similaires mais distants des autres membres sur la base de leur profil d'expression. Un grand nombre de mesures de similarités et dissimilarités existent, seules les plus connues seront présentées dans les paragraphes suivants.

Les algorithmes de classification se regroupent en deux grandes catégories: les **approches supervisées (*supervised learning*) et non supervisées (*unsupervised learning*)** (synthèses par Slomin (2002) et Leung et Cavalieri (2003)). Les méthodes non supervisées groupent les objets sans *a priori* (*data driven*). Ces techniques sont dites exploratoires (*exploratory techniques*) et sont essentiellement employées pour la découverte de classes (*class discovery*). A l'inverse, les méthodes supervisées utilisent de la connaissance *a priori*.

Elles établissent des règles et un modèle de classification à partir d'un jeu de données annotées, ou jeu d'apprentissage (*training set*), pour ensuite prédire la classification (*Class prediction*) de nouveaux cas appartenant à un jeu de données test.

2. Formatage des données et mesures de distance

Selon la question biologique posée, il peut être nécessaire de formater les matrices de données d'expression et/ou de choisir une mesure de distance plutôt qu'une autre, avant d'appliquer un algorithme de classification. En effet, les résultats des regroupements dépendent du type et de la qualité des variables employées. De même, le choix des mesures de distance conditionne les résultats des classifications.

2.1 Formatage des matrices d'expression

a) Types de variables

Suivant la question biologique posée, les variables peuvent être les gènes ou les échantillons biologiques. Les données d'expression à classer sont quantitatives ou semi-quantitatives. Par exemple, les mesures d'expression issues de la technologie des puces à ADN « 2 couleurs » sont généralement relatives, *i.e.* semi-quantitatives. Les données d'expression peuvent également être pondérées, par exemple suivant la confiance accordée à la mesure d'un gène ou d'un échantillon (Hughes, 2002). Enfin, il est possible d'ajouter aux valeurs d'expression différents types de variables comme des données cliniques ou des données de la littérature (Chaussabel et Sher, 2002). Dans ce cas, les données de type continue (*e.g.* poids, taille ou âges des patients) doivent généralement être transformées en variables discrètes.

b) Valeurs manquantes

Les matrices de données d'expression peuvent être incomplètes. Pour certains gènes, aucune valeur d'expression n'est attribuée et la donnée est notée « NA » (*Non Attributed* ou *Non Available*). Les données manquantes ont pour origine l'absence d'expression du gène dans l'échantillon ou la mauvaise qualité de la mesure (donnée écartée de l'analyse par les algorithmes de filtration). Ces données sont réparties de manière hétérogène dans les matrices d'expression. Elles peuvent influencer les résultats des calculs de distance ainsi que la stabilité des classifications obtenues. En effet, certaines mesures de distance, telle la distance euclidienne (*cf.* p. 89-90), sont pondérées en fonction de la taille des vecteurs comparés (*i.e.*

profils d'expression) (Oba *et al.*, 2003). De même, les distances inter-groupes sont plus ou moins sensibles aux valeurs manquantes (de Brevern *et al.*, 2004).

Plusieurs stratégies existent pour contrôler l'effet des valeurs manquantes. Par exemple, il est possible de modifier les algorithmes de groupement afin qu'ils prennent en compte cette absence de données et ré-équilibrent la matrice de données d'expression. Certains algorithmes permettent de compléter le jeu de données : moyenne sur le gène, moyenne sur les *k* voisins les plus proches (*K-Nearest Neighbors - KNN*) ou encore décomposition des valeurs d'expression en valeurs propres (Troyanskaya *et al.*, 2001). La méthode KNN est actuellement la méthode la plus simple à mettre en œuvre et semble relativement efficace (avec *k* compris entre 10 et 15 selon les auteurs). Toutefois, elle présente des difficultés à estimer les valeurs manquantes correspondant à des valeurs extrêmes (de Brevern *et al.*, 2004). Aussi, de nouveaux algorithmes basés sur le théorème de Bayes (Oba et al., 2003) ou des régressions locales (Kim *et al.*, 2005) semblent pouvoir répondre à cette limite.

c) Centrer - Réduire

Center et/ou réduire les données suivant les gènes et/ou les échantillons peut parfois être utile. Ces corrections permettent d'améliorer la comparaison des variations entre deux gènes ou deux échantillons par rapport à leurs paramètres de position respectifs (moyenne, médiane, mode).

Par exemple, dans le cas d'un plan expérimental à une référence unique, si les comparaisons sont faites entre les conditions, indépendamment de la référence, il peut être intéressant de **centrer les gènes sur la moyenne ou la médiane. Les ratios reflètent alors la variation des gènes par rapport à la moyenne (médiane) des différentes conditions**. Cette procédure se justifie moins lorsque la référence fait partie de l'expérience comme dans les analyses cinétiques. De même, il n'est pas utile de centrer les gènes lorsqu'il est important de savoir si un gène est sur- ou sous-exprimés par rapport à la référence et si la distance de ce gène aux autres gènes est importante (Fig. 23A-B). En effet, centrer les données tend à diminuer la distance entre les gènes en déplaçant les profils d'expression extrêmes vers le centre. Centrer selon les conditions peut permettre d'éliminer certains biais. De plus, cette transformation a peu d'influence sur la classification des gènes qui est réalisée sur la distance entre les gènes et non leur valeur absolue (Sturn, 2000).

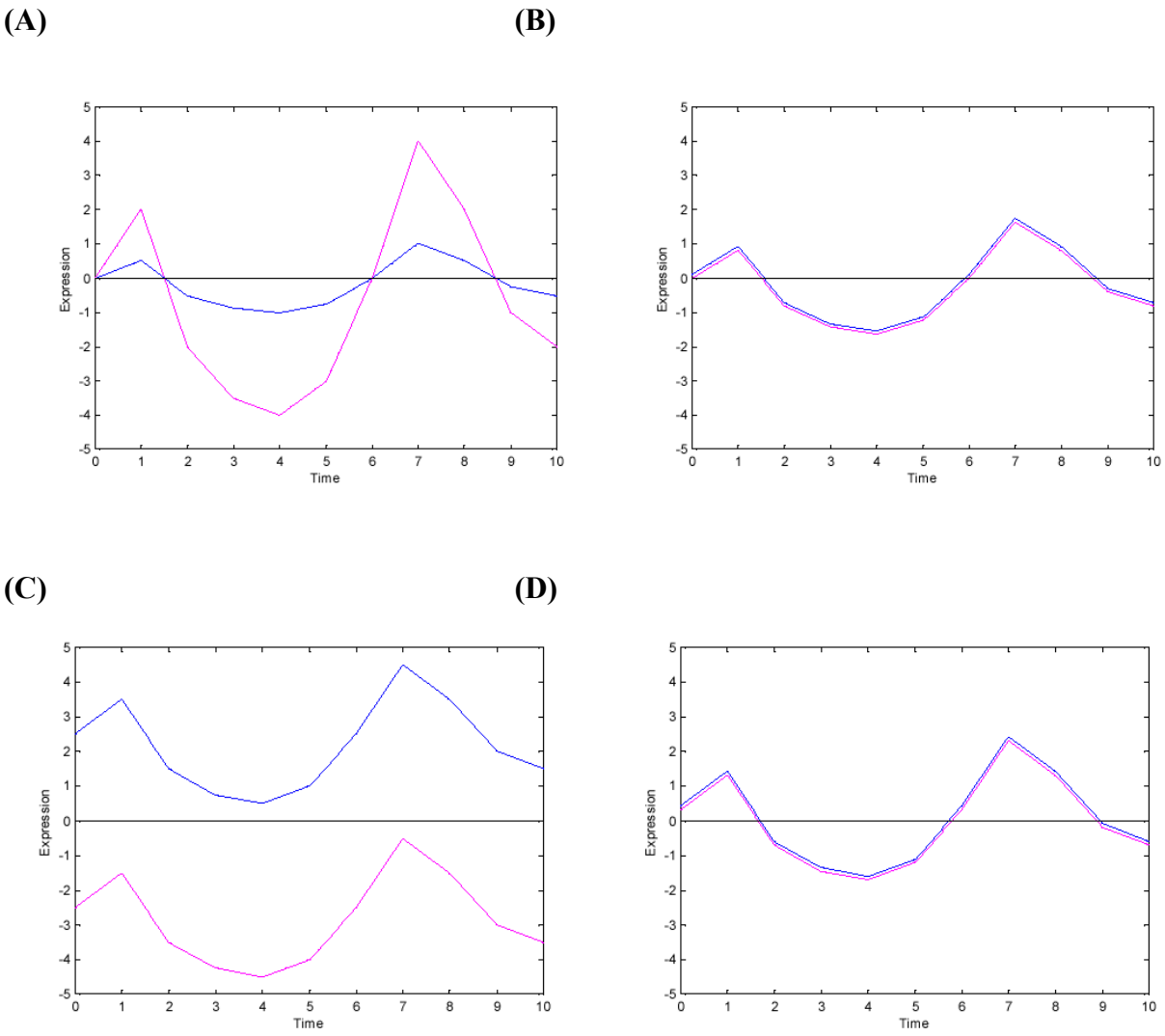


Figure 23. Centrer- réduire (d'après Sturn, 2001). Soit les profils d'expression de 2 gènes, (A-B) Centrer les profils d'expression des gènes sur la moyenne des profils d'expression : (A) avant transformation un gène est sur-exprimé et l'autre est sous-exprimé (B) après transformation les deux profils sont identiques. (C-D) Réduction des profils d'expression des gènes par l'écart type des profils d'expression : (C) un des profils d'expression est très faible tandis que le second a une amplitude de variation très importante, (D) après réduction du signal par l'écart type les deux profils sont identiques.

Une autre transformation courante est la **réduction des valeurs au moyen d'une division ou d'une multiplication par un scalaire** tel l'écart type. L'écart type caractérise la variabilité des données. Diviser par ce facteur a pour effet d'**amplifier les valeurs faibles et réduire les signaux élevés** (Fig. 23C-D). Cependant cela signifie également que l'amplitude d'un signal très bas, qui ne représente que du bruit, devient comparable à celles des forts signaux. Par conséquent, cette procédure peut créer des similarités qui n'existent pas.

Enfin, center et réduire simultanément les données d'expression revient à les standardiser (Draghici 2003). La transformation dite du « z-score », par exemple, combine la soustraction de la moyenne et la division par l'écart type. Une transformation de type « z-score » modifiée par la médiane et la déviation absolue de la médiane (MAD) est également possible. Ces estimateurs sont en général de meilleurs indicateurs de l'hétérogénéité des mesures. La standardisation des données doit être appliquée avec précaution. En effet, la standardisation des gènes génère un ensemble de gènes identiques. Un gène affecté uniquement pas le bruit environnant ne pourra par être distingué d'un autre gène qui varie beaucoup entre les expériences. La standardisation des échantillons est applicable dans un plus grand nombre de circonstances mais tend également à atténuer les différences entre les données.

2.2 Mesure de distance

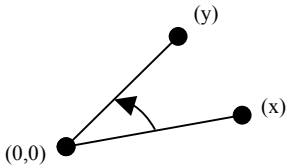
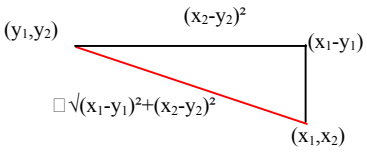
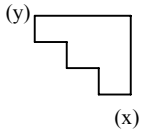
Une mesure de distance, ou métrique, est une formule qui calcule un nombre positif reflétant la proximité (similarité) ou l'éloignement (dissimilarité) entre deux points d'un espace des données (Tab. 10).

Une distance *d* entre deux points *x* et y se note *d(x, y)*. Cette distance possède trois propriétés:

- symétrie, $d(x, y) = d(y, x)$
- positivité, $d(x, y) >= 0$
- inégalité triangulaire, $d(x, y) \leq d(x, z) + d(z, y)$

L'inégalité triangulaire indique qu'une distance entre deux points doit être mesurée le long de la route la plus courte.

Tableau 10. Exemples de mesures de distance, avantages et inconvénients

Distance	Calcul	Schématisation	Avantages/Inconvénients
Coefficient de corrélation de Pearson	$d_r(x,y) = 1 - r_{xy} \quad \in [0\,;\,2]$ $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(yi - y)^2}}$		▪ Mesure de similarité[1] ▪ Non modifiée par une transformation scalaire des données ▪ Ne prend pas en considération l'amplitude de variation (*fold change*) entre les données ▪ Sensible aux valeurs aberrantes ▪ Suppose la linéarité des données
Angulaire	$d_A(x,y) = \cos(\theta) = \frac{x.y}{\|x\|\|y\|}$		▪ Non modifiée par les changements d'échelle
Euclidienne (d'après le théorème de Pythagore)	$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$		▪ Intuitive ▪ Mesure de dissimilarité[2] ▪ Ne permet pas la détection de corrélation négatives ▪ Sensible aux changements d'échelle ▪ Sensible aux valeurs manquantes
Manhattan	$d_M(x,y) = \sum_{i=1}^{n}\|x_i - y_i\|$		▪ Sensible aux valeurs aberrantes

[1] plus sa valeur est élevée, plus les objets se ressemblent ; [2] plus la distance est élevée plus les objets sont différents.

a) Corrélation de Pearson

Le coefficient de Pearson est une **mesure de similarité** : plus sa valeur est élevée, plus les objets se ressemblent. Cette distance évalue si les coordonnées des vecteurs comparés évoluent dans le même sens, *i.e.* si deux **gènes** sont **co-régulés** (Fig. 24). Le coefficient de corrélation n'est pas modifié par la transformation scalaire des données, *i.e.* un changement d'échelle (addition, soustraction, multiplication par une constante). De plus, l'amplitude de variation (*fold change*) entre les données a peu d'influence. Par exemple, deux gènes co-régulés seront considérés similaires même si l'un est 2 fois plus exprimé que l'autre. Cette métrique est relativement sensible aux valeurs aberrantes (Draghici 2003) et ne permet pas de mettre en évidence les corrélations négatives.

Pour palier à ces inconvénients, des variantes de cette métrique existent (Sturn, 2000). Le **coefficient de Pearson non centré**, par exemple, **prend en compte l'amplitude de la variation entre les données**. Le **coefficient de Pearson au carré** (centré ou non) permet, quant à lui, d'**identifier les profils d'expression réciproques**. Les vecteurs de sens opposés sont alors positifs et sont considérés comme identiques.

b) Distance euclidienne

Contrairement au coefficient de corrélation, la **distance euclidienne est une mesure de dissimilarité** : plus la distance est élevée moins les objets se ressemblent. La distance euclidienne est la méthode la plus utilisée pour l'analyse des profils d'expression (Quackenbush, 2001). Sa formule a pour origine le théorème de Pythagore (Tab. 10). Elle est sensible à la taille des vecteurs, *i.e.* aux valeurs manquantes. Elle est également sensible aux changements d'échelle : deux gènes ayant le **même profil mais des amplitudes de variation différentes** seront **très distants** (Fig. 24).

Il existe plusieurs variantes de la distance euclidienne (Draghici 2003). La distance euclidienne normalisée permet, par exemple, de pondérer chaque point de mesure (dimension) par une quantité inversement proportionnelle à sa variabilité. En d'autres termes, l'espace des dimensions est déformé : il est rétréci le long des axes à forte variabilité et s'étire le long des axes à faible variance. Cette transformation permet d'égaliser les variances le long des axes et de mieux refléter la structure des données.

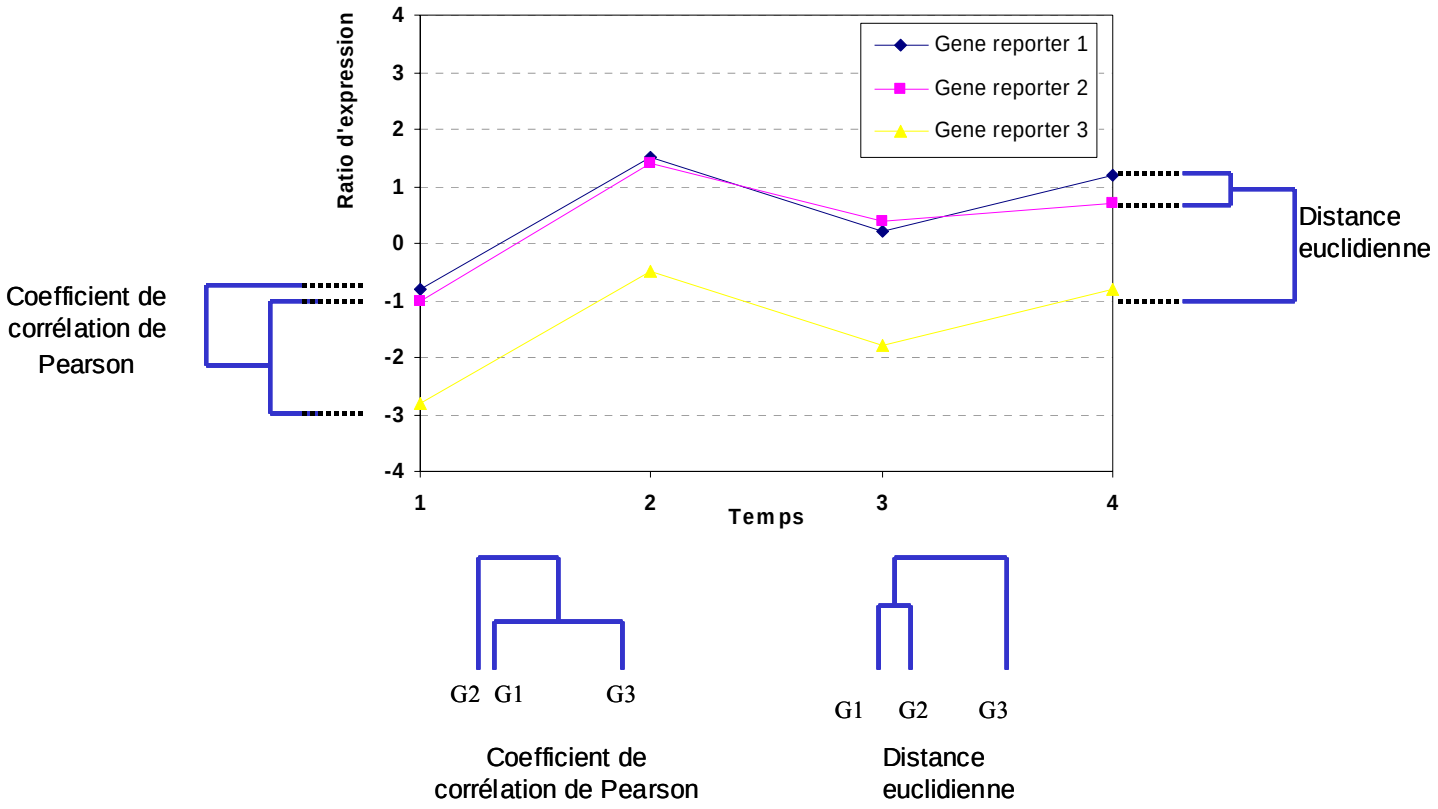


Figure 24. Coefficient de corrélation de Pearson *versus* distance euclidienne. L'application de différentes mesures de distance à des profils d'expression identiques donne des résultats différents. Le coefficient de corrélation de Pearson met en évidence les gènes co-régulés (dans le temps). La distance euclidienne privilégie l'amplitude des variations. Les gènes G1 et G3 sont co-régulés et sont donc proches selon la mesure du coefficient de corrélation de Pearson. En revanche, selon la distance euclidienne, les gènes G1 et G2 sont proches car leurs valeurs d'expression sont peu différentes.

La distance de Mahalanobis est une forme généralisée de la distance euclidienne normalisée. L'espace des données est déformé mais pas nécessairement le long des axes. Cette méthode améliore encore la définition de la structure des données. Enfin, la distance de Manhattan, ou *city-block distance* (en référence au plan des rues de New York), utilise les parallèles et non les diagonales entre les points. La distance de Manhattan est la somme des distances absolues entre deux vecteurs. Equivalente à la distance euclidienne, elle en possède les avantages et les inconvénients (Draghici 2003; Datta et Datta, 2003).

c) Distance angulaire

La distance angulaire est une mesure de similarité. Elle prend en considération uniquement les angles entre les vecteurs de données. L'amplitude des variations n'est pas prise en compte. Par conséquent, la distance angulaire n'est pas sensible au changement d'échelle et ne dépend pas de la taille des vecteurs.

d) Coefficients non paramétriques

Les coefficients non paramétriques se basent sur les rangs des observations pour chacune des variables. Par exemple, la corrélation des rangs de Spearman est la différence entre les rangs de chaque variable. Cette métrique est plus robuste face aux données bruitées que le coefficient de corrélation de Pearson. Toutefois, l'orientation des profils d'expression n'est pas prise en compte, aussi cette mesure n'est pas adaptée aux données de cinétiques.

Le Tau de Kendall évalue, quant à lui, la probabilité que deux variables (gènes) soient dans le même ordre pour les échantillons concernés. Cet indice examine de manière systématique les paires qu'il classe comme concordantes ou non. L'inconvénient majeur de cette approche est sa complexité, *i.e.* sa demande en puissance de calcul.

En bref

Centrer et/ou réduire les données d'expression ne doit pas être systématique, tout dépend de la question biologique posée. De même, il n'existe pas une distance de similarité (dissimilarité) meilleure qu'une autre. Le choix de la métrique à appliquer est fonction de la question posée. Les classifications qui en découlent peuvent être très différentes (Sturn, 2000; Draghici 2003).

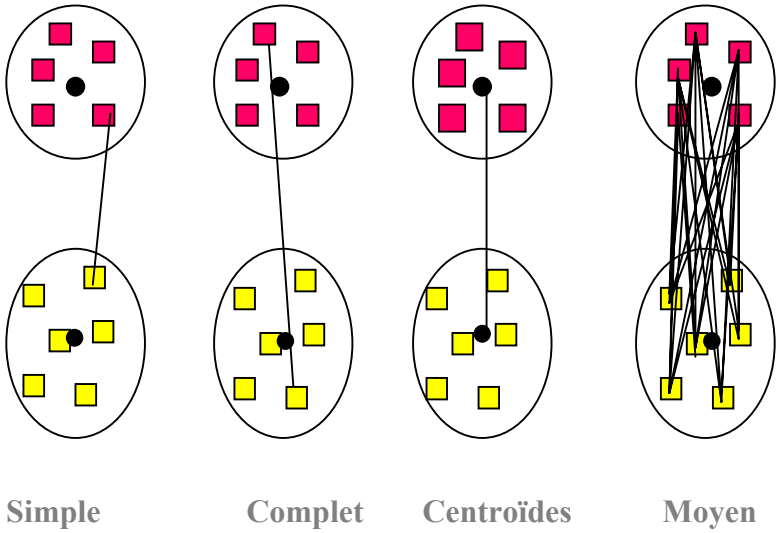


Figure 25. Distances inter-groupes dans le cadre d'une classification hiérarchique (d'après Draghici, 2003). Chaque cercle noir correspond à un cluster de gènes (carrés de couleur jaune ou fushia). La liaison entre les *clusters,* ou distance inter-groupes, peut être du type «lien simple », « lien complet », « un lien entre les centroïdes » (point noir) ou « lien moyen » entre tous les gènes.

3. Classification non supervisée

Les méthodes de classification non supervisée sont des techniques de regroupement (*clustering*) où un processus automatique sépare les données observées en groupes distincts sans aucune connaissance préalable des classes existantes. Par exemple, l'application de ces techniques a montrée que la taxonomie actuelle des cancers regroupe des maladies distinctes sur le plan moléculaire (Golub *et al.*, 1999; Alizadeh *et al.*, 2000). Ces travaux ont également mis en évidence des différences de réponses aux traitements thérapeutiques, associées à une différence d'évolution des cancers.

3.1 Classification hiérarchique

D'abord utilisées en phylogénie, les méthodes de classification hiérarchique sont aujourd'hui les techniques de classification non supervisée les plus utilisées pour étudier les profils d'expression de gènes ou d'échantillons. Elles génèrent des **suites de classes emboîtées qui définissent une hiérarchie de partitions** encore appelée classification hiérarchique. Les algorithmes de classification travaillent à partir des matrices de distances issues des matrices de données d'expression. Actuellement, il existe trois principales modalités de calcul de distances entre les classes (distance inter-groupes) qui permettent de générer deux grand types d'algorithmes de classifications hiérarchiques : les algorithmes ascendants et les algorithmes descendants.

a) Distances inter-groupes

Des règles de calcul, encore appelées **règles d'agglomération**, sont nécessaires pour estimer les liaisons entre les groupes disjoints. Les principales distances inter-groupes sont actuellement le lien simple, le lien complet et le lien moyen (Fig. 25).

Le **lien moyen** (*average linkag*e) ou UPGMA (***U**nweighted **P**air **G**roup **M**ethod of **A**gregation)* est l'approche la plus utilisée. La distance entre deux groupes est la **moyenne des distances entre toutes les paires d'objets** (gènes ou échantillons biologiques) **de ces deux groupes** (Fig. 25).

Le **lien simple** ou lien du saut minimum (*single linkag*e) est encore qualifié de lien du plus proche voisin (*nearest neighbor).* La distance entre deux groupes est déterminée par la **distance entre les deux éléments les plus proches***,* appartenant à deux groupes différents (Fig. 25). Lorsqu'il existe plusieurs distances minimales équivalentes entres des groupes, le lien simple est l'algorithme d'agrégation le plus approprié. En revanche, dans les autres cas,

cette approche a tendance à générer des singletons par agglomérations de groupes très différents mais dont les voisins sont très proches (Yeung *et al.*, 2001). Cette méthode est généralement peu applicable aux données de puces à ADN.

Le **lien complet** *(complete linkag*e) ou lien d'agrégation par le diamètre est encore qualifié de lien du voisin le plus distant (*furthest neighbor)*. La distance entre deux groupes est déterminée par la **distance entre les deux éléments les plus éloignés***,* appartenant à deux groupes différents (Fig. 25). Généralement les groupements générés sont de petites tailles et fusionnent très tard dans la hiérarchie. La méthode du lien complet est particulièrement efficace si les objets appartiennent naturellement à des groupes de données distants dans l'espace de données. Toutefois, cette approche est sensible aux valeurs manquantes, même en faible nombre.

D'autres distances, moins utilisées car plus complexes et rarement proposées dans les logiciels, sont par exemple la distance entre les centroïdes (centre de gravité) ou le lien de Ward (Ward's linkage). La liaison entre les centroïdes est la distance entre les centres calculés des groupes (Fig. 25). Le lien de Ward utilise, quant à lui, l'analyse de variance pour évaluer la distance qui minimise la somme des carrés entre les groupes. Cette approche tend à obtenir des groupes compacts et de petite taille. Cependant, cette technique est sensible aux valeurs manquantes.

b) Algorithmes ascendants

Les algorithmes ascendants ou agglomératifs (*bottom-up*) construisent des groupes par **agrégations successives** des éléments les plus proches deux à deux pour fournir une hiérarchie de partitions des objets (Fig. 26). Le terme *élément* désigne donc à la fois les individus ou objets à classer et les regroupements d'individus générés par l'algorithme. Au départ, chaque élément constitue un groupe de taille 1. Puis, à chaque étape, les deux groupes les plus proches sont recherchés et fusionnés jusqu'à ce qu'il n'y ait plus qu'un seul groupe.

(A)

	G1	G2	G3	G4
G1	0			
G2	0.57	0		
G3	0.39	0.46	0	
G4	0.03	0.62	0.43	0

(B)

	(G1G4)	G2	G3
(G1G4)	0		
G2	0.61	0	
G3	0.42	0.46	0

(C)

	(G1G4G3)	G2
(G1G4G3)	0	
G2	0.74	0

(D)

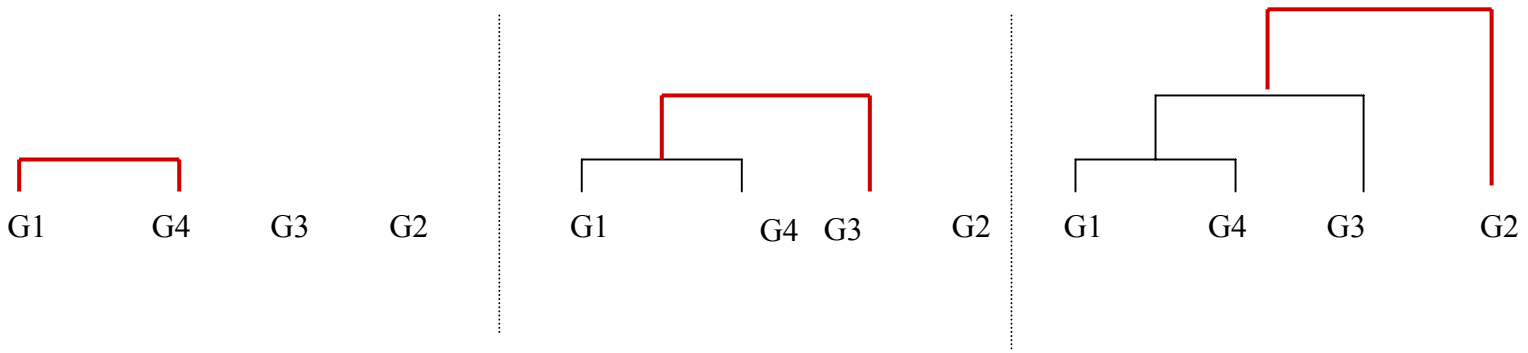


Figure 26. Principe de construction d'une classification hiérarchique ascendante. (A) Le tableau initial (gauche) présente les distances euclidiennes calculées entre 4 gènes (G1 à G4). La plus petite distance (en rouge) est choisie pour associer les gènes les plus similaires soit G1 et G4. (B) Les distances entre ce groupe et les autres gènes sont calculées. (C) L'algorithme fonctionne de manière itérative jusqu'à ce que l'ensemble des gènes appartienne à un seul et même groupe. (D) Le résultat des différentes partitions est représenté sous la forme d'un arbre, ou dendrogramme (bas de la figure).

La hiérarchie obtenue est généralement représentée sous la forme d'un arbre planaire hiérarchique, également appelé dendrogramme, contenant *n - 1* partitions (Fig. 26). Ce dendrogramme décrit de façon explicite la structure finale de la hiérarchie. Les individus qui se ressemblent le plus se regroupent dans le bas de l'arbre. La longueur des branches témoigne de leur éloignement. Cette représentation graphique permet également d'estimer le nombre de classes existant effectivement dans la population. Enfin, il faut noter l'aspect isomorphe d'un dendrogramme : les nœuds de l'arbre pivotent. Si *n* est le nombre de nœuds, il existe $2^{(n-1)}$ représentations possibles. Seule la longueur des branches aide à définir la proximité entre les groupes.

Associé à l'arbre, les matrices de données d'expression sont transformées en cartes « thermiques » (*heatmap*), colorées en fonction du niveau d'expression relatif des gènes. Généralement, les couleurs employées vont du vert (niveau bas) au rouge (niveau élevé) en passant par le noir (niveau médian).

Pour exemple, l'algorithme de classification hiérarchique ascendante a été appliqué à une étude réalisée à la demande de S. Nattel (Montreal Heart Institute, Canada) (Fig. 27). Leur travail visait à analyser le remodelage de l'expression des gènes cardiaques suite à différents modes de stimulation du cœur (Tab. 11). Le modèle utilisé est le chien. Les données analysées sont issues de puces à oligonucléotides Affymetrix® représentant le génome du chien et réalisées au Montreal Genome Center, Canada.

Tableau 11. Détails des échantillons des cœurs de chien analysés dans l'étude réalisée à la demande de S. Nattel (Montreal Heart Institute, Canada).

Nomenclature	Echantillons	Nombre de chiens par conditions	Durée
Contrôle	SHAM	5	-
Simulation rapide du ventricule gauche et mesure dans l'oreillette	VTPLA_24h VTPLA_2w	5 5	24 heures 2 semaines
Simulation rapide de l'oreillette gauche et mesure dans l'oreillette	RAP_24h RAP_1w	5 5	24 heures 1 semaine

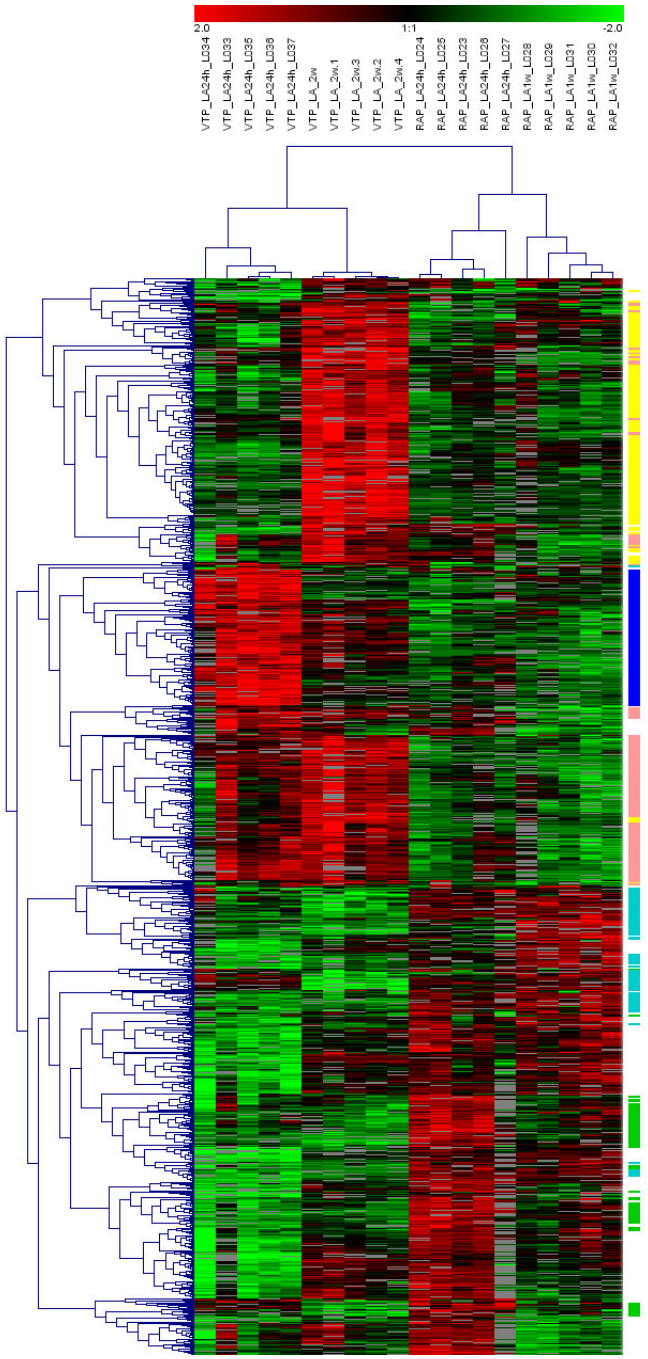


Figure 27. Classification hiérarchique ascendante (Etude réalisée à la demande de S. Nattel). Classification hiérarchique ascendante utilisant le coefficient de Pearson non centré pour mesurer de similarité et la moyenne entre les groupes (lien moyen) pour algorithme d'agrégation. La classification répartit les échantillons principalement en 2 groupes : les échantillons ayant subi des stimulations ventriculaires et les échantillons ayant subi des stimulations auriculaires. Au sein de ces regroupements, chaque temps (24 heures, 1 ou 2 semaines) est caractérisé par un ou plusieurs groupes de gènes significativement différentiellement exprimés.

Le traitement des données a été effectué en collaboration avec E. Libby et L. Glass à McGill University, Canada. Les méthodes RMA (Irizarry *et al.*, 2003) et dChip (Li et Wong, 2001), dédiées à l'analyse des données Affymetrix®, ont été utilisées pour la normalisation des données primaires. L'algorithme Zmad, implémenté dans MADSCAN (Le Meur *et al.*, 2004), a été utilisé pour rechercher les valeurs aberrantes. L'application SAM a été employée pour la mise en évidence des gènes différentiellement exprimés (Tusher *et al.*, 2001).

La classification obtenue est représentée sous la forme de deux dendrogrammes (un pour les gènes, un pour les individus) et d'une carte « thermique ». La classification répartit les échantillons principalement en 2 groupes : les échantillons ayant subi des stimulations ventriculaires et les échantillons ayant subi des stimulations auriculaires. De même au sein de ces regroupements, chaque temps (24 heures, 1 ou 2 semaines) est caractérisé par un ou plusieurs groupes de gènes significativement sur-exprimés.

c) Algorithmes descendants

Les algorithmes descendants ou divisifs (*top-down*) procèdent par dichotomies successives de l'ensemble des éléments pour fournir une hiérarchie de partitions. Ces algorithmes sont non déterministes, *i.e.* qu'il n'existe pas de solution finale unique.

Au départ, tous les objets appartiennent à un seul et même groupe. Un algorithme de partitionnement est utilisé pour diviser ce groupe en deux sous-ensembles. Cet algorithme est appliqué de manière récursive jusqu'à ce que tous les groupes soient de taille 1. A l'inverse des algorithmes ascendants hiérarchiques, dans le cas d'une représentation par dendrogramme, plus le groupe est bas dans l'arbre, moins bonne est la représentation de la structure des données.

Les algorithmes descendants sont peu utilisés, essentiellement par ce qu'ils demandent une plus grande ressource de calcul que les approches agglomératives. Ils sont donc peu représentés dans les logiciels d'analyse. Aucun exemple n'est donné ici. Nous pouvons tout de même citer la librairie *DIANA* (***DI****visive* ***ANA****lysis*) implémenté dans R.

3.2 Méthodes de partitionnement

L'objectif des méthodes de partitionnement est de minimiser la distance intra-groupe pour un nombre fixé, *K*, de groupes.

a) K-moyennes

La méthode dite des *k*-moyennes *(k-means),* introduite en 1967 par MacQueen, est une variante des méthodes d'agrégations autour de centres mobiles (techniques de ré-allocation dynamique des individus à des centres de classes eux-mêmes recalculés à chaque itération). Le but de cet algorithme est de **minimiser la distance de chaque objet** (*e.g.* gènes) par **rapport au centre du groupe auquel il appartient**.

La méthode des *k*-moyennes distribue les données en *k* groupes choisis *a priori* et répartis autour de *k* centres appelés noyaux ou centroïdes (Fig. 28).

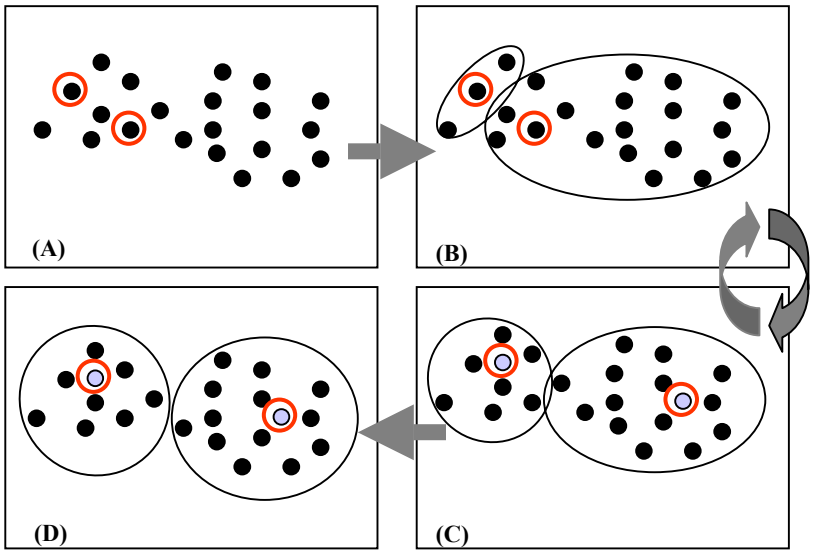


Figure 28. Algorithme des *K-moyennes* (A) Définition du nombre de classes et tirage aléatoire des centroïdes (cercle rouge). (B) Allocation des gènes aux classes. (C) Calcul des nouveaux centroïdes (violet), ré-allocation des gènes de manière itérative (double flèche) jusqu'à obtention de la partition finale (D).

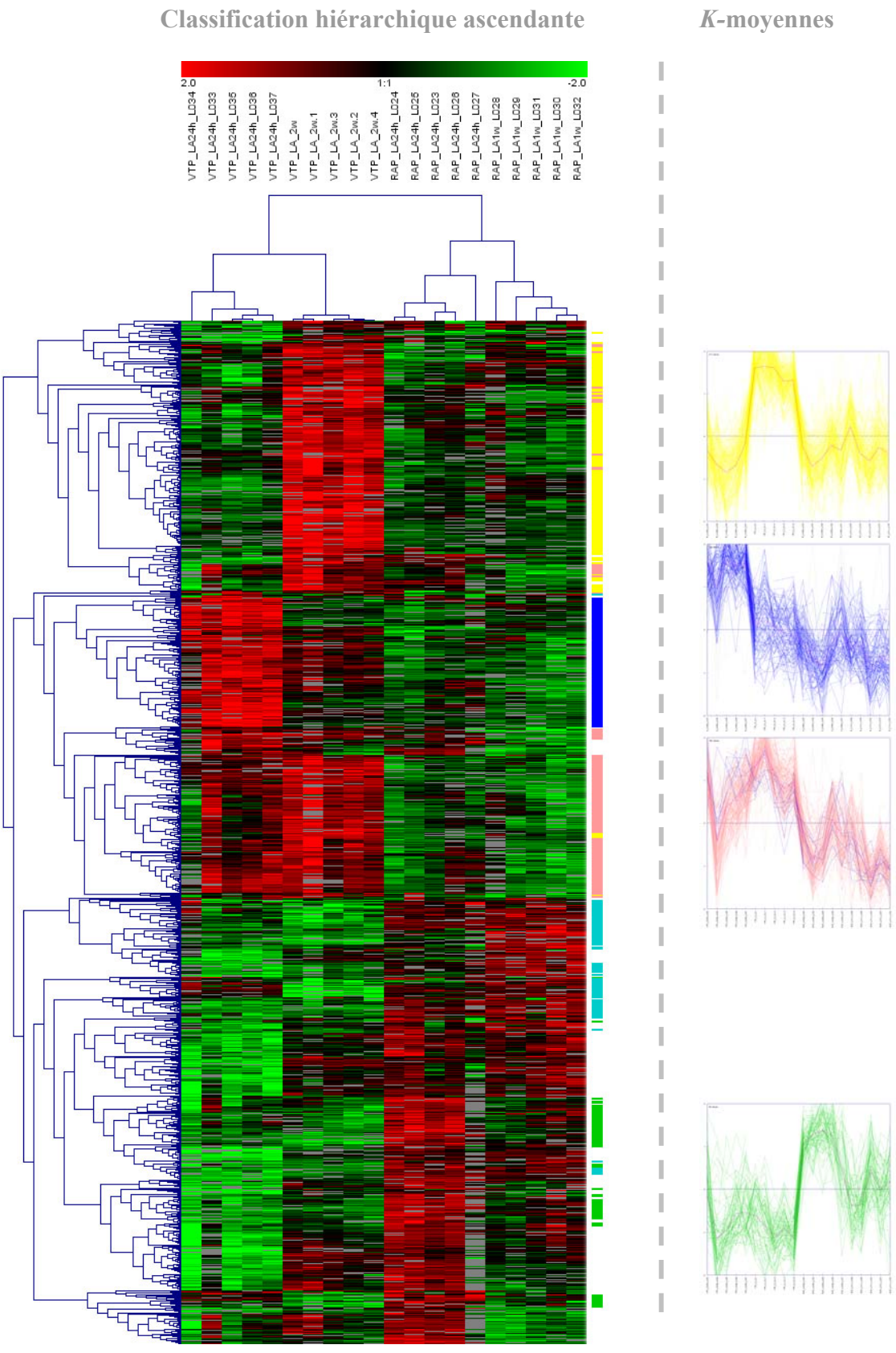


Figure 29. Classification hiérarchique ascendante *versus* *K*-moyennes (Etude réalisée à la demande de S. Nattel). (A) Classification hiérarchique ascendante utilisant le coefficient de Pearson non centré comme mesure de similarité et le lien moyen pour algorithme d'agrégation. (B) *K*-moyennes avec k=10; itération=50. Les groupes définis par l'algorithme de classification ascendante identifient clairement les différents échantillons. L'algorithme des *K*-moyennes apportent des informations complémentaires sur les gènes co-régulés qui ne sont pas toujours directement visibles.

Lors de l'initialisation, les *k* centroïdes sont tirés au hasard soit à partir de :

(i) l'ensemble des données,

(ii) des profils composites issus des données de départ,

(iii) d'un ensemble de données plus vaste représentant la population étudiée.

A partir de ces *k* centres, chaque individu est affecté à l'un des noyaux (le plus proche) ce qui permet de former *k* groupes. Le barycentre de chaque groupe est alors calculé pour constituer k nouveaux centres. L'opération est réitérée jusqu'à convergence.

Aujourd'hui, cette approche compte parmi les algorithmes de groupement les plus simples et les plus rapides en conséquence, l'un des plus utilisés (Tavazoie *et al.*, 1999). L'algorithme des *k*-moyennes a également été appliqué à une étude de S. Nattel (Montreal Heart Institute, Canada) (Fig. 29). Les résultats obtenus apportent des informations complémentaires sur les gènes co-régulés qui ne sont pas toujours directement visibles.

Toutefois, l'algorithme des *k*-moyennes est relativement sensible aux valeurs aberrantes. Kaufman et Rousseeuw (1990) proposent l'algorithme des *k*-médoïdes ou PAM (***P****artionning* ***A****round* ***M****édoïds*) qui permet de classifier les données de manière plus robuste. La médoïde d'un groupe est l'objet possédant la distance médiane la plus faible avec les autres objets du groupe. D'après ces auteurs, le calcul d'un centroïde peut se révéler peu significatif dans certains cas, en particulier lorsque les données sont de types catégorielles ou discrètes. Il est plus judicieux de choisir comme centre du groupe un objet présent dans le groupe et non un objet calculé.

Finalement, du fait de la phase d'initialisation au hasard, ces algorithmes sont dit non déterministes : l'algorithme appliqué plusieurs fois sur le même jeu de données peut produire des résultats différents. Le problème étant lié à l'optimisation d'une combinatoire, la solution trouvée sera rarement l'optimum global mais plutôt un des nombreux optimums locaux. D'après Draghici (2003), la position relative des profils d'expression mise en évidence par ces méthodes est rarement informative voir trompeuse.

b) Réseaux de Kohonen

Les réseaux de Kohonen (*Kohonen's map*), encore appelés cartes auto-organisatrices (***S****elf* ***O****rganisation* ***M****ap* - SOM) (Kohonen 1995), sont des **réseaux de neurones** qui utilisent une méthode d'apprentissage incrémentale dite compétitive (*data driven)*. Cette méthode est dérivée de l'approche *k*-moyennes sur laquelle des contraintes spatiales (topologiques) sont

ajoutées sous la forme d'un réseau virtuel (Fig. 30). Ce réseau, ou carte, **permet de réduire l'espace multidimensionnel des données d'entrée en un espace à 1 (ligne), 2 (grille) ou 3 (parallélépipède) dimensions**. Les cartes 1D et 2D sont les plus utilisées.

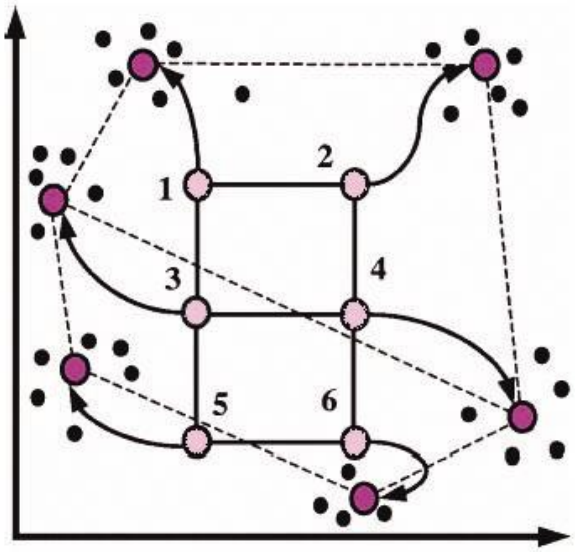


Figure 30. Réseaux de Kohonen (SOM) (d'après Tamayo *et al.*, 1999). La configuration initiale de la carte est un rectangle 3 x 2 nœuds (en rose), matérialisé par les lignes continues. Les trajectoires hypothétiques des nœuds lors de leur ajustement par itérations successives sont indiquées par les pointillés, les flèches et les points fushia. Les données d'expression sont les points noirs.

L'algorithme SOM, bien que non supervisée, nécessite la définition de plusieurs paramètres pour son initialisation. Tout d'abord, la topologie du réseau est définie par un ensemble de points interconnectés (nœuds ou neurones) (Fig. 30). Il est nécessaire de spécifier le nombre de nœuds et leur ordonnancement. Le nombre de nœuds correspond au nombre de groupes attendus. La configuration du réseau peut être rectangulaire ou hexagonale, en une ou plusieurs dimensions. De plus, la méthode d'apprentissage est basée sur deux paramètres : le facteur d'apprentissage α (*learning rate*) et la taille du voisinage *r* (*radius, neighborhood* ou *grid*). La fonction de voisinage *r* peut être de différentes formes : *bubble*, gaussienne, *cut* gaussienne, *epanechicov*…(Sturn, 2000). A noter que lorsque *r* =0, un seul prototype est modifié à chaque étape et l'algorithme de Kohonen est similaire à l'algorithme des *k*-moyennes.

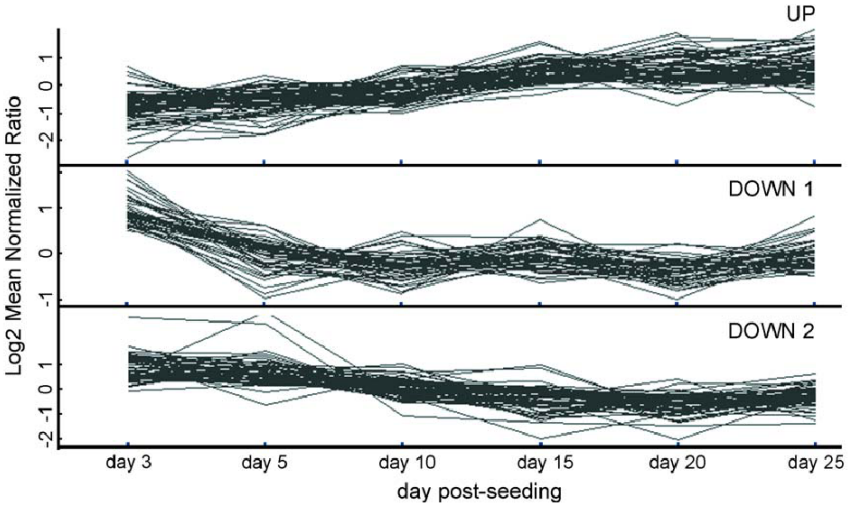


Figure 31. Exemple de classification par l'algorithme SOM des gènes principalement impliqués dans les mécanismes de différenciation des entérocytes (extrait de Bédrine-Ferran, H., Le Meur, N. *et al*, 2004) (GeneSight 3.0 Biodiscovery, Inc). Le jeu de données utilisé se compose des gènes différentiellement exprimés ayant au moins 4 points de mesures sur les 6 temps de la cinétique. Une première classification SOM en 4 groupes a été faite sur l'ensemble des données. Les graphiques représentent le résultat d'une seconde application de l'algorithme aux trois groupes possédant le plus grand nombre de gènes.

La procédure se déroule en 2 temps : la phase d'initialisation et la phase d'apprentissage. L'étape d'initialisation a pour but d'associer à chaque nœud un vecteur, établi à partir de profils d'expression tirés au hasard. La phase d'apprentissage tente d'agréger les profils d'expression aux différents vecteurs qui sont ré-estimés à chaque itération (Tamayo *et al.*, 1999).

L'algorithme de SOM a notamment été utilisé par Bédrine-Ferran, H *et al.* (2004) (*cf.* p.60) afin de mettre en évidence les gènes principalement impliqués dans les mécanismes de différenciation des entérocytes (Fig. 31). Ce mode de classification a permis d'identifier 3 groupes de gènes : des gènes « UP » dont l'expression croit de façon constante au cours de la différentiation ; des gènes « DOWN1 » dont l'expression décroît rapidement au début de la différenciation puis se stabilisent ; des gènes « DOWN2 » dont l'expression décroît faiblement mais de manière constante tout au long de la différentiation.

La propriété la plus intéressante de cette carte est la faculté de préserver les relations topologiques des données d'entrée. En effet, contrairement à la classification hiérarchique et à l'algorithme des *k*-moyennes, la position des groupes dans l'espace de résultat reflète le degré de similarité entre les données. Les données projetées dans un même voisinage ont des profils d'expression proches. Autre avantage, l'analyse est peu sensible aux valeurs manquantes.

Cette technique présente également quelques inconvénients. La phase d'initialisation est complexe et l'utilisation de points appartenant à l'ensemble des données pour définir les nœuds est fortement conseillée. Le choix de la taille et de la forme du réseau a une composante heuristique. Le nombre de neurones détermine la granularité des résultats, *i.e.* leur précision. Dans la majorité des cas, le choix est réalisé après des essais successifs ou une analyse descriptive préliminaire. De plus, tout comme l'algorithme des k-moyennes, l'algorithme SOM est non déterministe. Enfin, les résultats dépendent de la distance choisie (Draghici 2003).

3.3 Autres méthodes de regroupements non supervisés

Les matrices de données d'expression sont devenues une source pour le développement de nouvelles approches de classification. De nouvelles méthodes (et outils informatiques) apparaissent ainsi chaque mois. Moins utilisées que les méthodes décrites précédemment, les approches de logique floue, bipartition ou *gene shaving* peuvent être particulièrement intéressantes.

La classification floue ou *fuzzy algorithm* est très voisine des k-moyennes (Gasch et Eisen, 2002). Elle permet le classement des gènes dans plusieurs groupes à la fois, avec une probabilité associée. Selon Moloshok e*t al.* (2002), cette technique reflète mieux la biologie. Il existe peu d'outils proposés pour la classification floue, nous pouvons toutefois citer la librairie *Fanny*, implémentée dans R (Datta et Datta, 2003).

Les méthodes de bipartition (*biclustering*) cherchent à regrouper simultanément les lignes et les colonnes d'une matrice d'expression afin d'obtenir des sous-groupes homogènes et stables (Cheng et Church, 2000). Ces techniques généralisent les méthodes de regroupements classiques basées sur le regroupement indépendant des lignes et des colonnes. Ces techniques permettent d'identifier et organiser des sous-groupes de gènes co-exprimés dans des sous-groupes de conditions, autorisant les gènes à participer à plusieurs groupes. Ainsi, Kluger *et al.* (2003) ont montré que des gènes co-régulés dans un type de tumeur ne le sont pas nécessairement dans un autre. Peu connues, ces méthodes sont peu disponibles sous la forme de logiciel, nous pouvons tout de même citer SAMBA[37] (***S**tatistical-**A**lgorithmic **M**ethod for **B**icluster **A**nalysis*) (Sharan *et al.*, 2003).

Enfin, le *gene shaving* (Hastie *et al.*, 2000) permet d'identifier des petits groupes de gènes corrélés qui optimisent la variation entre les différents échantillons. L'algorithme commence par déterminer la composante principale des gènes. Pour chaque gène, la valeur absolue de sa corrélation avec la composante principale est calculée. La fraction *a* (en général 10%) des gènes les moins corrélés est supprimée. Ces étapes peuvent être répétées jusqu'à n'obtenir plus qu'un seul gène. La méthode du *gene shaving* génère une suite de groupes emboîtés à la manière d'une classification hiérarchique descendante. Toutefois, contrairement aux méthodes de classification hiérarchique et tout comme les approches de logique floue et de bipartition, un gène peut faire partie de plusieurs groupes. De plus, cette méthode peut être utilisée dans le cadre d'une classification supervisée.

3.4 Validation des regroupements

Les regroupements réalisés avec différents algorithmes et/ou métriques ne donnent pas nécessairement les mêmes résultats. De plus, ces algorithmes ne présentent pas les mêmes propriétés face aux caractéristiques atypiques des matrices de données d'expression. Trois

[37] http://www.cs.tau.ac.il/~rshamir/expander/expander.html

questions se posent alors : **Quelle est la qualité d'un groupe ? Est-il stable ? Quelle est la précision des regroupements obtenus ?**

Une manière d'évaluer la **qualité** d'un groupe est de comparer sa dispersion à la distance qui le sépare du groupe le plus proche (Draghici 2003). En effet, si la **distance inter-groupes** est plus grande que la **dispersion des objets au sein d'un groupe**, ces groupes sont bien disjoints. Le ratio de ces distances est un bon estimateur de la qualité d'un groupe. L'indice de Dunn, proposé par Azuaje (2002), reprend cette approche de façon itérative. Il permet, par exemple, de déterminer le nombre optimum de groupes *k* suite à plusieurs applications du même algorithme de partitionnment avec différents paramètres (*i.e.* algorithme des *k*-moyennes avec k variant). La moyenne des distances entre les membres d'un groupe et son centre peut également être un indicateur de qualité. Ceci se traduit par la hauteur des branches du dendrogramme, proportionnelle à la distance entre les objets. Enfin, le logiciel *Machaon* CVE (*Machaon Cluster Validation Environment*) propose un ensemble d'indices de validité des regroupements parmi lesquels les indices de Dunn, Jaccard ou Goodman-Kruskall (Azuaje, 2002; Bolshakova *et al.*, 2005). Cet outil permet non seulement d'évaluer la qualité des regroupements mais aussi d'estimer le nombre optimum de groupes.

A la question « **est-ce que le groupe est stable** ? », nous cherchons à savoir si nous obtenons les mêmes regroupements (de gènes et/ou d'échantillons) lorsque deux expérimentateurs évaluent les mêmes données et réalisent la même analyse. En d'autres termes, est-ce que les éléments du groupe sont réellement liés ou est-ce dû au hasard ou à l'erreur expérimentale ?

Les méthodes les plus connues pour évaluer la stabilité des groupes sont certainement les approches par ré-échantillonnage aléatoire de type ***bootstrap*** ou ***Jackknife***. Datta et Datta (2003) proposent, quant à eux, des indices de qualité basés sur la perturbation du jeu de données par retrait, au hasard, d'un échantillon du jeu de données **(*Leave-one-out*)**. Ils suggèrent ainsi d'évaluer la proportion de gènes (échantillons) mal classés après le retrait d'un échantillon. Grâce à cette technique, ces auteurs ont comparé 6 algorithmes de regroupement proposés dans divers librairies du projet R. Ils ont ainsi montré que les méthodes DIANA et *mclust* (Banfield et Raftery, 1993) sont les algorithmes les plus performantes.

D'autres méthodes, suggèrent l'**utilisation des mesures répétées** afin d'évaluer la stabilité et la précision des regroupements. Kerr et Churchill (2001) proposent une analyse de

variance des données répétées couplée à un ré-échantillonnage aléatoire. Dans ce but, les données d'origine sont perturbées en ajoutant du bruit sous la forme de paramètres d'erreur estimés à partir de la variabilité des données répétées. Ces données perturbées sont ensuite regroupées. Les regroupements ainsi obtenus sont comparés aux originaux. Cette méthode permet d'évaluer la reproductibilité des regroupements pour un jeu de données et un algorithme donné. Cependant, elle ne permet pas d'évaluer la précision des résultats (Yeung et Bumgarner, 2003). Une amélioration de cette approche est proposée par Dudoit et Fridlyand (2004). Ils démontrent que l'agrégation des résultats par ré-échantillonnages successifs (*bagging* pour ***b**oostrapp **agg**regat**ing***) permet non seulement d'évaluer leur stabilité mais aussi d'améliorer les performances de ces classifications. Enfin, l'outil Rosetta Luminator® (Rosetta Biosoftware, Inc.) propose la construction d'un modèle d'erreur dérivé des données répétées, obtenues par l'ensemble des expériences menées avec une technologie de puce à ADN donnée (Hughes, 2002). Rosetta Luminator® utilise ce modèle pour estimer l'erreur sur la mesure de chaque gène. Les algorithmes de classification peuvent ensuite utiliser ces estimations pour pondérer les mesures de distance. Les points de mesure incertains auront moins de poids dans le regroupement. Cette stratégie permet d'améliorer non seulement la stabilité mais aussi la précision des regroupements.

Enfin, Yeung *et al.* (2003) propose un modèle bayésien mixte (*Bayesian **I**nfinite **M**ixture **M**odel- IMM*). Cette approche permet d'estimer la probabilité *a posteriori* d'appartenir à un groupe et par conséquent d'évaluer la stabilité des groupes d'objets non corrélés.

En bref

Les méthodes de classification non supervisées (*clustering*) groupent les objets sans *a priori.* Les principales méthodes sont les algorithmes de classification hiérarchique et les méthodes de partitionnement. Ces techniques séparent les données observées en groupes distincts sur la base de leurs similarités ou dissemblances. De nouvelles approches permettent également qu'un objet (gène) appartienne à un ou plusieurs groupes.

La validation des regroupements est une étape importante. Elle permet d'évaluer la qualité, la stabilité et la précision des groupes.

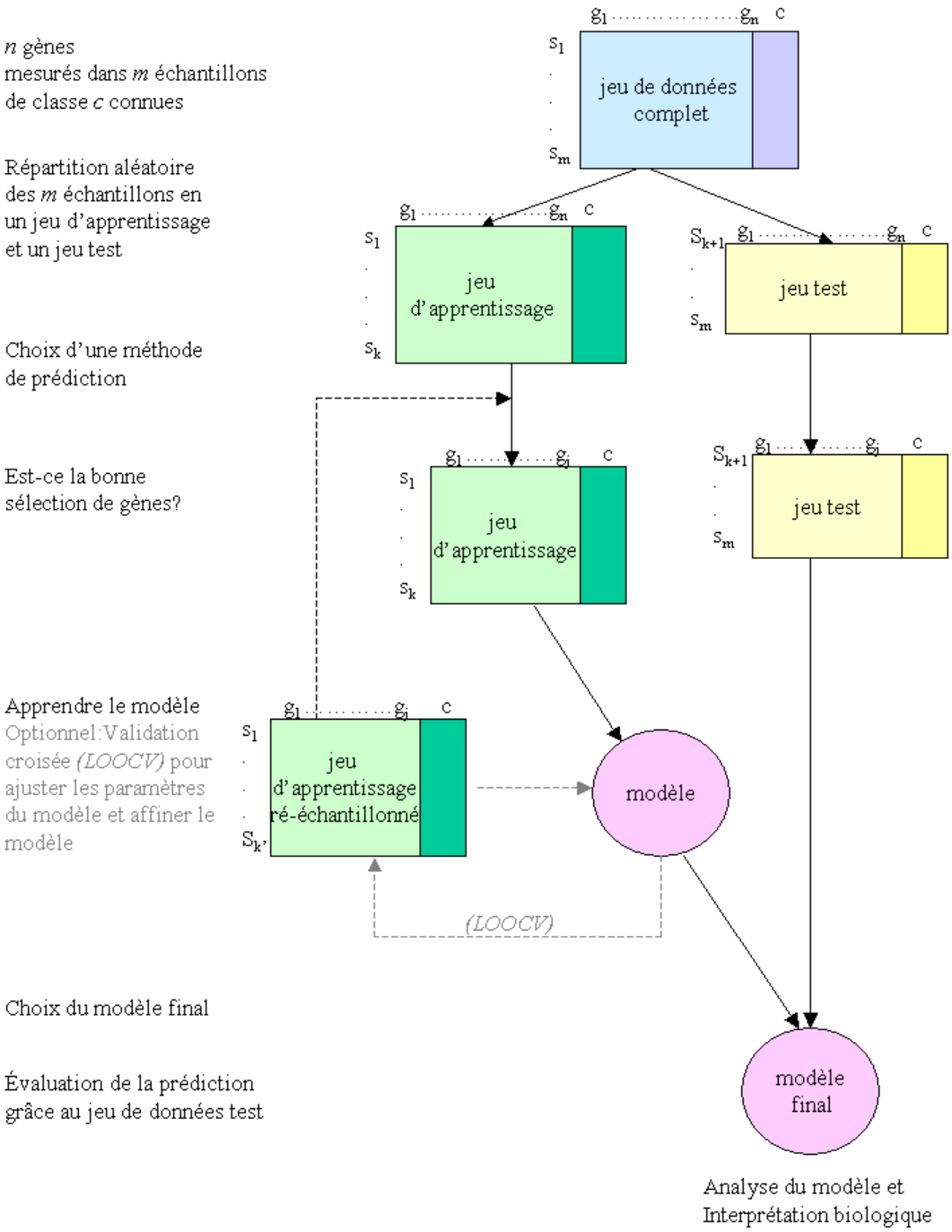


Figure 32.Exemple de processus de construction d'un modèle prédictif pour classer des échantillons (d'après Slomin, 2002). Le jeu complet de données d'expression est réparti au hasard en un jeu de données d'apprentissage et un jeu de données test. La majorité des algorithmes d'apprentissage nécessite l'optimisation de paramètres (tels le nombre de gènes, le nombre des plus proches voisins). L'ajustement des paramètres et du modèle est généralement fait au fur et à mesure des étapes de validation. Le principe de la validation est d'extraire de manière répétée de petits échantillons tests du jeu de données d'apprentissage, construire de nouveaux modèles et évaluer les performances du modèle. Par exemple, « *leave-one-out cross validation* » construit *n* modèles, chacun utilisant *n-1* exemples du jeu d'apprentissage, et test la précision de la prédiction des n échantillons. Finalement le modèle final, éventuellement choisi durant la phase de validation, est testé sur un nouveau jeu de données (jeu test) non utilisé durant sa construction.

4. Classification supervisée

Les méthodes de classification supervisée, contrairement aux méthodes de découvertes de profils, utilisent de la **connaissance *a priori* pour construire des classifications** (Slonim, 2002). Elles établissent des règles et un modèle de classification à partir d'un jeu de données connu et annotés, dit jeu d'apprentissage (*training set*), afin de prédire la classification de nouveaux cas appartenant à un jeu de données test (*Class prediction*). Ainsi, la plupart des méthodes de classification supervisée comprennent (Fig. 32) :

(i) une phase d'apprentissage sur des échantillons dont la classification est connue,

(ii) une phase de test au cours de laquelle l'algorithme de classification est généralisé pour prédire la classification d'autres échantillons.

Dans le domaine des puces à ADN, l'un des objectifs de l'analyse des données d'expression est de mettre en évidence les gènes marqueurs d'une classe afin de rendre possible le diagnostic (Golub *et al.*, 1999; Alizadeh *et al.*, 2000; Khan *et al.*, 2001) et/ou le pronostic (Alizadeh *et al.*, 2000; Sorlie *et al.*, 2001) sur la base des portraits moléculaires. Une fois encore les dimensions et l'aspect bruité des matrices de données d'expression sont une limite à l'application des algorithmes de classification par apprentissage. Le risque principal est alors le « surapprentissage » (*overfitting*), *i.e.* l'algorithme ne parvienne pas à généraliser la classification et par conséquent ne parvienne pas à classer les échantillons tests.

Comme pour les approches non supervisées, de nombreuses techniques de classification supervisées existent. Les caractéristiques atypiques des matrices de données d'expression (bruits, dimension) sont également source de nouveaux développements (Romualdi *et al.*, 2003). Aussi, seules quelques méthodes parmi les plus connues, et actuellement les plus performantes, sont présentées.

4.1 K plus proches voisins

La technique des *k* plus proches voisins ou KNN (***K N**earest **N**eighbor*) est la méthode de classification la plus simple. Soit un nombre d'échantillons appartenant à des classes connues, l'échantillon inconnu est associé à la classe qui possède les *k* échantillons qui lui sont le plus proches (similaires).

Deux paramètres sont à définir pour appliquer l'algorithme : *k* ou le nombre d'échantillons voisins rechercher et *l* ou la marge d'erreur. Plus précisément, si *k*=3 et *l* =3, cela signifie que les 3 plus proches voisins doivent être dans la même classes ; si *l* =2 au moins 2 des 3 plus proches voisins doivent être dans la même classe. L'algorithme des KNN peut facilement séparer plusieurs classes. Il peut également distinguer des classes de données non linéaires. Seul inconvénient majeur, cette méthode est sensible aux valeurs aberrantes.

4.2 Classification des centroïdes

La classification par l'analyse des centroïdes (*centroid classifier*) permet une **répartition rapide des données en plusieurs classes**. Pour chaque classe connue, le barycentre est calculé. Ensuite, toutes les distances possibles (le plus souvent distance euclidienne) entre l'échantillon à classer et les différents barycentres des différentes classes sont calculées. L'échantillon inconnu est alors agrégé à la classe pour laquelle la distance au barycentre est la plus faible. Cette approche possèdent deux inconvénients majeurs : elle est sensible aux données bruitées et sépare les données uniquement de façon linéaire.

L'approche par centroïdes « rétrécis » (*shrunken centroid*) est une amélioration de l'analyse des centroïdes, développée par Tibshirani *et al.* (2002). Cette méthode permet de définir les jeux de gènes qui caractérisent au mieux chaque classe. Les centroïdes sont estimés à partir d'un plus petit nombre de gènes et sont donc moins bruités. La prédiction est plus précise. Cette méthode a l'avantage de pouvoir être également appliquée aux approches de classification non supervisée. En effet, elle permet, par exemple, de mettre en évidence les gènes qui contribuent le plus aux regroupements obtenus par une classification hiérarchique. Cette analyse est accessible à partir des logiciels Excel® (macro complémentaire) et R sous les noms de PAM et *pamr* pour ***P**rediction **A**nalysis of **M**icroarrays (for **R**)* (Tibshirani *et al.*, 2002).

4.3 Analyse discriminante linéaire

Contrairement aux techniques de KNN et de classification par les centroïdes, l'analyse discriminante linéaire (LDA - ***L**inear **D**iscriminante **A**nalysis*) est paramétrique. La classification est **basée sur un modèle statistique** issu de données distribuées selon la loi Normale (approximativement). Dans un premier temps, une ligne droite ou un hyperplan est calculé afin de séparer au mieux deux classes connues. Cette **séparation est réalisée de telle sorte que la variation intra-classe soit minimale et que la variation inter-classe** soit

maximale. L'échantillon inconnu est alors positionné dans l'espace et associé à la classe dont il est le plus proche dans le plan.

Il existe de nombreuses variantes de LDA. Elles diffèrent par le mode de calcul du poids attribué à chaque gène. La plus ancienne est la méthode LDA de Fisher. Cette dernière utilise la matrice de corrélation entre toutes les paires de gènes pour les pondérer. Selon Dudoit *et al*. (2002), cet algorithme est peu efficace. Il est sensible à la dimension de la matrice et par conséquent peu stable, entraînant un risque de surapprentissage. L'analyse discriminante linéaire diagonale (DLDA) est un cas particulier de la LDA de Fisher où la corrélation entre les gènes est ignorée. Ceci minimise le nombre de paramètres à estimer et augmente les performances de la méthode. Ainsi, des études ont démontré que cette approche est plus efficace que certaines méthodes plus sophistiquées telles que les réseaux neuronaux ou les méthodes de partitionnements récursifs (Dudoit *et al.*, 2002; Romualdi *et al.*, 2003). Les classifieurs DLDA sont parfois qualifiés de classifieurs "bayésiens naïfs" car ils utilisent une approche bayéssienne pour associer un échantillon à la classe qui possède la probabilité maximale *a posteriori*. Les approches par *weighted voting* (Golub *et al.*, 1999; Alizadeh *et al.*, 2000) et *compound covariates* (Hedenfalk *et al.*, 2001) sont des méthodes similaires qui semble également donner de bons résultats. Cependant, leurs performances restent inférieures à la méthode DLDA (Dudoit et Fridlyand, 2002).

Finalement, de manière générale, les analyses discriminantes linéaires sont des techniques puissantes. Toutefois, elles ne sont pas directement applicables à plus de deux classes et nécessitent également que les données soient séparables linéairement.

4.4 Machines à vecteurs de support

Les machines à vecteurs de support (*SVM- **S**upport **V**ector **M**achine*) correspondent à une évolution majeure des algorithmes de classification supervisée (Vapnik 1998). Le principe d'une SVM est de dessiner une **droite ou un hyperplan afin de séparer au mieux deux ou plusieurs classes d'apprentissage présentes dans l'espace à *n* dimensions** des données d'expression (Fig. 33A). L'objectif est de **maximiser la distance des échantillons aux frontières de l'hyperplan**. Les frontières sont définies par un jeu de données appartenant aux données d'apprentissage. Ces frontières sont les vecteurs de support (*support vectors*) (Fig. 33B). Ils peuvent être définis de manière linéaire ou à partir d'une famille de fonctions (polynômes, *spline*). La distance entre les vecteurs de support des classes, ou frontière,

représente la marge de l'hyperplan. Lorsqu'il n'existe pas d'hyperplan capable de séparer les données, les échantillons sont transposés dans un espace de dimension supérieure.

(A) **(B)**

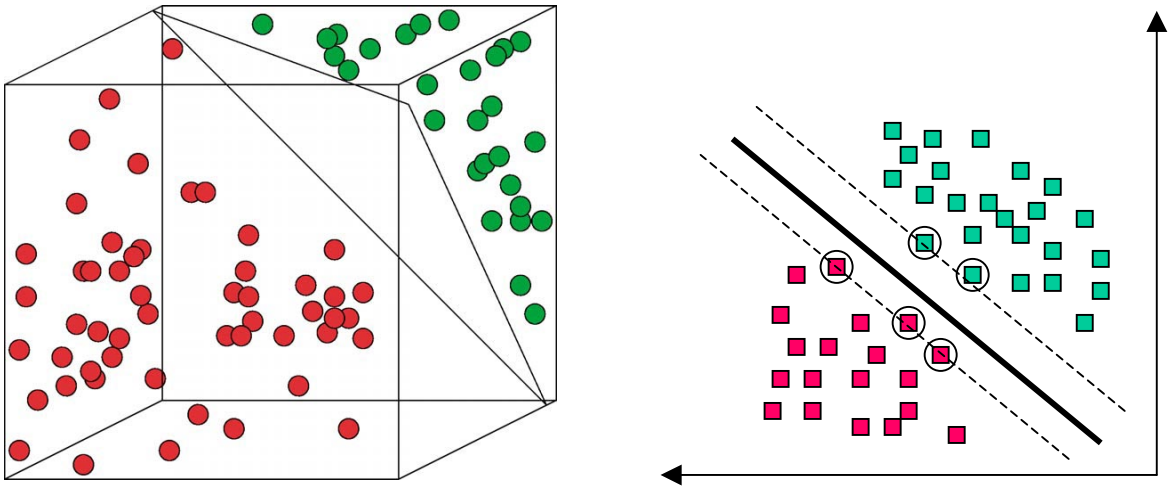


Figure 33. Machine à vecteurs de support, approche binaire. (A) Visualisation dans un espace à 3 dimensions. (B) Visualisation dans un espace à 2 dimensions. L'algorithme sélectionne un hyperplan (A - triangle ; B - ligne accentuée) qui maximise la largeur de l'écart (« marge ») entre 2 classes (points rouges et verts). L'hyperplan est défini par des instances d'apprentissages « frontières » appelées vecteurs supports (B- cercles noirs). Les nouveaux échantillons sont classés selon leur position par rapport à l'hyperplan.

Développées dans un premier temps pour une séparation binaire des données (Brown *et al.*, 2000), différentes approches SVM existent aujourd'hui pour une analyse multi-classes ou *MC-SVM* (***M**ulti**C**ategory **S**upport **V**ector **M**achines*) (Yeang *et al.*, 2001; Ramaswamy *et al.*, 2001; Statnikov *et al.*, 2004). Ces algorithmes donnent généralement des réponses précises. Ils minimisent le risque de surapprentissage permettant ainsi la généralisation de la règle de classification. Ces techniques présentent également l'avantage d'être particulièrement robustes aux problèmes de dimension (Statnikov *et al.*, 2004; Pavlidis *et al.*, 2004). De manière générale, ces techniques sont plus efficaces que les autres algorithmes de classification supervisée tels que la méthode des KNN ou les différentes approches de réseaux neuronaux. Ainsi, Brown *et al.* (2000) ont montré, grâce à une analyse par *SVM,* que des gènes peuvent être classés sur la base de leurs catégories fonctionnelles. Cette étude leur a également permis d'inférer une fonction à des gènes jusqu'alors inconnus. De même, Statnikov *et al.* (2004) démontre que les MC-SVMs peuvent être un outil puissant pour le diagnostic et le pronostic médical. Enfin, l'emploi de ces techniques ne se limite pas à

l'analyse des données d'expression de gènes. En effet, elles ont été utilisées afin de classer des régions promotrices de gènes (Pavlidis *et al.*, 2001), pour reconnaître le mode de repliement de protéines (Ding et Dubchak, 2001) ou encore pour prédire des interactions protéine-protéine (Bock et Gough, 2001).

4.5 Validation des regroupements

L'une des finalités de ces études est la création d'outils décisionnels pour l'aide au diagnostic et au pronostic. Or, dans le domaine médical, nous n'avons pas le droit à l'erreur. Il est donc important de connaître la précision de ces approches.

La méthode la plus classique est le partitionnement aléatoire du jeu de données en un jeu de données d'apprentissage (2/3 des données) et jeu de données test (1/3). Les méthodes de *bootstrap*, *Jackknife, Leave-one-out* et *bagging* (Dudoit et Fridlyand, 2002), présentées pour la validation des méthode de classification non supervisée, sont également applicables aux approches supervisées. La validation croisée par retrait (***L**eave **O**ne **O**ut **C**ross **V**alidation –LOOCV*) est l'approche la plus connue (Simon, 2003) (*cf.* Fig. 32 p. 103). Elle permet notamment d'estimer la confiance dans la prédiction.

En bref

Les méthodes de classification supervisée utilisent de la connaissance *a priori* pour construire des classifications. Elles établissent des règles et un modèle de classification à partir d'un jeu de données connu et annotés, dit jeu d'apprentissage (*training set*), afin de prédire la classification de nouveaux cas appartenant à un jeu de données test (*Class prediction*). Les méthodes se distinguent principalement par leur mode de séparation des données (1 ou plusieurs classes, linéaire ou non).

5. Analyses factorielles

Une matrice de données d'expression de *n* gènes et *m* échantillons représente un espace de données à *n* points et *m* dimensions (ou inversement). Quelque soit la configuration, le caractère multidimensionnel des données est difficile à appréhender. Actuellement, aucun outil de visualisation n'est capable d'analyser un si grand espace. Toutefois, les analyses factorielles, méthodes de statistiques descriptives, permettent de **réduire le nombre de dimensions de l'espace des données** et par conséquent la complexité du problème.

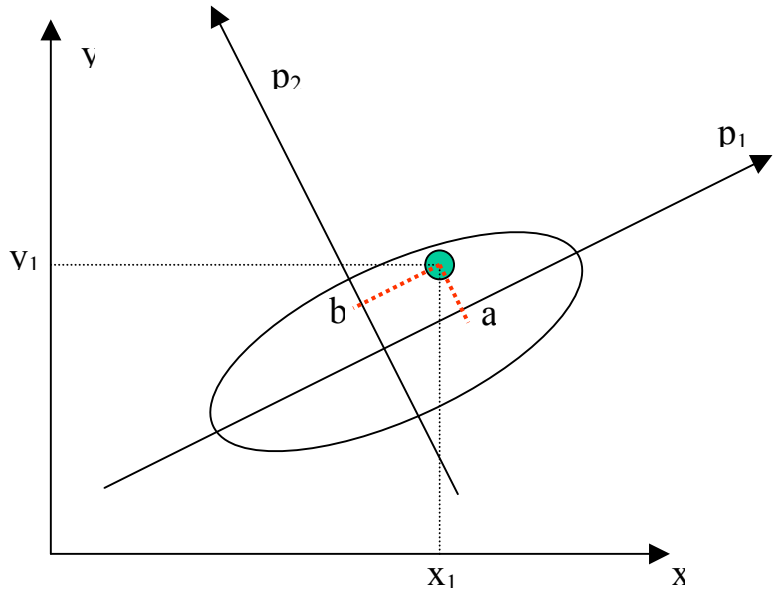


Figure 34. Principe de l'analyse en composante principale. Les composantes principales des données sont les projections de ces données dans un nouveau système de coordonnées. Le point vert, appartenant au nuage de points ovale, a les coordonnées (x_1,y_1) dans le système de coordonnées d'origine x0y. Le premier vecteur propre p_1, représente la direction de la plus grande variance. Le second vecteur propre lui est perpendiculaire. Les composantes principales du point (x_1,y_1) sont (a,b) par projection sur les vecteur propres p_1 et p_2.

Le principe des analyses factorielles est la projection des données dans un espace aux dimensions réduites. Cet espace est construit grâce à une combinaison linéaire des variables qui explique le mieux les variations des données d'origine, l'objectif étant de perdre le moins d'informations possible. Ces techniques reposent sur l'idée qu'une grande partie de la variation des données peut être expliquée par un petit nombre de variables transformées.

Les méthodes factorielles regroupent trois principales techniques, déjà employées dans l'analyse des données d'expression : l'analyse en composantes principales (Raychaudhuri *et al.*, 2000; Alter *et al.*, 2000), l'analyse des correspondances (Fellenberg *et al.*, 2001) et l'analyse des correspondances multiples (Khan *et al.*, 1998; Bittner *et al.*, 2000).

5.1 Analyse en composante principale

L'analyse en composante principale ou ACP (PCA - ***P**rincipal **C**omponent **A**nalysis*) est une **méthode statistique pour l'exploration de données multi-variées,** présentée pour la première fois par en 1933 par Hotelling (Legendre and Legendre, L. 1998). L'ACP est également connue sous les noms « d'analyse par décomposition des données en valeur singulière » (SVD - ***S**ingular **V**alue **D**ecomposition*) (Alter *et al.*, 2000; Holter *et al.*, 2000) dans le domaine du transcriptome ou « *Karhunen-Loève expansion »* en traitement du signal (Alter *et al.*, 2000). Cette méthode sert de fondement théorique aux autres méthodes de statistiques factorielles qui en apparaissent comme des cas particuliers.

L'objectif de l'ACP est de **réduire la dimension de l'espace des données en déformant le moins possible la réalité**. Pour cela, elle détermine une suite d'axes orthogonaux, non corrélés, conservant au mieux les distances entre les individus. Ces axes sont appelés axes principaux d'inertie ou composantes principales (Fig. 34). Les composantes principales sont définies par les vecteurs propres ou *eigenvector*. La conservation des distances (aux données d'origine) par chaque axe est mesurée par la variance des coordonnées des individus sur cet axe, encore appelée valeur propre ou *eigenvalue*.

Mathématiquement, à partir d'une matrice de données d'expression à m observations et n variables, l'ACP calcule un nouveau système de coordonnées (une nouvelle matrice). La matrice de données peut être centrée ou non, normalisée (basée sur la matrice de corrélations) ou non (basée sur la matrice de variance-covariance). Ceci définit les différents types d'ACP et expliquent également les différences de noms (Yeung et Ruzzo, 2001). Les directions du nouveau système de coordonnées sont les vecteurs propres de la matrice de variance-covariance ou de la matrice des corrélations des profils d'expression (Raychaudhuri *et al.*,

2000; Draghici 2003). La nouvelle matrice capture la forme des données. Par exemple, pour un nuage de points ovoïde à *n* dimensions, le vecteur propre de la matrice (composante principale) est le grand axe de l'ovale (Fig.34). Ce premier axe d'inertie est l'axe sur lequel la projection du nuage de points a le plus de variance. L'axe secondaire sera l'axe orthogonal au premier et calculé de la même manière. *p* axes principaux d'inertie sont ainsi calculés par rotation et dans l'ordre décroissant de la variation qu'ils expliquent. Généralement, les composantes principales utilisées sont les 2 ou 3 premières puisqu'elles témoignent des principales variations observées dans le jeu de données original. Les dernières composantes reflètent quant à elles les bruits résiduels. Cependant, Yeung et Ruzzo (2001) ont montré que les premières composantes d'une ACP ne permettent pas nécessairement de mettre en évidence la structure des données, i.e. la répartition en groupes.

Alter *et al*. (2000) ont montré que l'analyse de données d'expression sur le cycle cellulaire de la levure (Spellman *et al.*, 1998) par une approche SVD permet d'émettre des hypothèses sur le fonctionnement de certains régulateurs du cycle et de modéliser leur processus. En effet, ils ont mis en évidence des corrélations entre les profils d'expression de ces régulateurs et leurs cibles. Ils ont ainsi inféré un modèle d'ondes d'expressions ou *wave model* pour expliquer les différentes phases du cycle cellulaire. De la même manière, ils ont montré qu'il est possible de comparer le transcriptome de différents organismes par une approche généralisée de la SVD ou GSVD (***G****eneralized* ***S****ingular* ***V****alue* ***D****écomposition*) (Alter *et al.*, 2003). De plus, L'ACP (et ses variantes) peut s'avérer utile dans le choix du nombre de classes à définir *a priori* pour l'application des algorithmes de classification supervisée. Elle peut également être à la base de certains de ces algorithmes (Bicciato *et al.*, 2003). Toutefois, selon Yeung et Ruzzo (2001), l'ACP ne forme pas de groupe et encore moins de classification. Ils déconseillent l'utilisation directe des vecteurs propres (et ou valeurs propres) dans les algorithmes de classification. Selon ces auteurs, ils n'améliorent pas, voire dégradent, la qualité des regroupements.

5.2 Analyse factorielle des correspondances

L'analyse factorielle des correspondances ou AFC (*CA* **-** ***C****orrespondance* ***A****nalysis*), proposée par Benzécri dans les années 60, est une méthode exploratoire pour analyser des données qualitatives (tableaux de contingence, présence-absence, enquête). L'objectif est de recherche et d'étudier les associations entre variables. Tout comme l'ACP, elle représente les données dans un espace de dimension réduite, encore appelé carte. Elle permet de visualiser

les paramètres (variables gènes) et les objets (variables échantillons) dans le même espace, mettant en évidence d'éventuelles dépendances entre les deux.

Tout d'abord développée pour l'analyse des tableaux de contingence, l'AFC utilise la statistique du *Chi2*[38] pour évaluer le degré d'homogénéité des données. La valeur de la statistique est élevée lorsque la relation est étroite entre les paramètres (lignes) et les objets (colonnes) de la matrice (Fellenberg *et al.*, 2001). Graphiquement, les points sont représentés de telle sorte que la somme des distances à leur centroïde (aussi appelée inertie globale) soit proportionnelle à la valeur de *Chi2* du tableau de données. Cette distance est petite lorsque les profils de 2 vecteurs présentent la même tendance, indépendamment de la valeur absolue de leur expression. Grâce à l'AFC, Fellenberg *et al* (2001) ont mis en évidence des gènes associés à certaines phases du cycle cellulaire de la levure.

5.3 Positionnement multidimensionnel

Proposé par Shepard et Kruskal dans les années 60 (Legendre and Legendre, L. 1998), le positionnement multidimensionnel (MDS - ***M****ultidimensional* ***S****caling*) est une méthode d'analyse de données largement utilisée dans les domaines du marketing et de la psychométrie (particulièrement dans les pays anglo-saxons). Tout comme l'ACP, le principe de la méthode est de construire une représentation des individus dans un espace de dimension réduite. Toutefois, contrairement à l'ACP, la matrice de départ est une matrice de similarités/dissemblances (euclidienne, corrélation de Pearson etc.) d'où parfois le nom d'ACP de tableau de distances. Ceci permet d'établir des relations non linéaires entre les individus (Bittner *et al.*, 2000). Graphiquement, les échantillons sont représentés dans un espace euclidien à 2 ou 3 dimensions. Ceci permet d'estimer et visualiser le degré de corrélation entre les objets étudiés. Ainsi, Bittner *et al.* (2000) ont montré que la méthode MDS, couplée à des méthodes de classification, permet de distinguer des catégories de mélanomes jusqu'à présent non identifiées sur la base de critères cliniques. Cette technique permet donc d'estimer le nombre de classes à définir *a priori* pour l'application des algorithmes de classification supervisée. Elle peut également aider au choix de la distance à utiliser lors de l'application d'un algorithme de regroupement.

[38]La statistique du *Chi2* évalue l'importance des écarts entre des fréquences d'occurrence (ou des pourcentages) observées à l'intérieur d'échantillons aléatoires et des fréquences (ou des pourcentages) théoriques espérées qui devrait être observées si l'hypothèse nulle soumise au test était vraie.

En bref

Les analyses factorielles sont des méthodes de statistiques descriptives. Elles permettent de réduire le nombre de dimensions de l'espace des données et par conséquent la complexité du problème. Ces techniques reposent sur l'idée qu'une grande partie de la variation des données peut être expliquée par un petit nombre de variables transformées à partir des données.

6. Quelle(s) technique(s) de classification ? Quel(s) Outil(s) d'analyses et de visualisation ?

6.1 Quelle(s) technique(s) de classification ?

Les algorithmes de classification se sont montrés particulièrement efficaces pour regrouper et classer les gènes comme les échantillons. Au travers de l'analyse des profils d'expression, des fonctions inconnues de gènes ont pu être inférées (Eisen *et al.*, 1998) et la classification de certaines maladies a pu être affinée (Golub *et al.*, 1999; Alizadeh *et al.*, 2000).

De nombreuses études comparatives ont été réalisées afin de déterminer l'algorithme de classification de plus performant (Yeang *et al.*, 2001; Dudoit et Fridlyand, 2002; Romualdi *et al.*, 2003; Statnikov *et al.*, 2004). Actuellement aucun consensus n'est établi (Tab. 12).

Parmi les méthodes de classification non supervisée, la classification hiérarchique ascendante est certainement la technique la plus simple. Elle offre rapidement une vue d'ensemble des données. Elle permet de détecter des groupes de gènes et de patients et peut aider à la définition du nombre de groupes attendus dans les algorithmes de partitionnement ou les techniques de classification supervisée. Engin, grâce à la représentation en carte thermique, elle permet parfois de mettre en évidence des biais expérimentaux.

Concernant les approches supervisées, il semble que les méthodes SVMs soient particulièrement intéressantes (Yeang *et al.*, 2001; Statnikov *et al.*, 2004). Elles permettent de discriminer des données non linéairement séparables et certaines de ces approches offrent la possibilité de définir plusieurs classes.

Enfin, les analyses factorielles sont des méthodes de réduction des dimensions de l'espace des données qui facilitent leur visualisation. Ces méthodes ne sont pas des techniques de classification mais des méthodes descriptives. Elles peuvent toutefois être à la base de certaines techniques de classification (*biclustering*). Elles sont tout de même à utiliser avec

précautions : elles peuvent améliorer comme détériorer les performances de ces algorithmes de classification (Romualdi *et al.*, 2003).

Tableau 12. Avantages et inconvénients des principales méthodes de classification, exemples d'applications sur les données d'expression.

Méthodes	Exemples d'application	Avantages/Inconvénients de l'algorithme
Classification non supervisée		
Classification hiérarchique ascendante (agglomérative) (HCL)	Regroupement à partir des éléments individuels (Eisen *et al.*, 1998)	▪ Intuitive ▪ Déterministe ▪ Implémenté dans de nombreux outils ▪ Résultats dépendant des métriques et distances choisies ▪ Sensibles aux valeurs aberrantes
Classification hiérarchique descendante (divisive)	Regroupement à partir de l'ensemble des éléments (Alon *et al.*, 1999; Kerr et Churchill, 2001a)	▪ Intuitive ▪ Non déterministe ▪ Résultats dépendant des métriques et distances choisies ▪ Sensibles aux valeurs aberrantes
K-moyennes (*K-means*)	Partition des gènes/expériences en un nombre connu de groupes par la recherche de centroïdes (Tavazoie *et al.*, 1999)	▪ Intuitive ▪ Implémenté dans de nombreux outils ▪ Nécessité de spécifier le nombre de groupes attendus ▪ Non déterministe
Cartes auto-organisatrices (SOM)	Partition des gènes/expériences en un nombre connu de groupes par association à des nœuds (Tamayo *et al.*, 1999; Toronen *et al.*, 1999)	▪ Implémenté dans de nombreux outils ▪ Nécessité de spécifier le nombre de groupes attendus ▪ Non déterministe

Classification supervisée		
K plus proches voisins (KNN)	Recherche de gènes impliqués dans la voie oestrogenique des souris C2C12 (Theilhaber *et al.*, 2002).	▪ Intuitive ▪ Rapide (temps calcul) ▪ Capacité à séparer des données non linéaires ▪ Classification extensible à plus de 2 classes ▪ Sensibles au valeurs aberrantes
Classification par centroïdes	Mise en évidence de gènes caractérisant au mieux chaque classe pour une meilleure définition des classes (Tibshirani *et al.*, 2002)	▪ Intuitive ▪ Rapide (temps calcul) ▪ Classification extensible à plus de 2 classes ▪ Séparation de données linéaires uniquement
Analyse linéaire discriminante (LDA)	(Dudoit et Fridlyand, 2002)	▪ Basée sur un modèle statistique ▪ Puissante ▪ Séparation de données linéaires uniquement ▪ Pas directement applicable à plus de deux classes
Machine à vecteurs de support (SVM)	Algorithme d'apprentissage qui incorpore des données externes (Ramaswamy *et al.*, 2001)	▪ Implémenté dans de nombreux outils ▪ Capacité à séparer des données non linéaires ▪ Pas directement applicable à plus de deux classes
Analyses factorielles		
Décomposition en valeur singulière (SVD/PCA) (similaire à l'analyse en composante principale)	Réduction de la matrice des données des gènes/échantillons par d'expression en vecteurs propres et possibles mise en évidence de regroupements (Alter *et al.*, 2000; Holter *et al.*, 2000)	▪ Réduction de l'espace des données, peut faciliter la visualisation des données ▪ Pas de séparation directe des données en classes
AnalyseFactorielle des correspondances	Réduction de l'espace des données (carte), visualisation des variables et des échantillons dans un même espace et mise en évidence d'éventuelles dépendances. (Fellenberg *et al.*, 2001)	▪ Réduction de l'espace des données, peut faciliter la visualisation des données
Positionnement multidimensionnel	Classification des gènes dans l'ordre de leur capacité à minimiser le volume des groupes et maximiser la distance inter-groupe (de centre à centre) (Bittner *et al.*, 2000)	▪ Réduction de l'espace des données, peut faciliter la visualisation des données

6.2 Quel(s) Outil(s) d'analyses et de visualisation ?

Il existe une gamme importante d'outils pour l'application des méthodes de classification et la visualisation de leurs résultats. L'un des premiers ensembles logiciels utilisé pour les analyses par classification hiérarchique ascendante a été Cluster-TreeView de Eisen *et al*. (1998). Cluster propose les méthodes de classification hiérarchique ascendante, *k*-moyennes et ACP. TreeView permet la visualisation des résultats sous la forme d'une carte thermique (*heatmap*). Ce produit académique a fait de cette représentation une référence. Certains auteurs parlent même d'« Eisengramme ». Aujourd'hui, de nombreuses suites logiciels, académiques ou commerciales, proposent une implémentation des principales méthodes de classifications (classification hiérarchique, *k*-moyennes, SOM, SVM...), couplées à diverses représentations graphiques (cartes thermiques, *biplot,* graphique en 3 dimensions) [39,40] facilitant l'interprétation.

Les outils d'analyses se répartissent en deux grandes catégories : les logiciels intégrés et les solutions possédant des environnements de développement. Les logiciels intégrés, tels Genesis[41] (Sturn *et al.*, 2002), J-express[42] (Dysvik et Jonassen, 2001) ou GeneSpring®, sont relativement simples d'utilisation et possèdent généralement des fonctionnalités graphiques conviviales (interactives, dynamiques). Toutefois, ces outils sont peu flexibles car ils ne permettent pas l'ajout de méthodes d'analyses par les utilisateurs. A l'inverse, les solutions possédant un environnement de développement offrent de nombreuses possibilités. Elles permettent le développement de méthodes et de procédures d'analyse ainsi que l'ajout de modules (*plug-ins*) créés par la communauté des utilisateurs de ces outils. Ces logiciels sont par exemple R/BioConductor, SAS® ou Matlab®. Ces outils, dédiés en premier lieu aux analyses mathématiques et statistiques, s'avèrent très puissants pour l'analyse des données génomiques. Cependant, ils nécessitent une certaine expertise et sont globalement moins souples d'utilisation (script en ligne de commande, conception des graphiques peu interactive...). Des solutions commerciales intermédiaires et très puissantes existent. Par exemple, Spotfire® ou Rosetta Resolver® offrent la possibilité d'implémenter et d'exécuter des scripts R dans leur environnement d'analyse. Toutefois, ces logiciels sont généralement très coûteux en terme d'investissement et de mise à jour. Les produits académiques, bien que

[39] http://genopole.toulouse.inra.fr/bioinfo/microarray
[40] http://genomicshome.com
[41] http://genome.tugraz.at/Software
[42] http://www.molmine.com

parfois moins conviviaux, sont généralement plus souples et plus rapides en terme d'évolution.

III. Intégration des méta-données

Au cours des 30 dernières années, les chercheurs ont privilégié une approche réductionniste de l'analyse des mécanismes biologiques. Ils se sont focalisés soit sur un petit nombre d'objets biologiques (un gène, une protéine...) soit sur une population d'éléments (le génome, le transcriptome, le protéome...). Si l'analyse du transcriptome par la technologie des puces à ADN offre un aperçu des « corrélations » entre les gènes et les phénomènes biologiques (*« guilty by association »*), elle ne permet pas à elle seule de révéler la **causalité des mécanismes de régulation** (Quackenbush, 2003). Aussi, **l'intégration des méta-données** (*metadata*), i.e. des informations issues de différentes sources contrôlées comme les ontologies, les résumés d'articles scientifiques ou les banques de données protéiques (Balasubramanian *et al.*, 2004), est devenue indispensable pour interpréter les données issues des expériences de transcriptomique.

Généralement disponibles sous la forme de bases de données publiques, les méta-données sont multiples et de qualités hétérogènes. Des approches systématiques, associées à des outils bio-informatiques, sont nécessaires pour analyser et intégrer l'information de ces bases de données.

1. Ontologies pour la génomique

La biologie est un domaine qui manque encore de formalisme strict. Malgré les efforts du consortium HGNC (*HUGO Gene Nomenclature Committee*) pour standardiser la nomenclature des gènes, des améliorations étaient encore nécessaires pour définir les fonctions des gènes et de leurs produits (Ashburner *et al.*, 2000). Ceci a incité la communauté scientifique à développer **des ontologies pour annoter les gènes et leurs produits**.

« Ontologie est un terme philosophique qui veut dire « doctrine de l'être ». La science déploie une problématique ontologique lorsque vient à se poser la question du statut de la réalité des entités qui constituent le référent du discours scientifique » (Encyclopedia Universalis, 1991). Par extension, une ontologie est un vocabulaire structuré et contrôlé qui est une base au développement de la connaissance.

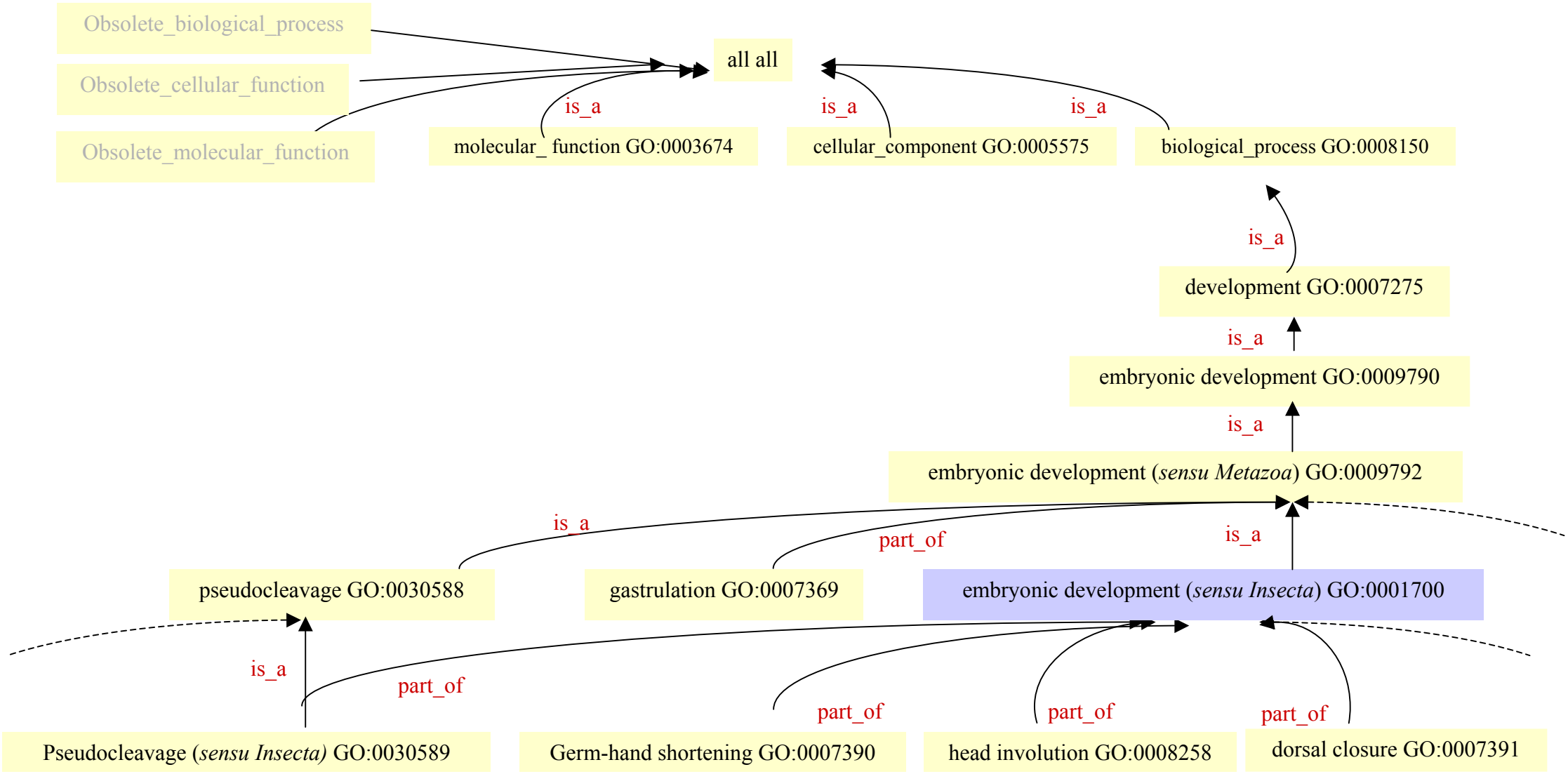


Figure 35.Extrait du graphe orienté acyclique de Gene Ontology. Le graphe a pour origine les trois ontologies *molecular function, cellular function, biological process*. Les mentions « is_a » et « part_of » indiquent le type de relation entre les termes parents et enfants. Le terme « embryonic development (*sensu Insecta*) GO :0001700» présente une limite de GO, à savoir la nécessité de définir des catégories spécifiques à certaines espèces. Les termes obsolètes sont matérialisés par les catégories « obsolete_molecular_function », « obsolete_cellular_function », « obsolete_biological_process».

1.1 Gene Ontology

a) Définition

Gene Ontology[43], communément appelé GO, a récemment été développé par le *Gene Ontology Consortium*, groupe de travail international basé à l'EBI, pour aider à l'annotation des génomes (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2001). **Son objectif est d'établir un vocabulaire structuré, contrôlé et dynamique pour décrire le rôle des gènes et produits de gènes de l'ensemble des eucaryotes**.

b) Composition et structure

GO se compose de **trois ontologies** qui définissent les **processus biologiques**, les **fonctions moléculaires** et la **localisation cellulaire** des produits de gènes. Le processus biologique fait référence à l'objectif biologique auquel un gène ou produit de gène participe (*e.g.* la croissance cellulaire ou la transduction du signal). Un processus biologique est le résultat d'une ou plusieurs fonctions moléculaires associées dans un ordre donné. La fonction moléculaire décrit l'activité biochimique ou l'action du produit d'un gène (*e.g.* enzyme transporteur, ligand). La localisation cellulaire présente l'endroit de la cellule où se trouve la forme active du produit d'un gène. GO permet d'exprimer les phénomènes de régulation et offre la possibilité de représenter des données incomplètes. Enfin, GO est un vocabulaire dynamique dont le nombre de termes évolue rapidement, avec actuellement quelques 17700 termes (Fig. 35) (le 02/04/05).

Les trois ontologies GO sont structurées sous forme d'un **graphe orienté acyclique** ou **DAG** (***D****irected* ***A****cyclic* ***G****raph*) (Fig. 35). Ce DAG est un réseau où chaque nœud représente un terme GO identifié sous la forme *GO : nnnnnnn*. Chaque terme GO peut être un « enfant » de un ou plusieurs « parents ». En effet, la majorité des gènes est pléiotrope, c'est-à-dire qu'un gène peut avoir plusieurs produits et les produits d'un gène possèdent une ou plusieurs fonctions biochimiques. Le terme « enfant » est toujours plus spécifique que le ou les termes « parents ». La relation entre un enfant et son « parent » peut être du type « est un(e) » (*is_a*), identifié par « % », lorsque le terme enfant est une instance du terme parent (Fig. 35). Elle peut aussi être de la forme « fait parti de » (*part_of*), représentée par « < », si le terme enfant est un élément du parent. Si un terme à plusieurs « parents », il peut avoir différentes relations avec chacun de ses « parents ». Chaque ontologie est multidimensionnelle.

[43] http://www.geneontology.org

Néanmoins, une des contraintes de GO est le respect de la règle « *True Path Rule* » (The Gene Ontology Consortium, 2001): si le terme « enfant » décrit un produit de gène alors tous ses termes parents doivent aussi pouvoir s'appliquer sur ce même produit de gène. Malgré la volonté du consortium GO d'établir une ontologie générique, valable pour tous les eucaryotes, des termes spécifiques à certaines espèces ou groupe d'espèces ont dû être définis pour respecter cette règle. Ainsi GO : 0001700 *embryonic development* (*sensu Insecta*) ne peut être utilisé que pour les insectes (Fig. 35).

Enfin, pour chaque annotation, un code qualifie le type d'évidence utilisé pour l'annotation (Tab. 13). La qualité de l'annotation dépend souvent de son origine : une annotation issue de références scientifiques a généralement plus de poids qu'une annotation par prédiction automatique.

Tableau 13. Origine de l'annotation des catégories *GeneOntology* (d'après *www.geneontology.org*).

Symbole	Evidence de l'annotation
IMP	Inferred from mutant phenotype
IGI	Inferred from genetic interaction
IPI	Inferred from physical interaction
ISS	Inferred from sequence similarity
IDA	Inferred from direct assay
IEP	Inferred from expression pattern
IEA	Inferred from electronic
TAS	Traceable author statement
NAS	Non-traceable author statement
ND	No biological data available
IC	Inferred by curator

c) Applications

La première application de GO est l'**annotation des génomes**. La base de donnée GO n'est pas constituée des produits des gènes mais uniquement des termes et concepts (association de termes) qui les caractérisent. Dans le but d'annoter les génomes, des correspondances (*mapping*) entre les termes GO et les gènes ou produits de gènes sont proposées par les différentes bases de données associées au projet[44] telles Flybase, SGD ou

[44] http://www.geneontology.org

GOA (Camon *et al.*, 2004). GOA[45] (*The **G**ene **O**ntology **A**nnotation*) est notamment un projet dirigé par EBI et le groupe SWISSPROT. Son but est de fournir, pour l'ensemble des organismes, une correspondance entre les termes GO et les produits de gènes (protéomes) référencés dans UniProt[46].

L'avantage d'un vocabulaire contrôlé est de pouvoir être interprété par une machine. Par conséquent, l'annotation des gènes peut se faire automatiquement grâce à des outils de bio-informatique. Afin de visualiser les annotations, le consortium GO propose AmiGO![47], une application Web qui permet d'explorer les liaisons entre GO et les bases de données associées au projet. Les requêtes peuvent être faites à partir des termes GO ou des produits de gènes. Dans le même esprit, certaines grandes bases de données, comme EBI, proposent des outils de recherche des termes GO sur leur propre base de données. Des éditeurs d'ontologies spécifiques, tels DAG-édit ou COBrA, ont également été développés pour faciliter la navigation dans les différentes ontologies (Aitken *et al.*, 2004). Enfin, des versions allégées de GO (*GO slims*) ont même été développées pour une annotation rapide des génomes ou d'une collection de cDNA.

Une seconde utilisation de GO est l'**interprétation des données de puces à ADN** et la mise en évidence de catégories fonctionnelles significativement représentées dans des profils d'expression. Dans ce but, les termes GO associés aux gènes estimés « différentiels » peuvent être comparés à l'ensemble des catégories fonctionnelles du génome étudié ou aux gènes présents sur la puce (Draghici 2003). Il existe de plus en plus d'outils[48] (plus ou moins sérieux) pour interpréter les données de puces à ADN *via* GO. Ces outils se distinguent par les données d'entrée et leur formatage, les organismes supportés, les méta-données utilisées (données d'expression, localisation chromosomique…), l'emploi ou non de statistiques et le type d'application (Pasquier *et al.*, 2004). Les outils les plus connus sont certainement GOMiner (Zeeberg *et al.*, 2003), Onto-Express (Draghici *et al.*, 2003a) et Fatigo (Al Shahrour *et al.*, 2004). Des modules d'analyse statistiques de GO sont également développés au sein du projet BioConductor (*e.g. GoStat*).

Enfin, GO offre la possibilité de caractériser et comparer la composition des puces à ADN en terme de catégories fonctionnelles. Draghici *et al.* (2003), grâce à leur outil

[45] http://www.ebi.ac.uk/GOA
[46] http://www.expasy.uniprot.org
[47] http://www.godatabase.org/cgi-bin/amigo/go.cgi
[48] http://www.geneontology.org /GO.tools.html

OntoCompare[49], ont ainsi montré que différentes puces commerciales dédiées à une même question biologique ne recouvrent pas nécessairement les mêmes catégories GO. Aussi, selon la question biologique posée, le contenu d'une puce peut être plus pertinent qu'un autre.

d) Limites et évolutions

GO est devenu, « **malgré lui », un standard** pour l'annotation des génomes. Aujourd'hui, la majorité des travaux réalisés avec la technologie des puces à ADN présente une annotation de leurs résultats avec GO. Cependant, GO possède de nombreuses limites.

Toute d'abord, **GO est construite de manière subjective.** La structure des catégories est établie manuellement, d'où un risque important d'erreurs. En effet, le nombre de termes obsolètes est impressionnant (~ 1000 termes). De plus, sa complexité et sa taille croissante entraînent des **difficultés dans sa maintenance**. La fréquence des mises à jour (quotidienne) rend difficile toute comparaison entre les différentes bases de données et/ou les outils GO. Aussi, dans le cadre de ***G**ene **O**ntology **N**ext **G**eneration project* (GONG), **la tendance actuelle est à l'utilisation de langages formels de représentations de connaissances** comme Protégé-2000 (Yeh *et al.*, 2003) ou DAML+OIL (Wroe *et al.*, 2003) pour une meilleure base logique de description. En effet, DAML+OIL (*DARPA Agent Markup Language + Ontology Inference Layer*) est un langage formel de description avec un modèle objet pour la classification de concepts et l'instanciation d'objets. Il permet une vérification de la cohérence des ontologies et vise à développer leur interopérabilité dans le cadre d'un « Web Sémantique » (marquage sémantique des sources Web en utilisant des ontologies) (Golbreich *et al.*, 2002).

Ensuite, le résultat de la mise en correspondance des termes GO et des gènes est un **réseau statique**. Il ne permet pas de visualiser la notion d'espace et de temps (Fraser et Marcotte, 2004). Ce réseau représente également une vision moyennée de l'expression des gènes obtenue dans différentes cellules, tissus et conditions. En effet, si chaque nœud du réseau correspond à un gène, les relations entre les nœuds (gènes) peuvent être issues de l'analyse d'un échantillon « sain » comme d'un mutant. Dans le but de résoudre certains de ces problèmes, Fraser et Marcotte (2004) proposent une analyse systématique de différents jeu de données pour une description probabiliste de l'annotation des gènes. Par cette approche, ils pensent également expliciter les interactions entre les gènes.

[49] http://vortex.cs.wayne.edu/projects.htm

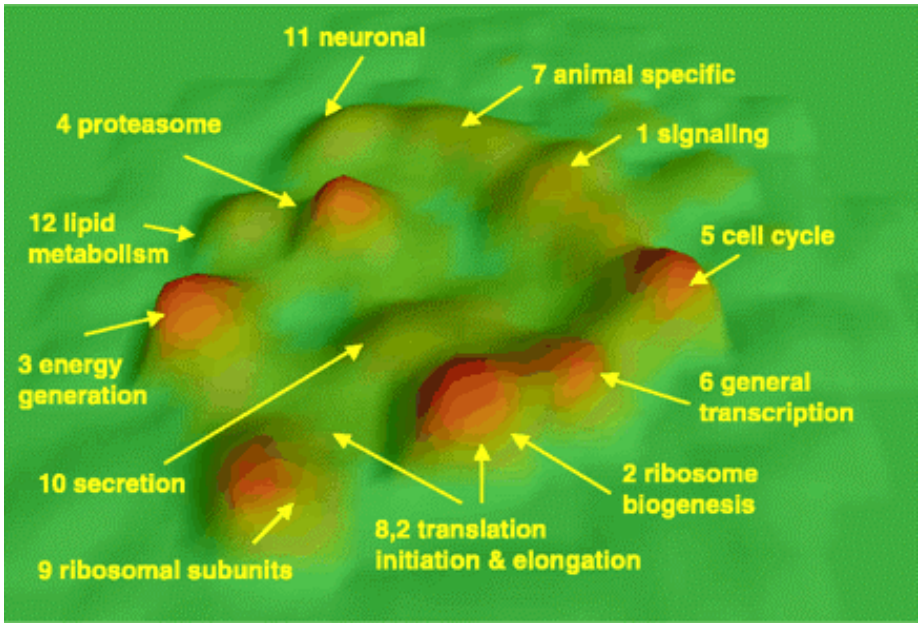


Figure 36. Carte topographique des *metagenes* d'après Stuart *et al.* (2003) et le logiciel VxInsight (Davidson *et al.*, 1998). Visualisation en 3 dimensions de 3416 *metagenes* (un *metagene* correspond à un groupe de gènes orthologues dans différentes espèces). Ces *metagenes* se répartissent 12 catégories fonctionnelles principales. Les sommets (rouge) témoignent d'une forte densité de *metagenes* dont la co-expression est fortement conservée au cours de l'évolution.

Enfin, l'une des difficultés dans l'interprétation de GO est **le niveau de spécificité à retenir** pour annoter les gènes. En effet, GO est structuré en 5 niveaux allant du plus global (*e.g.* all_all) au plus spécifique (*e.g.* GO:0035096, *larval midgut cell death*). Or pour un produit de gène donné, il est souvent délicat de choisir le niveau le plus informatif. Aussi, Stuart *et al*. (2003) introduisent la notion de « *metagene* » qu'ils définissent comme un groupe de gènes orthologues entre différentes espèces étudiées. Ce concept leur permet de définir les principaux processus biologiques conservés entre différentes espèces au cours de l'évolution (Fig. 36). Ils généralisent ainsi la classification GO à 12 grandes catégories. L'application de ce concept a également permis d'inférer des fonctions biologiques à des gènes non encore annotés (Stuart *et al.*, 2003; Puthier *et al.*, 2004).

1.2 Autres ontologies

GO appartient au **répertoire OBO**[50] (***O**pen **B**iological **O**ntologies*) dont l'objectif est de regrouper, en un site Web, les ontologies publiques et les vocabulaires contrôlés couvrant les domaines de la biologie. Ces ontologies, dont le nombre ne cesse de s'accroître (actuellement ~40), sont plus ou moins spécifiques. Pour exemple :

(i) SO (***S**equence **O**ntology*), appartenant également au projet GO, a pour but la description des séquences,

(ii) EVOC (*Expressed Sequence Annotation for Humans)* vise à définir les séquences exprimées dans des conditions d'expériences données (plate forme de puces à ADN, mode de préparation des tissus, traitements …)

(iii) MGED ontology working group[51] s'attache à standardiser l'annotation des expériences de puces à ADN (Stoeckert, Jr. *et al.*, 2002).

MGED ontology working group coordonne également ses développements avec MAGE-ML qui vise à établir un format standard pour l'échange des données issues des expériences de puces à ADN.

PANTHER/X *ontology* est également une ontologie pour la description des fonctions protéiques (Thomas *et al.*, 2003; Mi *et al.*, 2005). Elle se compose de trois catégories « fonction moléculaire », « processus biologique » et « voie biochimique ». Les catégories "fonction moléculaire" et "processus biologique" sont des DAG similaires aux classes GO du

[50] http://obo.sourceforge.net
[51] http://mged.sourceforge.net/ontologies/index.php

même nom. Elles sont toutefois largement simplifiées pour faciliter l'analyse et l'intégration d'un grand nombre de données. Le module « voie biochimique » est une représentation hiérarchique des objets.

En bref

Les ontologies pour la génomique sont des vocabulaires structurés et contrôlés pour décrire les objets biologiques tels que les séquences génomiques, les gènes ou encore les produits de gènes.

Le répertoire d'ontologies biologiques OBO (*Open Biological Ontologies*) contient notamment GO (*Gene Ontology*) devenu, « malgré lui », un standard pour l'annotation des génomes. GO se compose de trois ontologies qui définissent les processus biologiques, les fonctions moléculaires et les localisations cellulaires des produits de gènes. Les premières applications de GO sont l'annotation des génomes et l'interprétation des données de puces à ADN. Si GO possède le mérite de formaliser l'annotation des gènes et de leur produit, il présente quelques limites : sa construction est manuelle et par conséquence subjective, sa taille croissante entraîne des difficultés de maintenance et, enfin, son réseau reste statique. Aussi, la tendance actuelle est à l'association de plusieurs ontologies et à l'utilisation de langages formels de représentations de connaissances.

2. Littérature

La littérature (publications, revues, livres, rapports...) est l'une des premières sources d'informations scientifiques (parmi les banques de données de séquences, d'expression…). Son analyse est donc l'une des premières approches pour extraire de la connaissance des gènes.

2.1 Méthodologies

La fouille de textes (*text mining)* vise à automatiser l'analyse des textes écrits en langage naturel (non structuré) pour (re-)découvrir de l'information et de la connaissance dispersées. Ces techniques de traitement concernent l'ensemble des méthodes capables de convertir les documents bruts (oraux, manuscrits ou électroniques) en information exploitable par l'homme ou la machine.

Dans le domaine biomédical, l'analyse des documents électroniques nous intéresse plus particulièrement (Shatkay et Feldman, 2003). En effet, la principale source de documents scientifiques se trouve sur le Web avec notamment la base de données bibliographiques MEDLINE[52] et son moteur de recherche PubMed. MEDLINE référence à elle seule près de 15,000,000 articles publiés, depuis les années 50, dans plus de 4 600 journaux différents. PubMed permet d'explorer MEDLINE et de présenter sous la forme de citations ou de résumes les résultats d'une requête.

Les techniques d'analyse des documents électroniques sont plus ou moins complexes. Elles vont de l'analyse statistique des **co-occurrences de termes** (ou mots clés) (Chaussabel et Sher, 2002; Chaussabel *et al.*, 2003) à l'utilisation des techniques d'**analyse du langage naturel**, en passant par l'utilisation de modèle de Markov caché pour l'identification et la classification de termes. Chaussabel et Sher (2002) montrent ainsi qu'il est possible d'annoter, voire d'inférer de la connaissance, sur les gènes grâce à la recherche de co-occurrences de termes dans les résumés des articles référencés dans MEDLINE. Leur approche consiste à extraire les termes les plus fréquemment employés pour qualifier des gènes étudiés et construire une matrice de co-occurrence de termes (Fig. 37). Cette matrice est ensuite combinée et comparée à une matrice de données d'expression de gènes. Des hypothèses peuvent alors être émises quant à la fonction des gènes qui ne sont pas associés à des mots clés mais qui sont co-exprimés avec des gènes annotés.

[52] http://www.pubmed.org

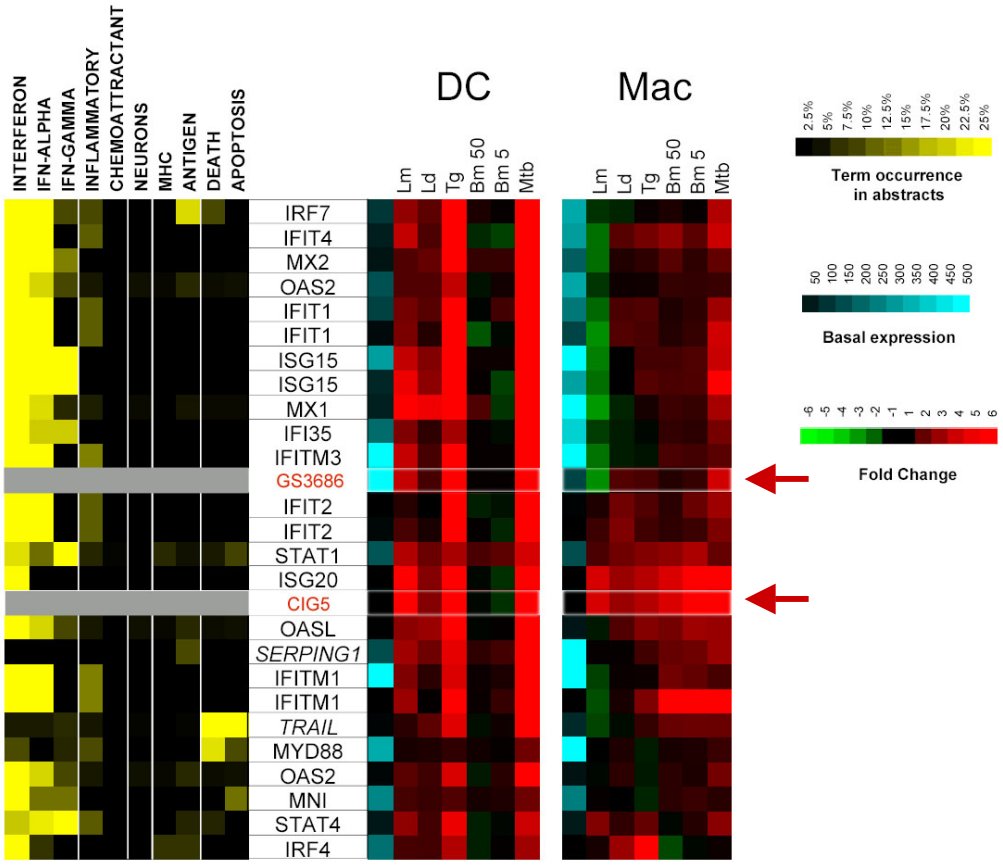


Figure 37. Interprétation des données d'expression de puces à ADN par leurs fréquences de co-occurrence avec des termes issus des résumés de Medline (d'après Chaussabel *et al.*, 2003). Des puces Affymetrix® humaines (HU95) ont été utilisées pour étudier le niveau d'expression des gènes dans des cellules dendritiques (DC) et des macrophages (Mac) stimulés par différents agents pathogènes (Lm, *Leishmania major* ; Ld, *leishmania donova* ; Tg, *taxoplasma gondii* ; Bm50(5) ; *Brugia malayi* 50(5) larves ;Mt, *Micobacterium tuberculosis*). Les niveaux d'intensités des gènes sont corrélés avec leur niveau d'expression de base (échelle de bleus), leur amplitude de variations (échelle du vert au rouge), et leur fréquence d'occurrence avec des termes de la littérature (échelle de jaune). Les gènes listés en rouge (GS3686, CIG5) n'appartiennent pas au groupe d'origine et ne sont pas annotés. Leur classification sur la base de leur profil d'expression suggère une liaison avec les mécanismes de fonctionnement des interférons.

2.2 limites

Dans le domaine de la biologie, la fouille de texte est particulièrement complexe et difficilement entièrement automatisables.

La première difficulté dans l'automatisation est la reconnaissance des noms de gènes dans les résumés. En effet, un nom de gène possède le plus souvent plusieurs synonymes et,

malgré la mise en place d'une nomenclature officielle par HGNC[53], il reste de nombreuses ambiguïtés. Par exemple, le NPPA, gène codant pour le peptide précurseur A du facteur natriurétique, est aussi appelé ANP, *natriuretic peptide precursor A*, *atriopeptin*, *cardionatrin*, CDD-ANF, PND, *pronatriodilatin*, *atrial natriuretic factor*, *atrial natriuretic polypeptides*, *atrionatriuretic factor* et *atrial natriuretic peptide* (source :MADSENSE[54] *cf.* p37, 119).

Le second problème est la mise en évidence des relations entre plusieurs gènes co-cités dans un résumé. L'extraction automatique des règles de grammaire, telles que la recherche de pronom, d'anaphore ou d'apposition, est rapidement source d'erreurs. Enfin, l'analyse du contexte dans lequel interviennent ces relations (tissus, organes, conditions expérimentales) est également très important et encore plus complexe à analyser.

2.3 Outils

L'un des premiers outils développé pour associer données d'expression et la littérature a été MedMiner (Tanabe *et al.*, 1999). Ce service Web propose de lier les gènes présents sur une puce à ADN et des concepts (*e.g.* contraction musculaire) avec les informations disponibles sur GeneCards[55] et PubMed. Grâce à GeneCards, il évalue si les gènes de la puce appartiennent aux concepts et s'il existe d'autres gènes associés à ces concepts. Au travers de PubMed, MedMiner sélectionne et « digère » les résumés concernant l'association des gènes et des concepts. La principale limite de cet outil est la difficulté pour l'utilisateur de définir les concepts. En effet, un concept doit correspondre à un ensemble de mots clés (*e.g.* inhibe, réprime…) qui doit permettre de retrouver l'ensemble des articles pertinents. Un trop petit nombre de mots clés entraîne le risque de perte d'information tandis q'une liste trop importante noie l'information.

Afin de faciliter l'analyse, de nombreux auteurs proposent d'associer les données de la littérature à des thésaurus de termes et de concepts biologiques (Jenssen *et al.*, 2001; Masys *et al.*, 2001; Glenisson *et al.*, 2004). En effet, le domaine biomédical est caractérisé par l'existence de nombreux standards terminologiques et thésaurus plus ou moins généralistes (*e.g.* MeSH, UMLS, GO) qui se sont développés au fur et à mesure des besoins. Ces standards sont partagés par la communauté scientifique et notamment par les banques de documents généralistes comme MEDLINE pour l'indexation de ses articles. Masys *et al.* (2001)

[53] http://www.gene.ucl.ac.uk/nomenclature/
[54] http://www.madtools.org
[55] http://bioinfo.weizmann.ac.il/cards/index.shtml

suggèrent ainsi l'utilisation de la hiérarchie des termes MeSH (***Me**dical **S**ubject **H**eading*), associés aux articles de MEDLINE, pour mettre en évidence des éléments de similarité entre des gènes co-exprimés dans des données d'expression. De la même manière, Jenssen *et al.* (2001) ont proposés PubGene[56] (aujourd'hui solution commerciale) pour rechercher dans MEDLINE des réseaux de co-citations de gènes pour annoter et représenter les gènes co-exprimés dans les expériences de puces à ADN. Dans le but d'améliorer l'annotation, ces auteurs associent les relations de co-citations aux mots clés MeSH et aux termes GO. Enfin, Glenisson *et al.* (2004) présentent TXTGate, un outil basé sur l'utilisation simultanée de plusieurs vocabulaires spécifiques (GO, MeSH, eVOC, OMIM) pour associer différents concepts aux gènes.

Des outils commerciaux comme Bibliosphere®[57] de Genomatix (Fig. 38) et PathwayAssist®[58] d'Iobion proposent également des solutions très intéressantes. Par exemple, pour un gène donné, ces deux outils extraient les phrases d'intérêt en accentuant les relations « gène-gène » (*e.g. regulating, decreased, enhanced*), « gène-facteur de transcription » (*e.g.* GATA4, Tbx5) ou le contexte de la relation (*e.g. heart, plasma*) (Fig 38A). Ils offrent également, pour chaque gène, une visualisation 3D dynamique sous la forme d'un réseau de co-citations (Fig. 38B-C). Enfin, Bibliosphere® propose des tableaux de contingence « gènes –facteur de transcription » liés à d'autres d'outils de la suite Genomatix comme MatInspector pour l'analyse des sites de fixation des facteurs de transcription.

En bref

La fouille de textes (*text mining)* vise à automatiser l'analyse des textes écrits en langage naturel (non structuré) pour (re-)découvrir de l'information et de la connaissance dispersées. Les techniques d'analyse vont de l'analyse statistique des co-occurrences de termes (ou mots clés) à l'utilisation des techniques d'analyse du langage naturel. Les principales difficultés dans l'automatisation sont la reconnaissance de vocabulaire spécialisé comme les noms de gènes, les règles de grammaire et l'hétérogénéité des sources. Afin de faciliter l'analyse, les données de la littérature peuvent être associées à des thésaurus de termes et de concepts biologiques.

[56] http://www.pubgene.com/
[57] http://www.genomatix.de/products/BiblioSphere/
[58] http://www.stratagene.com/products/displayProduct.aspx?pid=559

(A)

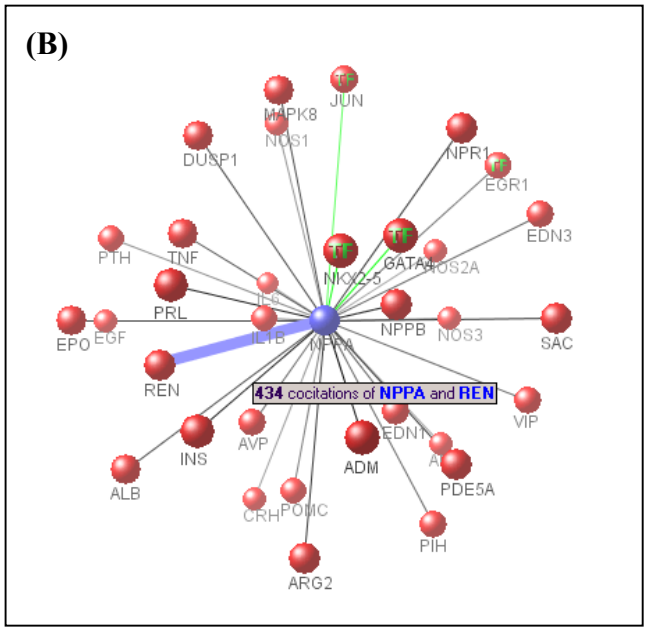
All citations of NPPA:

View next page

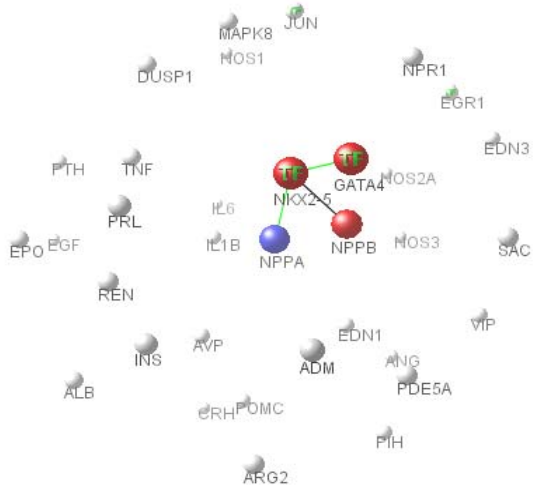
Colors: TF GENE TISSUE FUNCTION

PubMed	Article
14997707	IL-6 plays a key role in **regulating** **cardiac** hypertrophy and the development of **heart** failure, and SERCA, ANP [->NPPA] and BNP are all **cardiac** hormones with regulatory properties.
14988225	CNP and SIN-1 **decreased** atrial stroke volume and myocytic ANP [->NPPA] release.
14978031	In contrast, **Tbx20** **represses** ANF [->NPPA] promoter activity and also **inhibits** the **activation** **mediated** by **Tbx5**.
	Of the two T-box binding consensus sequences in the promoter of ANF [->NPPA], only T-box binding element 1 (TBE1) is **required** for the **synergistic** **activation** of ANF [->NPPA] by **Tbx5** and **GATA4**, but TBE2 is **required** for **repression** by **Tbx20**.
14767494	Chronically measured BP and HR; **plasma/blood** volume; wet and dry ventricular weights; body fat/water; and hormonal profile (**plasma renin** **activity**, aldosterone, cortisol, **atrial natriuretic peptide** [->NPPA], adrenaline, and noradrenaline).
14757752	Cal itself possessed the transcription-promoting **activity**, and cotransfection of Cal **enhanced** CSX/NKX2-5-induced **activation** of **atrial natriuretic peptide** [->NPPA] gene promoter.

(B)



(C)



PMID:15464586 Surprisingly, expression levels of the **putative** **GATA4** **target** **genes** **ANF**, **BNP**, MEF2C, **Nkx2-5**, cyclin D2, and BMP4 were **unchanged** in mutant **hearts**, suggesting that **GATA4** is not a dose-limiting regulator of the expression of these genes during later stages of embryonic cardiac development.

Figure 37. Analyse de la littérature avec Genomatix®. Interrogation du gène NPPA (*natriuretic peptide precursor atrial*) (A) Extrait des relations de co-citation du gène NPPA avec d'autres gènes (rouge), mise en évidence du type de relation tels « inhibits » ou « activation » (gras noir) et localisation tissulaire de cette relation (bordeaux). (B) Réseau de co-citations des gènes cités au moins 6 fois avec NPPA, tel le gène codant pour la renine (REN) ou les facteur transcription (TF) JUN, NKX2-5, et GATA4. (C) Zoom des relations NPPA entre les facteurs de transcription NKX2-5, GATA4 et le gène NPPB.

3. Banques de données pour la biologie

Comme pour les articles scientifiques dans les années 80, des banques de données sont aujourd'hui devenues indispensables pour **sauvegarder** et **structurer** les informations issues des expériences de biologie moléculaire et plus particulièrement des données générées par les différentes technologies à haut débit (*e.g.* puce à ADN, SAGE). Un autre avantage des banques de données (surtout publiques) est d'améliorer le **partage** des informations dans le but de faciliter la (re-)découverte de connaissances.

3.1 Banques de données publiques

Le numéro spécial du journal *Nucleic Acids Research* sur les banques de données pour la biologie moléculaire référence, **au début de l'année 2005, 719 banques de données publiques**, soit 171 de plus que l'année précédente (Galperin, 2005). Cette collection se répartit actuellement en 14 catégories (3 de plus que l'an passé) allant des banques de données de séquences nucléotidiques à celles des données immunologiques en passant par les banques de données d'expression et de voies métaboliques. Ces banques de données sont plus ou moins généralistes, *i.e.* dédiées à un ou plusieurs organismes, une ou plusieurs organelles. Cependant, cette liste de ressources d'information biologiques est loin d'être exhaustive. Les banques de données ainsi référencées doivent être publiques et directement accessibles par les utilisateurs *via* le Web. Aussi, de nombreuses banques de données passées dans le domaine privé ou nécessitant l'installation de logiciel en local ne sont pas recensées.

3.2 Banques de données d'expression de gènes

Parmi les banques de données publiques, les banques données d'expression de gènes sont particulièrement importantes et intéressantes en terme de partage des connaissances. Ces banques de données se répartissent globalement en 2 catégories plus ou moins généralistes.

Les **banques de données généralistes pour le dépôt des données d'expression de gènes** (***repository***) ont été développées dans le but de partager les données d'expression de gènes (notamment issues des expériences de puces à ADN) au niveau de la communauté scientifique internationale. L'une de leur priorité est le respect par les biologistes du standard international MIAME (Brazma *et al.*, 2001) pour uniformiser les données et faciliter leur diffusion.

Les trois principaux banques de données généralistes pour le dépôt des données d'expression de gènes sont ArrayExpress[59] à l'EBI (Brazma *et al.*, 2003), GEO[60] au NCBI (Edgar *et al.*, 2002) et Cibex (***C**enter for **I**nformation **B**iology gene **EX**pression database*)[61] au DDBJ (***D**NA **D**ata **B**ank of **J**apan*) (Ikeo *et al.*, 2003). Ces *repositories* sont d'importance grandissante puisque, aujourd'hui, la majorité des journaux scientifiques requièrent, pour toutes publications dans le domaine des puces à ADN, le dépôt des données d'expression dans au moins une des banques de données publiques conforme au standard international MIAME.

Les *repositories* permettent de comparer les dessins expérimentaux réalisés pour répondre à diverses questions biologiques. Ils offrent la possibilité de confronter des matrices de données d'expression générées par différentes équipes, sur différents modèles et/ou différentes plates-formes. Les résultats de ces comparaisons permettent, entre autre, d'améliorer l'annotation et la connaissance sur les gènes dans les différentes conditions (Stuart *et al.*, 2003; McCarroll *et al.*, 2004). L'autre intérêt de ces banques de données généralistes est la mise à disposition des jeux de données aux communautés de chercheurs en bio-informatique, mathématiques et statistiques pour le développement de nouvelles méthodologies d'analyse. La conférence internationale Pacific Symposium on Biocomputing[62] est un exemple de réussite de ce type d'approche. De même, de nombreux articles, parus dans les journaux traitant de bio-informatique, présentent des algorithmes testés sur des jeux de données extraits des différentes banques de données publiques. Les résultats obtenus offrent même parfois un complément d'information sur les résultats biologiques.

Les **banques de données dédiées** ciblent principalement des organismes, des organes ou des conditions physiologiques (-pathologiques) précises. Certaines de ces banques sont également des *repositories*. MGD[63] (***M**ouse **G**enome **D**atabase*), par exemple, regroupe des informations sur le génome murin (Ringwald *et al.*, 2001) et propose la diffusion des données expérimentales obtenues par différentes technologies comme les puces à ADN, la technique d'hybridation fluorescente (FISH), ou encore des analyses de polymorphisme. *Gene expression in tooth*[64] présente, quant à elle, le niveau d'expression des gènes dans les différents épithéliums dentaires. T1D|base[65] intègre l'ensemble des données disponibles

[59] http://www.ebi.ac.uk/arrayexpress/
[60] http://www.ncbi.nlm.nih.gov/geo/
[61] http://cibex.nig.ac.jp/index.jsp
[62] http://psb.stanford.edu/
[63] http://www.informatics.jax.org/
[64] http://bite-it.helsinki.fi
[65] http://t1dbase.org

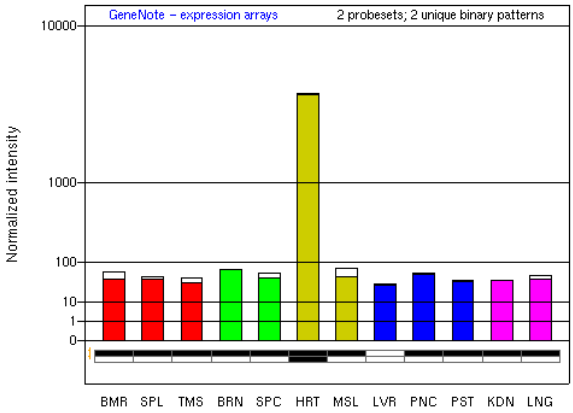
concernant le diabètes de type 1 (Smink *et al.*, 2005). Enfin, GeneNote66 (***Gene Normal Tissue Expression***) présente les niveaux d'expression de base des gènes, obtenus dans divers tissus par différentes technologies (SAGE, Affymetrix, electronic nothern) (Fig. 39) (Shmueli *et al.*, 2003).

3.3 Interopérabilité et extraction de connaissances des banques de données publiques

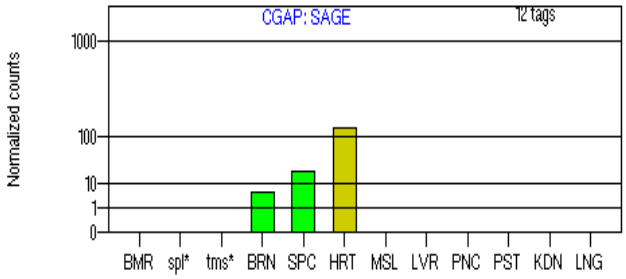
L'intégration des méta-données passe par l'interopérabilité des différentes sources. Cependant, les données biologiques dispersées dans les différentes banques de données sont le plus souvent hétérogènes, parfois redondantes et généralement de qualité inégale. De plus, l'absence ou le non-respect des standards de nomenclature pour annoter les objets biologiques rend délicat l'interopérabilité, l'intégration et la comparaison des différentes sources d'informations. Par exemple, suite à une recherche dans PubMed des articles parlant du précurseur A du peptide natriurétique depuis le début 2005 (requête du 16/03/05), 3 articles de génomique emploient le nom officiel (nomenclature du HGNC) de NPPA contre 187 articles qui utilisent l'alias ANP. Dès lors, la principale difficulté est de savoir **où et comment trouver les « bonnes » informations**.

Des outils d'aide à la décision proposent aujourd'hui des scénarios de requêtes sur les différentes banques de données publiques (Etzold *et al.*, 1996; Teusan, 2002; Boulakia *et al.*, 2004). Par exemple, MADSENSE, développé au sein du laboratoire par Raluca Teusan, est un service Web qui intègre des informations biologiques et bibliographiques dans un seul système d'aide à la compréhension des gènes (*cf.* p. 37) (Fig. 40).
Les outils d'aide à la décision prennent en considération les préférences des utilisateurs et présentent, pour certains, l'utilisation d'indices de confiance pour estimer la qualité des données recueillies. Par exemple, sur une échelle de 1 à 10, l'utilisateur peut attribuer un niveau de confiance de 9 aux données concernant les protéines annotés dans la banque de données Swiss-Prot et un indice de 7 aux informations de GenBank Protéine, les données soumises ou contenues dans la banque de donnée Swiss-Prot étant validées plus rapidement que dans GenBank (Boulakia *et al.*, 2004).

[66] http://genecards.weizmann.ac.il/cgi-bin/genenote/home_page.pl

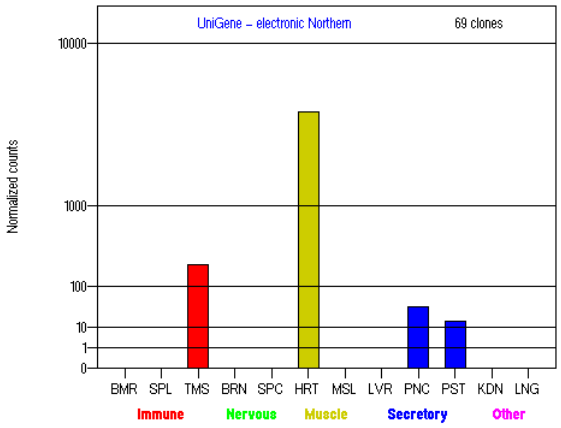


(B)



(C)

Tissue		Tags per gene	Total tags
BMR	Bone marrow	0	36,577
spl*	Spleen	0	0
tms*	Thymus	0	0
BRN	Brain	2	427,603
SPC	Spinal cord	1	54,785
HRT	Heart	9	83,063
MSL	Skeletal muscle	0	107,836
LVR	Liver	0	66,308
PNC	Pancreas	0	43,040
PST	Prostate	0	123,335
KDN	Kidney	0	40,993
LNG	Lung	0	88,708



Tissue		Clones per gene	Total clones
BMR	Bone marrow	0	17,521
SPL	Spleen	0	13,072
TMS	Thymus	2	9,088
BRN	Brain	0	183,446
SPC	Spinal cord	0	587
HRT	Heart	64	14,040
MSL	Skeletal muscle	0	69,597
LVR	Liver	0	50,284
PNC	Pancreas	2	51,565
PST	Prostate	1	61,004
KDN	Kidney	0	74,913
LNG	Lung	0	70,848

Figure 38. Niveau d'expression du gène NPPA dans différents tissus d'après GeneNote (Shmueli *et al.*, 2003). (A) Niveau moyen d'expression de deux sondes Affymetrix® issues des puces GeneChips HG-U95A-E. (B) Niveau d'expression selon la technologie SAGE. (C) Northern électronique à partir la banque de données Unigene (NCBI). Pour un gène donné et dans un tissu donné, le niveau d'expression correspond au nombre de clones (non redondant) pour ce gène sur le nombre de clones présents dans le tissu.

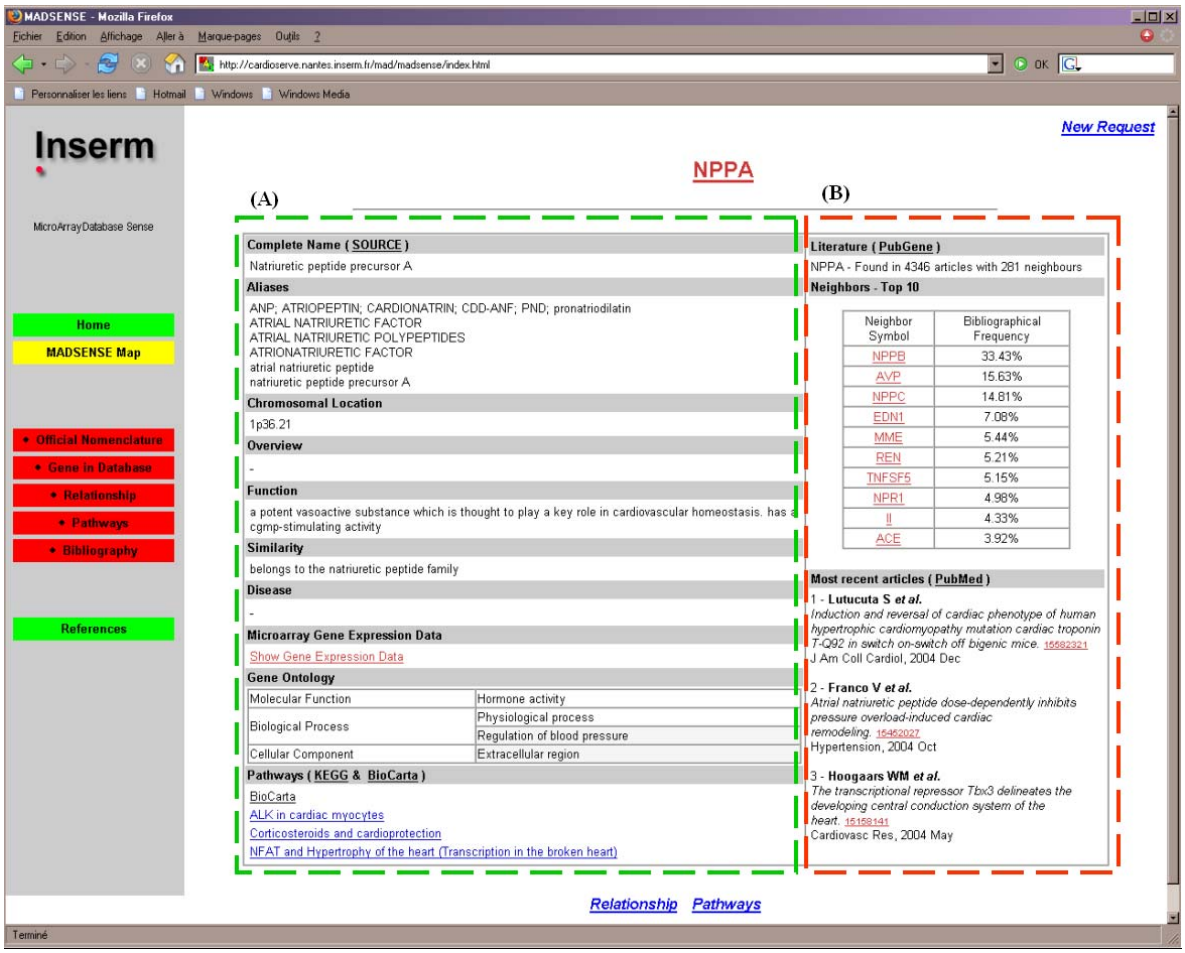


Figure 39. MADSENSE (*MicroArray Data SENSE*) : service Web pour l'intégration des données biologiques et bibliographiques. (A) Informations issues des banques de données biologiques SOURCE, KEGG, BIOCARTA. (B) Informations issues des banques de données bibliographiques PubMed et PubGene.

Des langages, tel XML et ses dérivés, sont développés pour faciliter l'interopérabilité des différentes sources électroniques d'informations. Par exemple, MAGE-ML (Spellman *et al.*, 2002) ou SBML[67] (***S**ystem **B**iology **M**arkup **L**anguage*) (Hucka *et al.*, 2003) sont des langages basés sur le format XML et dédiés au domaine de la biologie. MAGE-ML vise à formaliser et faciliter la présentation des données issues des expériences de puces à ADN. Une de ses applications est notamment le transfert automatique les informations contenues dans les bases de données des laboratoires (bases de données locales), telle BASE, vers les banques de données publiques d'expression de gènes comme ArrayExpress ou GEO. SBML est, quant à lui, un format pour représenter des modèles de réseaux de réactions biochimiques. SBML s'applique par exemple à la description des voies métaboliques, des mécanismes de signalisation cellulaire ou encore des réseaux de régulation. Ce langage est actuellement supporté par près de 75 logiciels, tels KEGG2[68] ou PANTHER pathway[69] .

Enfin, des projets plus ambitieux présentent des architectures *middleware* pour interroger et synthétiser les informations issues de différentes banques de données. Le *middleware*, littéralement ``élément du milieu", est une interface de communication universelle entre processus. Son objectif principal est d'unifier l'accès et la manipulation de l'ensemble des services logiciel disponibles sur le réseau, afin de rendre l'utilisation de ces derniers presque transparente. Par exemple, Biomediator (Shaker *et al.*, 2004) propose d'optimiser les requêtes entre bases de données à l'aide du langage PQL (Mork *et al.*, 2002). Ce langage, basé sur du XML, permet la construction de requêtes complexes pour l'analyse de sources XML.[my]Grid (Stevens *et al.*, 2003) est, quant à lui, un projet de recherche « *e-Science »* (e pour électronique) pour des expériences *in silico.* L'objectif de ce projet est de développer des outils afin de tester et valider des hypothèses par l'interrogation et l'analyse informatique des multiples banques de données biologiques.

[67] http://sbml.org/index.psp
[68] http://www.genome.jp/kegg/kegg2.html
[69] https://panther.appliedbiosystems.com/pathway/

4. Vers la biologie intégrative

Outre l'annotation des gènes et de leurs produits, **l'intégration des méta-données avec les données d'expression offre la possibilité d'expliquer la co-expression des gènes**. En effet, plusieurs hypothèses « croisées » permettent expliquer ce phénomène. Les gènes peuvent, par exemple, être **co-localisés sur les chromosomes**, **co-régulés par un même facteur et/ou appartenir à un même réseau de régulation génétique**.

4.1 Co-localisation chromosomique

La co-localisation chromosomique est une des hypothèses émises pour expliquer la co-expression des gènes. Dans les génomes procaryotes notamment, les gènes adjacents, souvent sous forme d'opérons, participent généralement à la même fonction biologique. De nombreuses analyses réalisées chez divers organismes eucaryotes, tels la drosophile (Cohen *et al.*, 2000; Spellman et Rubin, 2002), le nématode ou l'homme (Lercher *et al.*, 2002), ont également montré que des gènes adjacents sur le génome présentent souvent des profils d'expression similaires. Il existe plusieurs explications possibles à cette régionalisation de l'expression des gènes (Williams et Bowles, 2004). Tout d'abord, au cours de l'évolution, les **gènes dupliqués** restent souvent voisins et, compte tenu de leur ancêtre commun, ont tendance à avoir des profils d'expression similaire. De plus, même en l'absence de régulation concertée, certains gènes adjacents des génomes eucaryotes peuvent partager des éléments **cis-régulateurs** à l'origine de profil d'expression similaire. Enfin, certains suggèrent également une organisation supérieure des génomes dans laquelle l'ordre des gènes le long des chromosomes serait **corrélé avec les tissus** dans lesquels ils s'expriment (Roy *et al.*, 2002) ou permettrait l'**expression des gènes de « ménage »** indispensables au fonction de « base » de la cellule (Lercher *et al.*, 2002).

4.2 Facteurs de transcription

Comme souligné précédemment, la co-expression des gènes peut s'expliquer par des phénomènes de co-régulation. La r**echerche de régions cis-régulatrices et/ou de facteurs de transcription** communs à plusieurs gènes a notamment été réalisée au travers d'analyses de puces à ADN. Les premiers résultats ont été obtenus chez *Saccharomyces cerevisae.* DeRisi *et al.* (1997) ont montré une corrélation entre des groupes de gènes co-régulés et la présence de sites de fixation de facteurs de transcription en amont de la séquence codante de ces gènes. De nombreuses études se sont ensuite succédées (essentiellement chez les procaryotes car

moins complexe) pour rechercher de nouveaux sites de fixation de facteurs de transcriptions en combinant données d'expression et la découverte de motifs dans les séquences (synthèse par Vilo et Kivinen (2001)). Ainsi, grâce à l'utilisation d'algorithmes de classification (type K-moyenne) et au logiciel AlignACE (Roth *et al.*, 1998), Tavazoie *et al.* (1999) ont mis en évidence, chez la levure, 15 groupes de gènes dont les séquences en amont du cadre de lecture présentent des motifs similaires.

La découverte d'un profil de séquences est basée sur la recherche *a priori* d'un motif inconnu (chaîne de caractère, expression régulière, matrice pondérée) qui est statistiquement sur-représenté dans un jeu de séquences données. Les méthodes de découvertes de profil sont réparties en 2 catégories basées sur les séquences (*sequence-driven*) ou les profils (*pattern-driven*). Cette dernière est plus efficace. Toutefois, la recherche de motifs consensus à plusieurs séquences reste très sensible aux faux positifs (les séquences issues d'un même cluster ne sont pas nécessairement co-régulées).

Les profils peuvent être décrits sous la forme de profils probabilistes (*e.g.* matrices pondérées) ou discrets (*e.g.* expression régulières). La première représentation offre généralement de meilleurs résultats malgré sa complexité (difficile à implémenter). L'approche probabiliste la plus connue est l'algorithme EM ou ***E**xpection **M**aximisation* (Lawrence et Reilly, 1990) qui recherche dans les séquences des motifs consensus de taille connue. Cet technique est à l'origine de nombreux modèles et outils tels que Gibbs Motif Sampling, MEME, BioProspect, TFEM (Kechris *et al.*, 2004) ou Pattern Explorer d'Atragene™. Des sites Web, comme MATCH™ ou TESS[70], permettent également la recherche de motifs grâce à la comparaison des séquences avec des bases de données de promoteurs comme TRANSFAC[71] (Wingender *et al.*, 2001). La recherche de profils discrets est, quant à elle, basée sur une structure des données en arbre des suffixes (McCreight, 1976; Ukkonen, 1995), alors indépendante de la longueur des séquences testées (Vilo et Kivinen, 2001).

[70] http://www.cbil.upenn.edu/tess/
[71] http://www.gene-regulation.com/index.html

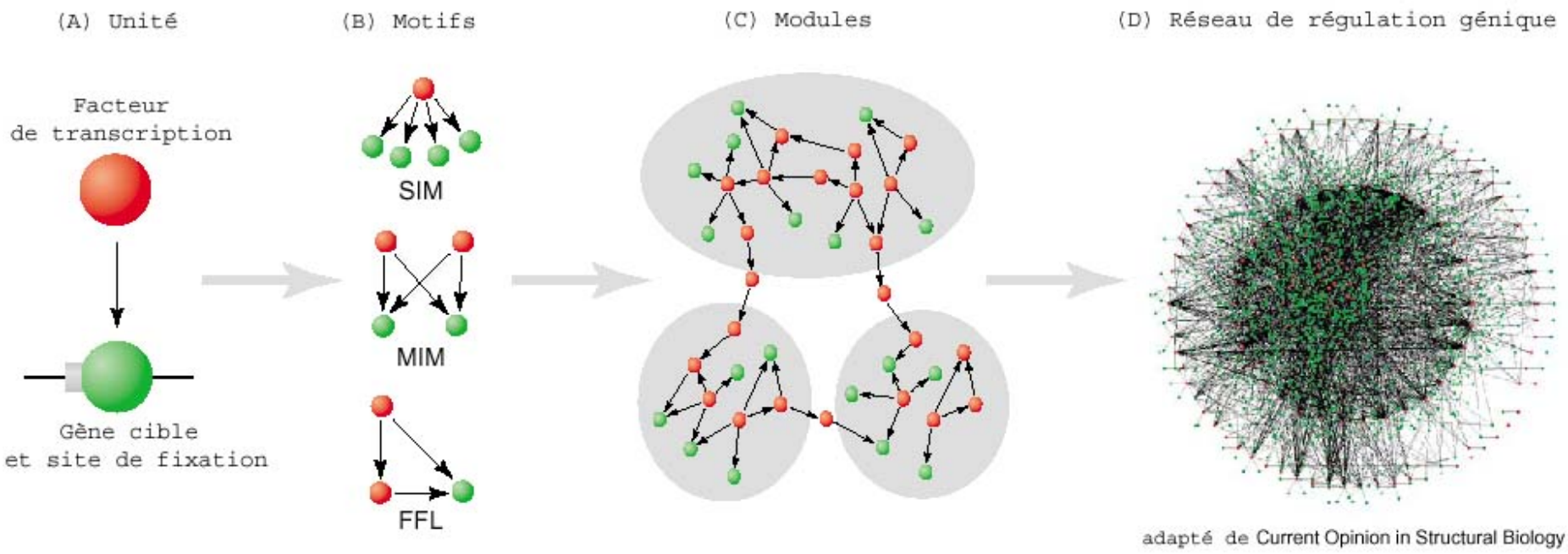


Figure 40. Organisation structurale des réseaux de régulation de la transcription (adapté de Babu *et al.*, 2004). (A) L' « unité » de base est le facteur de transcription, son gène cible avec le site de reconnaissance et la fonction de régulation entre les deux. (B) Les unités de base sont, la plupart du temps, organisées en « motifs » avec des profils de régulation spécifiques. Par exemples, un facteur de transcription est capable d'interagir avec plusieurs gènes cibles (SIM ou *Single input*), deux gènes cibles peuvent être reconnus par les deux même facteurs de transcriptions (MIMI ou *Multiple input*) et un facteur de transcription peut agir directement ou indirectement sur un gène cible (FFL – *Feed-forward loop*). (C) Les motifs peuvent être interconnectés pour former des modules semi-indépendants. (D) Réseau de régulation génique.

La limite de ces différentes approches est l'absence de prise en considération de l'aspect dynamique des mécanismes de régulation. En effet, la co-régulation des gènes est « condition-dépendante » (Chiang *et al.*, 2001; Pilpel *et al.*, 2001). Chiang *et al.* (2001) proposent donc une approche « par groupe de séquence » avec une métrique appelée GMEP (***G**enome-**M**ean **E**xpression **P**rofile*) où le niveau de contrôle des mécanismes de régulation est évalué sur la base des profils d'expression des gènes qui possèdent les mêmes motifs en *cis*. Ils ont ainsi mis en évidence 9 groupes de motifs (et les gènes associés) dont le niveau de contrôle varie avec les conditions expérimentales. Pilpel *et al.* (2001) également suggèrent que différentes combinaisons d'un petit nombre de facteurs de transcriptions sont à l'origine de la majorité des mécanismes de régulation de la transcription.

Ces derniers travaux introduisent la notion de réseaux de régulations. En effet, les mécanismes de régulation sont non seulement dynamiques mais aussi multiples et se complexifient avec le niveau de complexité de l'organisme (Pilpel *et al.*, 2001; Babu *et al.*, 2004).

4.3 Réseaux de régulations génétiques

Un réseau de régulation génétique est représenté par un DAG, *i.e.* graphe dans lequel chaque régulateur et chaque gène est représenté par un nœud et les liens entre les nœuds représentent les interactions entre les gènes et son (ses) régulateur(s) (Fig. 41) (synthèse sur la structure et l'évolution des réseaux de régulation génétiques par Babu *et al.* (2004)). De nombreuses approches existent pour mettre en évidence les modules et réseaux de régulation géniques (synthèse par van Someren *et al.* (2002)). Outre la recherche de régions cis-régulatrices (Roth *et al.*, 1998; Tavazoie *et al.*, 1999; Pilpel *et al.*, 2001), il est nécessaire de définir les liens entre les motifs de régulation.

a) Réseaux bayésiens

La statistique bayésienne est une approche prometteuse pour l'inférence de réseaux d'interaction géniques (Fig. 41) (Friedman *et al.*, 2000). Particulièrement adaptée à la prise en compte de l'incertitude, elle convient aux données bruitées (procédures expérimentales, précision des mesures, phénomènes biologiques aléatoires) que sont les données issues des expériences de puces à ADN. De plus, les méthodes bayésiennes peuvent être couplées à des algorithmes d'apprentissage supervisé pour sélectionner le sous-ensemble de gènes permettant de caractériser au mieux un groupe (Hartemink *et al.*, 2001; Pe'er *et al.*, 2001). La principale problématique est la définition de la distribution des probabilités conditionnelles.

(*e.g*. multinomiale pour des variables discrètes, gaussienne pour des variables continues). Elles peuvent être décrites manuellement par les experts du domaine (Troyanskaya *et al.*, 2003) ou générées automatiquement par apprentissage. Ce dernier requiert toutefois l'analyse d'un grand nombre de conditions et l'application d'algorithmes d'apprentissage (chaînes de Markov, permutation et ré-échantillonnage) souvent coûteux en temps de calcul. Ainsi Segal *et al.* (2003) proposent un modèle probabiliste basé sur les réseaux bayésiens pour identifier des modules de régulation sur la base de profils d'expression. Ils découvrent alors, chez la levure, des groupes de gènes co-régulés, leurs régulateurs et les conditions pour lesquelles ces mécanismes sont actifs (Fig. 42).

b) Génomique comparative

Chez les organismes supérieurs, tel que l'homme, la tâche est plus complexe compte tenu de la taille du génome et des nombreux mécanismes de régulation, notamment en *trans* (Morley *et al.*, 2004). Une des solutions proposées est la génomique comparative. En effet, les gènes et facteurs de régulation d'organismes relativement proches sont en partie conservés au cours de l'évolution. Il est donc possible d'inférer, à partir d'organismes connus, la fonction de certains gènes voire identifier des réseaux de régulation d'organismes moins bien étudiés (Baliga *et al.*, 2004). Pour exemple, McCaroll *et al.* (McCarroll *et al.*, 2004) ont développé une méthode systématique de comparaison des profils d'expression des gènes de différents organismes (*Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* et *Homo sapiens*). Ils ont ainsi mis en évidence des mécanismes d'expression génétiques analogues entre les différents organismes, mécanismes qui partagent des profils de régulation entre gènes orthologues. Les objets biologiques de ces profils peuvent être identifiés comme des éléments très conservés caractérisés par des catégories GO. De la même manière, Stuart *et al.* (2003), ont montré la conservation de groupes de gènes co-régulés entre 4 génomes eucaryotes dont l'homme. Grâce au concept de « *metagene* » (*cf.* p. 105), ils démontrent que les gènes conservés au cours de l'évolution ont des relations fonctionnelles. Ils ont ainsi pu inférer une fonction biologique à des gènes non encore annotés. Enfin, cette étude suggère également que les facteurs de transcription sont moins bien conservés que les gènes cibles, ce qui signifie que la régulation des gènes évolue plus vite que les gènes eux-mêmes.

2. FALSE

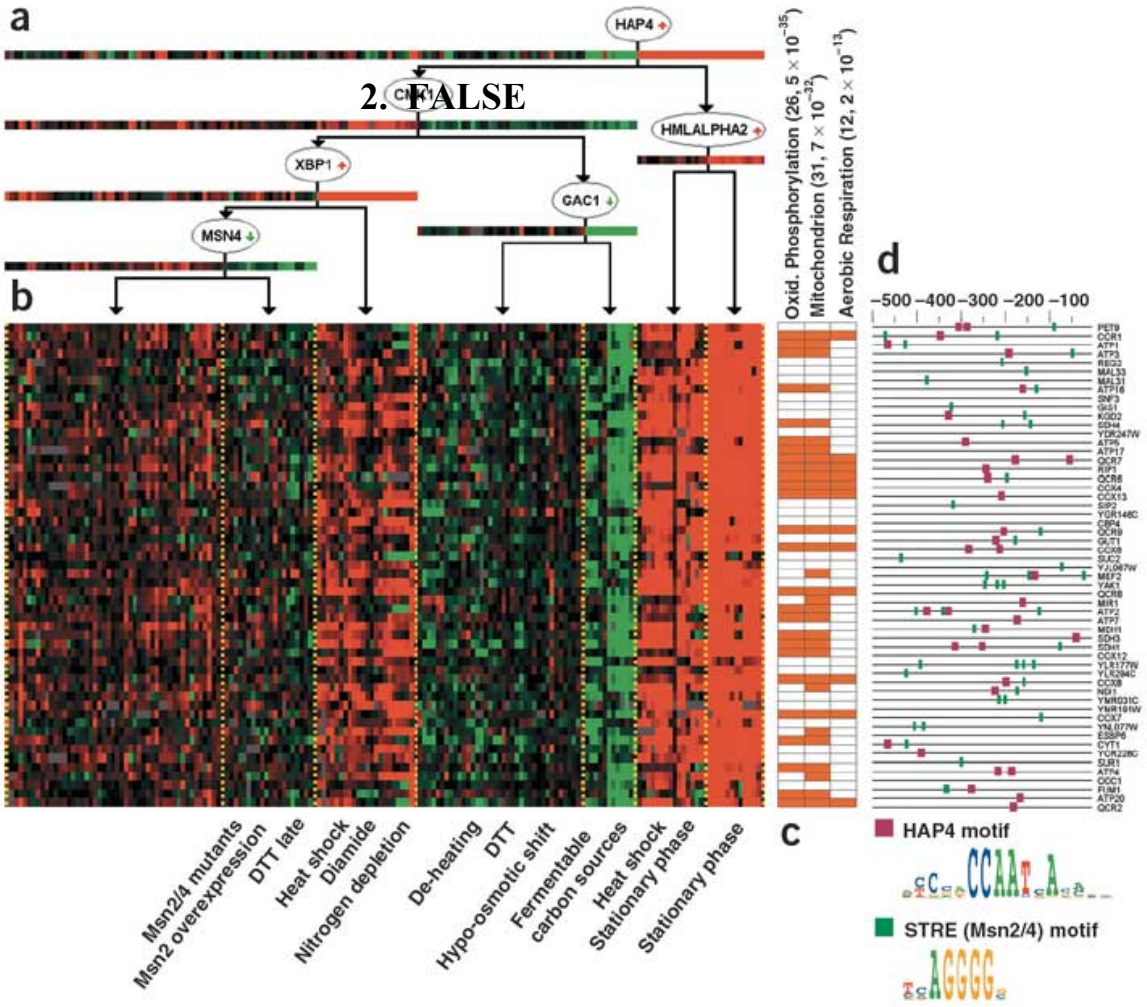


Figure 41. Mise en évidence de modules de régulation selon Segal et al. (2003). Module composé de 55 gènes impliqués dans la régulation du carbone et de la respiration. (a) Arbre du programme de régulation. Chaque nœud de l'arbre représente un régulateur (exemple HAP4) et une question sur son niveau d'activité (par exemple une flèche rouge vers le haut correspond à la question « Est-ce que Hap4 est sur-exprimé ? ». A droite de l'arbre la réponse à la question posée est vraie, à gauche elle est fausse. Le niveau d'expression de chaque régulateur est présenté sous l'arbre. (b) Profiles d'expression des gènes En ligne les gènes, en colonne les différentes conditions. Les puces à ADN sont ordonnées suivant l'arbre de régulation. Par exemple, les puces les plus à droite sont celles pour qui les profiles d'expression de HAP4 et HMLALPHA2 sont sur-exprimés. (c) Annotation des gènes d'après Gene Ontology Catégories fonctionnelles les plus significativement représentées (entre parenthèse est indiqué le nombre de gènes annotés et la *p-value* associée) dans le module (de gauche à droite, processus biochimique, localisation cellulaire, processus physiologique). (d) Analyses des promoteurs. Motifs issus de la base de données TRANSFAC (Wingender *et al.*, 2001). Les lignes représentent les 500 pb en amont du codon *start* des gènes. Les boîtes colorées indiquent la présence d'un motif cis-régulateur. Les motifs HAP4 (fushia) et STRE (vert) sont notamment sur-représentés comme l'indique également la représentation en logo en bas de la figure (Schneider et Stephens, 1990).

En conclusion, l'intégration des méta-données est indispensable à l'annotation et à la compréhension des mécanismes de fonctionnement des gènes. Elle offre surtout des perspectives dans le domaine de la biologie intégrative pour la compréhension du fonctionnement des systèmes biologiques (*systems biology*) ; un système biologique étant défini comme un ensemble (plus ou moins grand) d'éléments variés, aux fonctionnalités différentes, qui interagissent de manière sélective et non linéaire pour engendrer des mécanismes biologiques cohérents (Kitano, 2002).

En bref

Les méta-données sont les informations issues de différentes sources comme les ontologies, la littérature scientifiques, ou les banques de données d'expression...

L'élaboration d'une ontologie permet la définition de concepts stricts et introduit de la logique. *Gene Ontology* est devenu un standard pour l'annotation des génomes. GO se compose de trois ontologies qui définissent les processus biologiques, les fonctions moléculaires et la localisation cellulaire des produits de gènes. Cette ontologie facilite le partage des connaissances entre les experts et est utilisable par les machines. Par ailleurs, elle permet de mettre à jour les contradictions et les manques dans la connaissance actuelle.

La littérature est la première source d'information scientifique. Son analyse est l'une des premières approches pour extraire de la connaissance sur les gènes. La fouille de textes (*text mining)* vise à automatiser l'analyse des textes écrits en langage naturel (non structuré) pour (re-)découvrir de l'information et de la connaissance dispersées.

L'intégration des méta-données (séquences, GO, littérature…) à partir des banques de données publiques est indispensable à l'annotation et à la compréhension des mécanismes de fonctionnement des gènes. L'intégration des méta-données passe par l'interopérabilité des différentes sources de données qui est devenu un des nouveaux enjeux de la bio-informatique.

Conclusions & Perspectives

Conclusions & Perspectives

Notre équipe développe des études de génomique fonctionnelle dans le domaine cardiovasculaire et neuromusculaire au moyen à la technologie des puces à ADN. Le traitement et l'analyse des données d'expression de gènes issues des expériences menées au moyen des nouvelles technologies à haut débit, et plus particulièrement des puces à ADN, nécessitent des outils bio-informatiques nouveaux. Compte tenu des caractéristiques atypiques (dimension et aspect bruité) des matrices de données d'expression issues des puces à ADN, l'utilisation de méthodes mathématiques et statistiques spécifiques est incontournable. De nombreux outils ont dû être développés de l'extraction des données par les logiciels d'analyse d'images à la recherche des réseaux moléculaires, en passant par la normalisation et la validation des données. Les bases biologiques et informatiques indispensables à la compréhension de ces nouvelles approches sont résumées dans la première partie de ce mémoire.

La partie centrale de ce manuscrit traite des problématiques suivantes :

(i) analyse d'images des puces à ADN,

(ii) métrologie (plan expérimentaux, qualité des données primaires…),

(iii) transformations des données primaires en données « consolidées ».

Ce travail a abouti au développement du service Web MADSCAN[72] (*MicroArray Data Suite of Computed Analysis)*. Cet outil permet de filtrer, normaliser et valider statistiquement les données primaires issues des logiciels d'analyse d'images. MADSCAN est actuellement utilisé internationalement, avec déjà plus de 300 visiteurs et 200 analyses pour les 3 premiers mois de l'année 2005 (consultation au 16/03/05). Un progrès récent est l'interfaçage de MADSCAN avec la base de données BASE, réalisé au sein du laboratoire par Audrey Bihouée**.** Ce *plug-in* a été proposé à la communauté BASE et sera prochainement intégré à un article dédié aux *plug-ins* développés pour BASE. De ce fait, les codes sources de MADSCAN (version *plug-in* BASE) sont disponibles sous la licence GPL. Nous sommes en train d'étendre les fonctionnalités de l'outil : traitement des données de puces mono-couleur, mise en évidence de gènes différentiellement exprimés et stables entre plusieurs conditions.

[72] http://madtools.org

La plupart de ces analyses sont déjà implémentées dans des librairies du projet BioConductor que nous adaptons à l'outil MADSCAN. Nous pensons également ajouter de nouvelles fonctionnalités graphiques pour offrir une plus grande souplesse à l'utilisateur avec, par exemple, des graphiques interactifs.

Si les étapes d'acquisition, de traitement et de validation statistique des données sont aujourd'hui maîtrisées au sein du laboratoire (en partie grâce à MADSCAN), nos efforts portent désormais sur l'analyse et l'intégration des données. Aussi, la troisième partie de ce manuscrit présente une étude bibliographique des principales méthodes de :

(i) mise en évidence des gènes différentiellement exprimés,

(ii) classification des données d'expression,

(iii) intégration des méta-données.

La figure 43 résume certaines de ces analyses et les outils bio-informatiques associés. De manière générale, les analyses et les outils à appliquer dépendent de la question biologique posée. Toutefois, l'intégration des méta-données est devenue une étape essentielle à la compréhension des données d'expression. Plus important encore, l'intégration des méta-données participe à l'émergence de la biologie intégrative dont l'objectif est la description et la compréhension du fonctionnement des systèmes biologiques. Elle est devenue l'un des nouveaux enjeux de la bio-informatique.

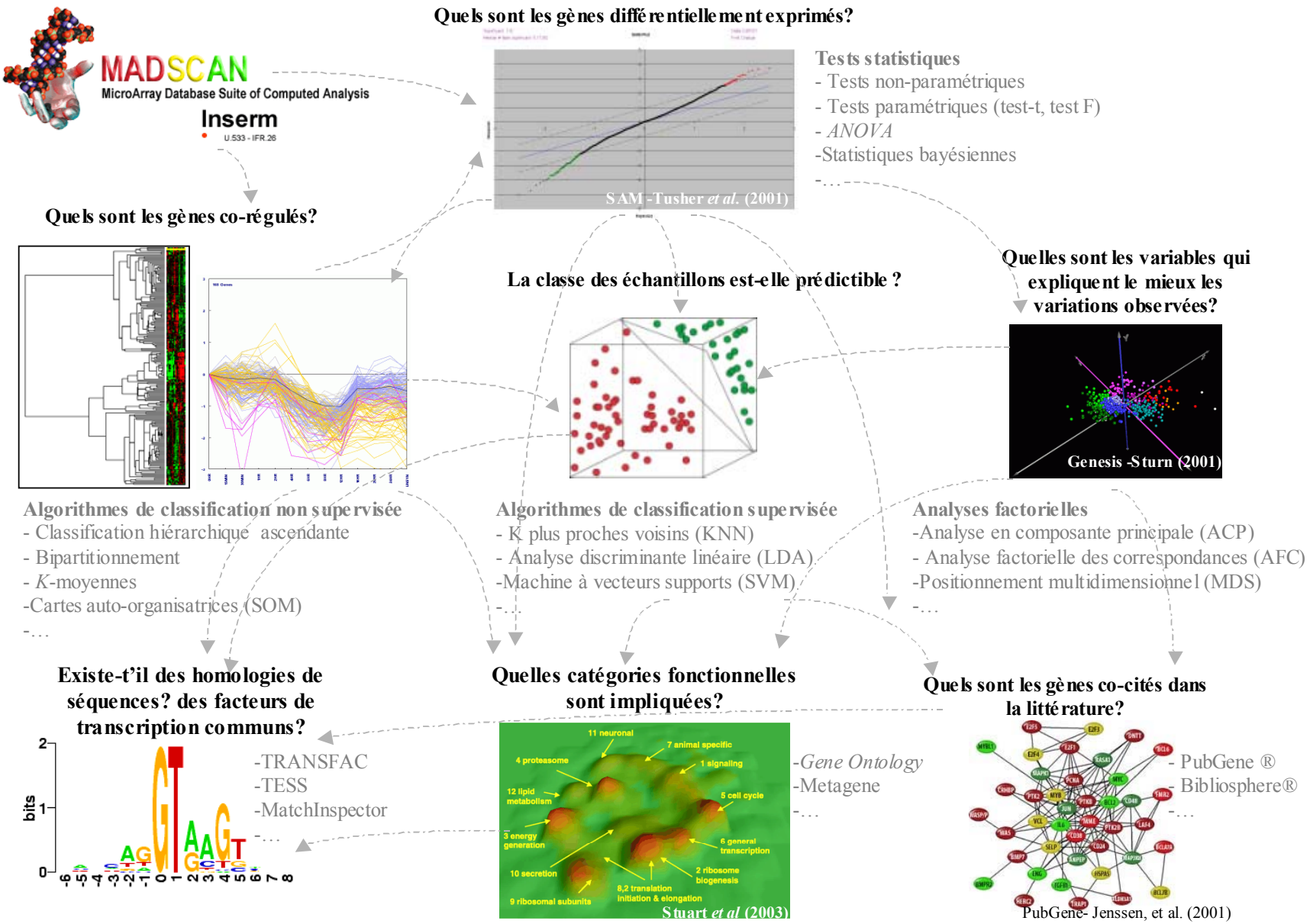


Figure 43. Exemples d'analyses et d'outils disponibles pour l'extraction des connaissances des données de puces à ADN.

Glossaire

Glossaire

Abréviations

abr. : abréviation
ang. : terminologie anglaise
fr. : terminologie française
info. : ayant trait à l'informatique
algo. : ayant trait à algorithmique
math. ayant trait aux mathématiques
stat. : ayant trait aux statistiques
bio. : ayant trait à la biologie

*
* *

Acide nucléique (fr., bio.) Acides organiques constituants des noyaux cellulaires dont dépendent les caractères géniques des chromosomes.

ACP (abr., fr., stat.) *Analyse en Composante Principale*, technique mathématique pour la description de données muti-dimensionnelles. (ang. : PCA *Principal Component Analysis*)

ADN (abr., fr., bio.) *Acide DésoxyriboNucléique.* Support de l'hérédité chez l'homme. (ang. : DNA)

ADNc (abr., fr., bio.) *Acide DésoxyriboNucléique complémentaire.* (ang. : cDNA)

AFC (abr., fr., stat.) Analyse Factorielle des Correspondances, technique mathématique pour la description de données muti-dimensionnelles.

Algorithme (fr., info.) Suite finie d'opérations élémentaires constituant un schéma de calcul ou de résolution d'un problème.

ANOVA (abr., ang., stat.) *ANalysis Of Variance.* Analyse de variance. Analyse statistique pour comparer et tester l'égalité des moyennes.

API (abr., ang., info.) *Application Programming Interface.* Interface pour langages de programmation, matérialisées par des primitives, permettant à une application d'accéder à des programmes système pour, par exemple, communiquer ou extraire des données.

Apprentissage (fr., algo., info.) doté de l'aptitude à modifier ses réponses futures en fonction de son expérience passée.

ARN (abr., fr., bio.) *Acide RiboNucléique* acide nucléique formé d'un enchaînement de nucléotides constitués de ribose et d'une base purique ou pyrimidique. (ang : RNA

Barycentre (fr., math.) Point unique d'un espace affine associé à une collection de points de cet espace affectés chacun d'un scalaire, et défini par extension de la notion de centre de gravité.

Batch (ang.) ou **Traitement par lots** (fr.) Fonctionnement d'un système où les tâches à accomplir ne sont pas traitées au fur et à mesure de leur arrivée, mais d'abord groupées dans une file d'attente avant d'être exécutées en une seule séquence continue.

Biopsie (fr., bio.) Prélèvement sur un être vivant d'un fragment de tissu en vue d'un examen microscopique.

bp (abr., ang., bio.) *base paire.* Paire de base. Unité de mesure de la distance sur la carte physique de l'ADN en paires de bases. A noter : La taille d'un ARN est donnée en nt (pour nucléotides).

Bug (ang, info.) Erreur ou bogue dans un programme, dans la programmation ou le câblage d'un composant électronique, entraînant des anomalies de fonctionnement et rarement désirées du système. (voir déboguer)

cf. (abr.) *confer*, abréviation latine qui signifie « comparer, rapprocher ». Indication invitant le lecteur à se référer à ce qui suit.

CGI (abr., ang., info.) *Common Gateway Interface.* Spécification concernant l'interfaçage d'un serveur Web avec une application.

Chromosome (fr., bio.) Porteurs des facteurs déterminants de l'hérédité.

Cis (fr., bio.) **Région cis-régulatrice** Régulation de l'expression d'un gène par une séquence d'ADN non codante, située sur le même chromosome que ce gène (ex: enhancer, promoteur).

Client-serveur (fr., info.) Architecture qui s'appuie sur un concept de répartition des traitements et des données sur un ensemble de systèmes comprenant à la fois des serveurs centraux ou départementaux et des micro-ordinateurs ou des réseaux locaux.

Cluster (ang., math., info.) Groupement d'un petit nombre d'objet.

CNGs (abr., ang., bio.) *Conserved Non-Genic sequences.* Séquences non-codantes très conservées.

Code génétique (fr., bio.) Succession de nucléotides qui code pour la synthèse d'une protéine.

Compiler (fr., info.) Action de traduire en langage d'assemblage ou langage machine (fichier objet) un programme écrit en langage évolué (fichier source) en vue de son exécution.

Cy3(5) (abr., bio.) Marquer fluorescent du nom de Cyanine 3 ou Cyanine 5.

DAG (abr., ang., math.) *Direct Acyclic Graph.* Graph direct acyclique.

DDBJ (abr., ang., bio., info*) DNA Data Bank of Japan* Banques de données nationale japonaise pour le dépôts des séquences d'ADN.

Déboguer (fr., info.) De l'anglais *debbug*, corriger une erreur dans un programme informatique. (voir bug)

Dendrogramme (fr., math.) Représentation graphique sous la forme d'un arbre planaire hiérarchique.

Déterministe (fr., algo.) Algorithme dont le résultat est une solution finie.(par opposition à indéterministe)

DOE (abr., ang.) *U.S. Department of Energy*, Département américain de l'énergie.

DTD (abr., ang*.,* info*.) Document Type Definition.* Description d'un type de document en SGML.

EBI (abr., ang., bio., info.) *European Bioinformatics Institute,* Institut européen de bio-informatique.

e.g. (abr.) *exempli gratia*, abréviation latine qui signifie « par (le moyen de l'...) exemple ».

Enzyme (fr., bio.) Substance protéique qui catalyse, accélère une réaction biochimique.

EST (abr., ang., bio.) *Express Sequence Tag.* Séquence de fractions exprimées dans un génome.

EVOC (ang., bio.) *Expressed Sequence Annotation for Humans,* ontology pour l'annotation des sequences.

FISH (ang., bio.) *fluorescent in situ hybridisation.* Hybridation fluorescente pour repérer la présence de molécules cibles par un système couplé anticorps-fluorophore.

Formulaire (HTML) (fr., info.) document présentant des champs permettant d'entrer des infos qui seront envoyées au serveur.

Génomique (fr., bio.) Analyse systématique de la composition génétique entière d'un organisme. Elle inclut l'étude des fonctions des gènes dans les cellules, les organes et les organismes.

Génomique inverse (fr., bio.) Identification des gènes à partir des protéines.

GIF (abr., ang., info.) *Graphics Interchange Format***,** format d'images en 256 couleurs très répandu sur Internet car reconnu par tous les navigateurs.

GNU (abr., ang., info.) *GNU in not Unix*, Projet de la *Free Software Foundation* visant à concevoir, réaliser et distribuer un système d'exploitation libre et complet inqpiré d'Unix.

GO (abr., ang., bio.) Gene Ontology. Vocabulaire contrôlé pour décrire les gènes et leurs produits en terme de processus biologiques, fonctions moléculaires, et localisation cellulaire.

GPL (abr., ang., info .) *General Public Licences.* Licence publique et gratuite pour les logiciels informatiques. Elle permet de protéger juridiquement les logiciels libres de l'apposition d'un *copyright* par une société. Cette licence indique que l'utilisateur peut copier, modifier et redistribuer la version modifiée à condition que celle-ci soit également libre : principe du *copyleft*.

GSVD (abr., ang., stat.,) *Generalized Singular Value Decomposition*, Technique de statistiques descriptives visant à réduire les dimensions de l'espace des données par leur décomposition (généralisée) en valeur singulière.

Heuristique (fr., info.) Technique consistant à apprendre petit à petit, en tenant compte de ce que l'on a fait précédemment pour tendre vers la solution d'un problème. L'heuristique ne garantit pas du tout qu'on arrive à une solution quelconque en un temps fini. Opposé à algorithmique.

HGNC (abr., ang., bio.) *HUGO Gene Nomenclature Committee.* Consortium internationale pour standardiser la nomenclature des gènes.

HGP (abr., ang., bio.) *Human Genome Project.* Projet Génome Humain.

Homologue (fr., bio.) Gène ou protéine qui dérive d'un ancêtre commun et qui présente des homologies de structure.

HTML (abr., ang., info.) *HyperText Markup Language.* Langage caractérisé par des balises, utilisé pour écrire des pages Web (forme très simplifiée de SGML).

HTTP (abr., ang., info.) *HyperText Transfert Protocol.* Protocole de niveau applicatif destiné à réaliser des systèmes hypermédia distribués. HTTP est le principal protocole utilisé par le Web.

HUGO (abr., ang., bio.) *Human Genome Organization.*

i.e. (abr.) *id est,* abréviation latine qui signifie « c'est à dire »

Instance (fr., info.) Définitions variables selon le langage de programmation: « une variable de type classe » (Object Pascal), « un objet » (C++), « une variable de type object » (Smalltalk).

Internet (abr., ang., info .) *INTERconnected NETworks.* Réseau international de réseaux interconnectés.

Interpréteur (fr., info.) Programme qui traduit une à une, en langage binaire, au fur et à mesure de leur exécution, les instructions établies en langage évolué.

LIMS (abr., ang., info.) *Laboratory Information Management System.* Système de gestion de l'information dans les laboratoires.

Macro complémentaire Microsoft Excel® (fr., info.) Composants pouvant être installés sur votre ordinateur pour ajouter des commandes et des fonctions à Excel. Ces macros sont propres au programme Excel. Les autres macros complémentaires disponibles sur Excel ou Microsoft Office sont des macros COM (Component Object Model, modèle d'objet composant).

MAD (abr., ang., math.) *Median Absolute Deviation.* Déviation absolue de la médiane.

MAGE-ML (abr., ang., bio., info.) *MicroArray Gene Expression Markup Language*. *DTD XML* uniformisant la description des données expérimentales de puces à ADN, en accord avec le standard international *MIAME*, ayant pour but de faciliter les échanges.

MDS (abr., ang., stat.) *MultiDimensional Scaling.* Technique de statistiques descriptives visant à réduire les dimensions de l'espace des données par un positionnement multidimensionnel.

mer (abr., ang., bio.) Abréviation d'oligomères, dans ce contexte unité de longueur d'un oligonucléotide (1 mer = 1 base nucléotidique)

MeSH (abr., ang., bio.) *Medical Subject Headings.* Thésaurus de référence dans le domaine biomédical, MeSH fournit les descripteurs ou « sujets » pour indexer les articles dans MEDLINE. Outil d'indexation, de recherche et de classement, il est produit par la NLM (National Library of Medicine, U.S.A.) avec la participation de l'INSERM pour la traduction française.

MIAME (abr., ang., bio.) *Minimum Information About a Microarray Experiment*. Standard international définissant les données minimales nécessaires à l'annotation des expérimentations de puces à ADN.

Middleware (ang., info.) Classe de logiciels qui assurent l'intermédiaire entre les applications et le transport des données par les réseaux.

NCBI (abr., ang., bio., info.) *The National Center for Biotechnology Information.* Centre national (américain) pour l'information dans le domaine des biotechnologies.

NHGRI (abr., ang., bio.) *National Human Genome Research Institute* Institut national (américain) de la recherche sur le génome humain.

NIH (abr., ang., bio.) *National Institut of Health.* Institut national (américain) de la santé.

Nucléotides (fr., bio.) Unité de construction des acides nucléiques,

résultant de l'addition d'un sucre (ribose pour l'ARN et désoxyribose pour l'ADN), d'un groupement phosphate et d'une base azotée à l'origine de l'information. Il existe quatre nucléotides différents pour l'ADN : adénine (A), thymine (T), guanine (G), cytosine (C) et quatre nucléotides différents pour l'ARN : uracile (U), guanine (G), cytosine (C), adénine (A). C'est la succession des bases résultant de l'enchaînement des nucléotides dans l'acide nucléique qui constitue le message génétique.

OMIM (abr.,ang., bio.) *Online Mendelian Inheritance in Man.* Service Web recensant les maladies humaines.

Ontologie (fr.) Vocabulaire structuré et contrôlé.

Open-source (ang., info.) Les codes sources du programme sont disponibles auprès du public qui peut les redistribuer et les modifier pour améliorer le programme.

Opéron (fr., bio.) Unité de transcription constituée par un promoteur, un opérateur et un ou plusieurs gènes de structure. Dans le cas de plusieurs gènes de structure, l'ARN messager ainsi produit est polycistronique : ensemble de gènes localisés les uns à la suite des autres, transcrits de façon coordonnée et dont les produits interviennent dans une même voie métabolique.

Pages dynamiques (fr., info.) Pages dont le code source HTML est produit par un programme qui s'exécute lors de la réception de la requête HTTP par le serveur.

Pangénomique (fr., bio.) Représentatif de l'ensemble d'un génome.

PCR (ang., bio.) *Polymerase Chain Reaction.* Réaction d'amplification en chaîne.

Peer-review (ang.) visé par les pairs.

Pléiotropie (fr., bio.) Propriété d'un gène d'agir sur plusieurs caractères.

PHP (abr., ang., info.) *Personal Home Page*, puis *Hypertext PreProcessor*. Langage de script orienté objet permettant de gérer un site Web.

Plug-in (ang., info.) Aussi appelé « greffon ». Logiciel tiers venant se greffer à un logiciel principal afin de lui apporter de nouvelles fonctions. Le logiciel principal fixe un standard d'échange d'informations auquel ses greffons se conforment. Le greffon n'est généralement pas conçu pour fonctionner seul.

Portail (fr., info.) Site Web qui offre un annuaire et/ou un moteur de recherche, des infos et des articles divers et variés, une galerie commerciale.

Programme (fr., info.) Enchaînements de fonctions sauvegardées dans un fichier.

Protéomique (fr., bio.) Analyse systématique de la composition et de la structure des protéines.

Protocole (fr., info.) spécification de la vitesse d'une communication, ainsi que de son codage, son établissement et sa fin.

Replicates (ang.) points de mesure répétés.

SAGE (abr., ang., bio.) *Serial Analysis of Gene Expression.* Méthode basée sur le séquençage et permettant une analyse quantitative de l'expression de gènes.

Séquençage (fr., bio.) Procédé utilisé pour déterminer l'ordre (la séquence) des acides aminés d'une protéine ou des bases dans les acides nucléiques (ADN et ARN).

Séquenceur (fr., bio.) Appareil de séquençage robotisé.

SGBD (abr., ang., info.) Système de Gestion de Base de Données (ex MySQL, PostgreSQL, Oracle).

SGML (abr., ang., info.) *Standard Generalized Markup Language,* méta-langage utilisé pour définir de façon générale des langages définissant des documents hypertextes de toutes sortes, normalisé sous le nom d'ISO 8879. HTML et XML en sont des dérivés simplifiés.

Shell (ang., info.) Interpréteur de ligne de commande, la partie du système d'exploitation utilisé comme interface avec l'utilisateur.

SNP (abr., ang., bio.) *Single Nucleotide Polymorphism.* Modification (substitution, délétion ou insertion) d'un nucleotide.

Spline (ang. math.) Courbe cousine de la courbe de Bézier, passant par un ensemble de points.

SSH (abr., ang., bio.) *Suppressive Substractive hybridization* Hybridation Soutractive Suppressive. Technique permet tant l'obtention de banques soustraites d'ADNc qui sont générées par l'élimination de la plupart des gènes mutuellement représentés entre les situations et/ou les tissus comparés.

SVD (abr., ang., math.) *Singular Value Decomposition.* Décomposition en valeur singulière.

Transcriptome Ensemble des ARN messagers présents dans un type cellulaire donné à un moment donné et dans une condition biologique.

Typage Opération consistant à donner un type à une donnée. Il peut être fort ou faible en fonction de la nécessité de le respecter ou non.

UMLS (abr., ang., bio.) *Unified Medical Language System.* Métathesaurus (réseau sémantique) constitué d'environ 800 000 concepts du domaine médical, définis à partir d'une centaine de terminologies médicales.

UNOS (abr., ang., bio.) *United Network for Organ Sharing. Réseaux unifié pour les transplantation d'organes.* Organisation à but non lucratif pour la gestion des transplantations d'organes.

XML (abr., ang., info.) *eXtensible Markup Language*. Norme d'échange de documents informatisés. XML est un méta-langage permettant de marquer la structure de documents texte de manière arborescente en insérant des "balises" (markup) dans le corps des documents.

Web (abr., ang., info.) *World Wide Web*, « Toile d'araignée mondiale », système basé sur des liens hypertextes, permettant l'accès aux ressources du réseau Internet.

WWW (abr., ang., info.) *World Wide Web.*

Références Bibliographiques

Références bibliographiques

Adams, M. D., Celniker, S. E., Holt, R. A. et al. (2000). The genome sequence of Drosophila melanogaster. *Science*, **287**: 2185-2195.

Adams, R. et Bischof, L. (1994). Seeded region growing. *Conference Proceeding: IEE Transactions on Pattern Analysis and Machine Intelligence,* **16:** *641-647.*

Aitken, S., Korf, R., Webber, B., et Bard, J. (2004). COBrA: a bio-ontology editor. *Bioinformatics*, **21**: 825-826.

Al Shahrour, F., Diaz-Uriarte, R., et Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**: 578-580.

Alizadeh, A. A., Eisen, M. B., Davis, R. E. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**: 503-511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc.Natl.Acad.Sci.U.S.A*, **96**: 6745-6750.

Alter, O., Brown, P. O., et Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc.Natl.Acad.Sci.U.S.A*, **97**: 10101-10106.

Alter, O., Brown, P. O., et Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc.Natl.Acad.Sci.U.S.A*, **100**: 3351-3356.

Ashburner, M., Ball, C. A., Blake, J. A. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet.*, **25**: 25-29.

Auer, H., Lyianarachchi, S., Newsom, D., Klisovic, M. I., Marcucci, G., Kornacker, K., et Marcucci, U. (2003). Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat.Genet.*, **35**: 292-293.

Azuaje, F. (2002). A cluster validity framework for genome expression data. *Bioinformatics*, **18**: 319-320.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., et Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr.Opin.Struct.Biol.*, **14**: 283-291.

Badlou, G (2004). Intégration d'un module de normalisation de données transcriptomiques dans l'application M@IA. DESS Compétence Complémentaire en Informatique, Rennes.

Balasubramanian, R., LaFramboise, T., Scholtens, D., et Gentleman, R. (2004). A graph theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, **20**: 3353-3362.

Baldi, P. et Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t - test and statistical inferences of gene changes. *Bioinformatics*, **17**: 509-519.

Baliga, N. S., Bjork, S. J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M. C., Hood, L., et DiRuggiero, J. (2004). Systems level insights into the stress

response to UV radiation in the halophilic archaeon Halobacterium NRC-1. *Genome Res.*, **14**: 1025-1035.

Banfield, J. D. et Raftery, A. E. (1993). Model-based gaussian and Non-gaussian clustering. *Biometrics*, **49**: 803-821.

Barrans, J. D., Stamatiou, D., et Liew, C. (2001). Construction of a human cardiovascular cDNA microarray: portrait of the failing heart. *Biochem.Biophys.Res.Commun.*, **280**: 964-969.

Bedrine-Ferran, H., Le Meur, N., Gicquel, I. et al. (2004). Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption. *Genomics*, **83**: 772-789.

Bertucci, F., Finetti, P., Rougemont, J. et al. (2004). Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer Res.*, **64**: 8558-8565.

Bicciato, S., Luchini, A., et Di Bello, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*, **19**: 571-578.

Bittner, M., Meltzer, P., Chen, Y. et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**: 536-540.

Bock, J. R. et Gough, D. A. (2001). Predicting protein--protein interactions from primary structure. *Bioinformatics*, **17**: 455-460.

Bolshakova, N., Azuaje, F., et Cunningham, P. (2005). An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, **21**: 451-455.

Bolstad, B. M., Irizarry, R. A., Astrand, M., et Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**: 185-193.

Boulakia, S. C., Lair, S., Stransky, N., Graziani, S., Radvanyi, F., Barillot, E., et Froidevaux, C. (2004). Selecting biomedical data sources according to user preferences. *Bioinformatics*, **20 Suppl 1**: I86-I93.

Brazma, A., Hingamp, P., Quackenbush, J. et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat.Genet.*, **29**: 365-371.

Brazma, A., Parkinson, H., Sarkans, U. et al. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**: 68-71.

Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biol.*, **4**: R41-1-R41-9.

Brosseau, J. (2002). Migration vers un système intégré de stockage et de traitement des données expérimentales de "puces à ADN". « Technologies Avancées des Sciences du Vivant » . Université Louis PASTEUR, **Strasbourg**.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., et Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc.Natl.Acad.Sci.U.S.A*, **97**: 262-267.

Brown, S. M. (2003). Bioinformatics becomes respectable. *Biotechniques*, **34**: 1124-1127.

Burke, S. (2001). Missing Values, Outliers, Robust Statistics & Non-parametric Methods. Statistics and data analysis.

Statistics and data analysis.LC.GC Europe online Supplement., **59**: 19-24.

Butcher, E. C., Berg, E. L., et Kunkel, E. J. (2004). Systems biology in drug discovery. *Nat.Biotechnol.*, **22**: 1253-1259.

Butler, D. (2004). Science searches shift up a gear as Google starts Scholar engine. *Nature*, **432**: 423-

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., et Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res.*, **10**: 2022-2029.

Camon, E., Barrell, D., Lee, V., Dimmer, E., et Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico.Biol.*, **4**: 5-6.

Chaussabel, D., Semnani, R. T., McDowell, M. A., Sacks, D., Sher, A., et Nutman, T. B. (2003). Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, **102**: 672-681.

Chaussabel, D. et Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol.*, **3**: RESEARCH0055-1-RESEARCH0055-16.

Cheng, Y. et Church, G. M. (2000). Biclustering of expression data. *Proc.Int.Conf.Intell.Syst.Mol.Biol.*, **8**: 93-103.

Chiang, D. Y., Brown, P. O., et Eisen, M. B. (2001). Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, **17 Suppl 1**: S49-S55.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., et Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**: 699-705.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat.Genet.*, **32 Suppl**: 490-495.

Claverie, J.-M., Audic, S., et Abergel, C. (1999). La Bioinformatique: une discipline stratégique pour l'analyse et la valorisation des génomes. *Conference Proceeding: Rencontres de Luminy,*

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J.Amer.Statist.Assoc.*, **74**: 829-836.

Cohen, B. A., Mitra, R. D., Hughes, J. D., et Church, G. M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat.Genet.*, **26**: 183-186.

Datta, S. et Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**: 459-466.

Datta, S., Satten, G. A., Benos, D. J., Xia, J., Heslin, M. J., et Datta, S. (2004). An empirical bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **20**: 235-242.

Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E., et Wylie, B. N. (1998). Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, **11**: 259-285.

de Brevern, A. G., Hazout, S., et Malpertuy, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC.Bioinformatics.*, **5**: 114-

DeRisi, J. L., Iyer, V. R., et Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**: 680-686.

Dermitzakis, E. T., Reymond, A., et Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat.Rev.Genet.*, **6**: 151-157.

Diatchenko, L., Lau, Y. F., Campbell, A. P. et al. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc.Natl.Acad.Sci.U.S.A*, **93**: 6025-6030.

Diehn, M., Sherlock, G., Binkley, G. et al. (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**: 219-223.

Ding, C. H. et Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**: 349-358.

Draghici, S. (2003). Data Analysis Tools for DNA Microarrays. First Edition; Chapman & Hall, Boca Raton, Florida.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., et Tainsky, M. A. (2003a). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**: 3775-3781.

Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S., et Tainsky, M. A. (2003b). Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics*, **19**: 1348-1359.

Dudoit, S. et Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, **3**: RESEARCH0036-1-RESEARCH0036-21.

Dudoit, S., Gentleman, R. C., et Quackenbush, J. (2003). Open source software for the analysis of microarray data. *Biotechniques*, **Suppl**: 45-51.

Dudoit, S., Yang, Y. H., Callow, M., et Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, **12**: 111-139.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., et Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nat.Genet.*, **21**: 10-14.

Dysvik, B. et Jonassen, I. (2001). J-Express: exploring gene expression data using Java. *Bioinformatics.*, **17**: 369-370.

Eberwine, J. (1996). Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. *Biotechniques*, **20**: 584-591.

Edgar, R., Domrachev, M., et Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**: 207-210.

Efron, B., Tibshirani , R., Storey, J. D., et Tusher, V. G. (2001). Empirical Bayes analysis of microarray experiment. *Journal of the American Statistical Association*, **96**: 1151-1160.

Eisen, M. B. et Brown, P. O. (1999). DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**: 179-205.

Eisen, M. B., Spellman, P. T., Brown, P. O., et Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc.Natl.Acad.Sci.U.S.A*, **95**: 14863-14868.

Encyclopedia Universalis (1991) Ontologie, In *Encyclopedia Universalis*. **16**:902-910. Universalis, Paris.

Etzold, T., Ulyanov, A., et Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**: 114-128.

Faller, D., Voss, H. U., Timmer, J., et Hobohm, U. (2003). Normalization of DNA-microarray data by nonlinear correlation maximization. *J.Comput.Biol.*, **10**: 751-762.

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., et Vingron, M. (2001). Correspondence analysis applied to microarray data. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 10781-10786.

Fleischmann, R. D., Adams, M. D., White, O. et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**: 496-512.

Fraser, A. G. et Marcotte, E. M. (2004). A probabilistic view of gene function. *Nat.Genet.*, **36**: 559-564.

Friedman, N., Linial, M., Nachman, I., et Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J.Comput.Biol.*, **7**: 601-620.

Galperin, M. Y. (2005). The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res.*, **33 Database Issue**: D5-24.

Gasch, A. P. et Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**: RESEARCH0059-

Gentleman, R. C., Carey, V. J., Bates, D. M. et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**: R80-

Gilbert, D. R., Schroeder, M., et van Helden, J. (2000). Interactive visualization and exploration of relationships between biological objects. *Trends Biotechnol.*, **18**: 487-494.

Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., et De Moor, B. (2004). TXTGate: profiling gene groups with text-based information. *Genome Biol.*, **5**: R43-1-R43-12.

Golbreich, C., Dameron, O., Gibaud, B., and Burgun, A. (2002). Standards et ontologies biomédicales pour le Web Sémantique. Technical report,

Golub, T. R., Slonim, D. K., Tamayo, P. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**: 531-537.

Gress, T. M., Hoheisel, J. D., Lennon, G. G., Zehetner, G., et Lehrach, H. (1992). Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm.Genome*, **3**: 609-619.

Gundavaram, S. (1996). CGI Programming on the World Wide Web. First; O'Reilly

Haab, B. B., Dunham, M. J., et Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.*, **2**: RESEARCH0004-

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., et Young, R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac.Symp.Biocomput.*, 422-433.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., et Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**: RESEARCH0003-

Hedenfalk, I., Duggan, D., Chen, Y. et al. (2001). Gene-expression profiles in hereditary breast cancer. *N.Engl.J.Med.*, **344**: 539-548.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., et Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**: 609-614.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., et Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, **106**: 413-415.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., et Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc.Natl.Acad.Sci.U.S.A*, **97**: 8409-8414.

Houseman, B. T. et Mrksich, M. (2002). Carbohydrate arrays for the evaluation of protein binding and enzymatic modification. *Chem.Biol.*, **9**: 443-454.

Howbrook, D. N., van der Valk, A. M., O'Shaughnessy, M. C., Sarker, D. K., Baker, S. C., et Lloyd, A. W. (2003). Developments in microarray technologies. *Drug Discov.Today*, **8**: 642-651.

Hucka, M., Finney, A., Sauro, H. M. et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics.*, **19**: 524-531.

Hughes, T. R. (2002). Yeast and drug discovery. *Funct.Integr.Genomics*, **2**: 199-211.

Hurowitz, E. H. et Brown, P. O. (2003). Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. *Genome Biol.*, **5**: R2-

Ideker, T., Thorsson, V., Siegel, A. F., et Hood, L. E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J.Comput.Biol.*, **7**: 805-817.

Ihaka, R. et Gentleman, R. (1996). R: a language for data analysis and graphics. *J.Comput.Graph.Statist.*, **5**: 299-314.

Ikeo, K., Ishi-i J, Tamura, T., Gojobori, T., et Tateno, Y. (2003). CIBEX: center for information biology gene expression database. *C.R.Biol.*, **326**: 1079-1082.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**: 860-921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**: 931-945.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.*, **4**: 249-264.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., et Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**: 533-538.

Jenssen, T. K., Laegreid, A., Komorowski, J., et Hovig, E. (2001). A literature network of human genes for high-

throughput analysis of gene expression. *Nat.Genet.*, **28**: 21-28.

Johnson, K. et Lin, S. (2001). Call to work together on microarray data analysis. *Nature*, **411**: 885-885.

Jordan, B. R. (1998). Large-scale expression measurement by hybridization methods: from high-density membranes to "DNA chips". *J.Biochem.(Tokyo)*, **124**: 251-258.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**: 818-821.

Kaufman, L. et Rousseeuw,P.J. (1990). Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, Inc., New York.

Kechris, K. J., van Zwet, E., Bickel, P. J., et Eisen, M. B. (2004). Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biol.*, **5**: R50-

Kenyon, G. L., DeMarini, D. M., Fuchs, E. et al. (2002). Defining the mandate of proteomics in the post-genomics era: workshop report. *Mol.Cell Proteomics.*, **1**: 763-780.

Kerr, M. K. et Churchill, G. A. (2001a). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 8961-8965.

Kerr, M. K. et Churchill, G. A. (2001b). Statistical design and the analysis of gene expression microarray data. *Genet.Res.*, **77**: 123-128.

Kerr, M. K., Martin, M., et Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J.Comput.Biol.*, **7**: 819-837.

Khan, J., Simon, R., Bittner, M. et al. (1998). Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**: 5009-5013.

Khan, J., Wei, J. S., Ringner, M. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat.Med.*, **7**: 673-679.

Kim, H., Golub, G. H., et Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics.*, **21**: 187-198.

Kim, H. K., Choi, I. J., Kim, H. S. et al. (2004). DNA microarray analysis of the correlation between gene expression patterns and acquired resistance to 5-FU/cisplatin in gastric cancer. *Biochem.Biophys.Res.Commun.*, **316**: 781-789.

Kitano, H. (2002). Computational systems biology. *Nature*, **420**: 206-210.

Kluger, Y., Basri, R., Chang, J. T., et Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**: 703-716.

Kohonen, T. (1995). Self Organised maps. 2nd edition, Berlin.

Kononen, J., Bubendorf, L., Kallioniemi, A. et al. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat.Med.*, **4**: 844-847.

Kooperberg, C., Fazzio, T. G., Delrow, J. J., et Tsukiyama, T. (2002). Improved background correction for spotted DNA microarrays. *J.Comput.Biol.*, **9**: 55-66.

Lamirault, G., Steenman, M., Le Meur, N., Demolombe, S., Trochu, J. N., et Leger, J. J. (2004). DNA chip technology in

cardiovascular research. *Arch.Mal Coeur Vaiss.*, **97**: 1251-1255.

Lander, E. S. (1999). Array of hope. *Nat.Genet.*, **21**: 3-4.

Lawrence, C. E. et Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**: 41-51.

Le Bouter, S., Demolombe, S., Chambellan, A. et al. (2003). Microarray analysis reveals complex remodeling of cardiac ion channel expression with altered thyroid status: relation to cellular and integrated electrophysiology. *Circ.Res.*, **92**: 234-242.

Le Bouter, S., El Harchi, A., Marionneau, C. et al. (2004). Long-term amiodarone administration remodels expression of ion channel transcripts in the mouse heart. *Circulation*, **110**: 3028-3035.

Le Meur, N. (2001). Etude comparative des logiciels d'analyse d'images: extraction de données de puces à ADN. Rapport de DEA génomique et Informatique, Rennes.

Le Meur, N., Lamirault, G., Bihouee, A., Steenman, M., Bedrine-Ferran, H., Teusan, R., Ramstein, G., et Leger, J. J. (2004). A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values: importance of replication. *Nucleic Acids Res.*, **32**: 5349-5358.

Lee, M. L., Kuo, F. C., Whitmore, G. A., et Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc.Natl.Acad.Sci.U.S.A*, **97**: 9834-9839.

Legendre, P. et Legendre,L. (1998). Numerical Ecology. second english edition, Amsterdam.

Lercher, M. J., Urrutia, A. O., et Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat.Genet.*, **31**: 180-183.

Leung, Y. F. et Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends Genet.*, **19**: 649-659.

Li, C. et Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 31-36.

Liang, P. et Pardee, A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**: 967-971.

Lieb, J. D., Liu, X., Botstein, D., et Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat.Genet.*, **28**: 327-334.

Liew, C. C. et Dzau, V. J. (2004). Molecular genetics and genomics of heart failure. *Nat.Rev.Genet.*, **5**: 811-825.

Lockhart, D. J., Dong, H., Byrne, M. C. et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat.Biotechnol.*, **14**: 1675-1680.

Lönnstedt, I. et Speed, T. (2002). Replicated microarray data. *Statistica sinica*, **12**: 31-46.

Ma, J. et Liew, C. C. (2003). Gene profiling identifies secreted protein transcripts from peripheral blood cells in coronary artery disease. *J.Mol.Cell Cardiol.*, **35**: 993-998.

Mainard, S. (2003). Etude comparative de méthodes et de logiciels statistiques : analyse de gènes différentiels.

Département Statistique et Traitement Informatique desDonnées, Vannes, Université de Bretagne sud.

Manduchi, E., Scearce, L. M., Brestelli, J. E., Grant, G. R., Kaestner, K. H., et Stoeckert, C. J., Jr. (2002). Comparison of different labeling methods for two-channel high-density microarray experiments. *Physiol Genomics*, **10**: 169-179.

Mantripragada, K. K., Buckley, P. G., de Stahl, T. D., et Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends Genet.*, **20**: 87-94.

Masys, D. R., Welsh, J. B., Lynn, F. J., Gribskov, M., Klacansky, I., et Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics.*, **17**: 319-326.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I., et Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat.Genet.*, **36**: 197-204.

McCreight, E. (1976). A Space-Economical Suffix Tree Construction Algorithm. *Journal of Association Computing Machinery*, **23**: 262-272.

Mi, H., Lazareva-Ulitsky, B., Loo, R. et al. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33 Database Issue**: D284-D288.

Moloshok, T. D., Klevecz, R. R., Grant, J. D., Manion, F. J., Speier, W. F., et Ochs, M. F. (2002). Application of Bayesian decomposition for analysing microarray data. *Bioinformatics*, **18**: 566-575.

Mork, P., Shaker, R., Halevy, A., et Tarczy-Hornoch, P. (2002). PQL: a declarative query language over dynamic biological schemata. *Proc.AMIA.Symp.*, 533-537.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., et Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**: 743-747.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., et Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polmerase chain reaction. *Cold Spring Harbor Symposium on Quantitative Biology*, **51**: 263-273.

Nadon, R. et Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**: 265-271.

Napoli, C., Lerman, L. O., Sica, V., Lerman, A., Tajana, G., et de Nigris, F. (2003). Microarray analysis: a novel research tool for cardiovascular scientists and physicians. *Heart*, **89**: 597-604.

Neuwirth, E. et Baier, T. (2001). Embedding R in Standard Software, and the other way round. *Conference Proceeding: Proceedings of the 2nd International Workshop on Distributed Statistical Computing, 1-10.*

Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P., et Jordan, B. R. (1995). Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics*, **29**: 207-216.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K., et Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics.*, **19**: 2088-2096.

Pan, W. (2002). A comparative review of statistical methods for discovering

differentially expressed genes in replicated microarray experiments. *Bioinformatics.*, **18**: 546-554.

Pan, W., Lin, J., et Le, C. T. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol.*, **3**: RESEARCH0022-

Park, T., Yi, S. G., Lee, S., Lee, S. Y., Yoo, D. H., Ahn, J. I., et Lee, Y. S. (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics.*, **19**: 694-703.

Pasquier, C., Girardot, F., Jevardat, d. F., et Christen, R. (2004). THEA: ontology-driven analysis of microarray data. *Bioinformatics.*, **20**: 2636-2643.

Pavlidis, P. (2003). Using ANOVA for gene selection from microarray studies of the nervous system. *Methods*, **31**: 282-289.

Pavlidis, P., Furey, T. S., Liberto, M., Haussler, D., et Grundy, W. N. (2001). Promoter region-based classification of genes. *Pac.Symp.Biocomput.*, 151-163.

Pavlidis, P., Li, Q., et Noble, W. S. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics.*, **19**: 1620-1627.

Pavlidis, P., Wapinski, I., et Noble, W. S. (2004). Support vector machine classification on the web. *Bioinformatics.*, **20**: 586-587.

Pe'er, D., Regev, A., Elidan, G., et Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics.*, **17 Suppl 1**: S215-S224.

Pennisi, E. (2003). Human genome. A low number wins the GeneSweep Pool. *Science*, **300**: 1484-

Perou, C. M., Sorlie, T., Eisen, M. B. et al. (2000). Molecular portraits of human breast tumours. *Nature*, **406**: 747-752.

Pilpel, Y., Sudarsanam, P., et Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat.Genet.*, **29**: 153-159.

Pinkel, D., Segraves, R., Sudar, D. et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat.Genet.*, **20**: 207-211.

Puthier, D., Joly, F., Irla, M., Saade, M., Victorero, G., Loriod, B., et Nguyen, C. (2004). A general survey of thymocyte differentiation by transcriptional analysis of knockout mouse models. *J.Immunol.*, **173**: 6109-6118.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat.Rev.Genet.*, **2**: 418-427.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat.Genet.*, **32 Suppl**: 496-501.

Quackenbush, J. (2003). Genomics. Microarrays--guilt by association. *Science*, **302**: 240-241.

Ramaswamy, S., Tamayo, P., Rifkin, R. et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 15149-15154.

Raychaudhuri, S., Stuart, J. M., et Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac.Symp.Biocomput.*, 455-466.

Reiner, A., Yekutieli, D., et Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics.*, **19**: 368-375.

Ren, B., Robert, F., Wyrick, J. J. et al. (2000). Genome-wide location and function of DNA binding proteins. *Science*, **290**: 2306-2309.

Ringwald, M., Eppig, J. T., Begley, D. A., Corradi, J. P., McCright, I. J., Hayamizu, T. F., Hill, D. P., Kadin, J. A., et Richardson, J. E. (2001). The Mouse Gene Expression Database (GXD). *Nucleic Acids Res.*, **29**: 98-101.

Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., Valle, G., et Lanfranchi, G. (2003). Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum.Mol.Genet.*, **12**: 823-836.

Roth, F. P., Hughes, J. D., Estep, P. W., et Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat.Biotechnol.*, **16**: 939-945.

Roy, P. J., Stuart, J. M., Lund, J., et Kim, S. K. (2002). Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature*, **418**: 975-979.

Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., et Peterson, C. (2002). BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol.*, **3**: SOFTWARE0003-1-SOFTWARE0003-6.

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M., et Wisman, E. (2001). Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell*, **13**: 113-123.

Schena, M., Shalon, D., Davis, R. W., et Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**: 467-470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., et Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc.Natl.Acad.Sci.U.S.A*, **93**: 10614-10619.

Schmutz, J., Wheeler, J., Grimwood, J. et al. (2004). Quality assessment of the human genome sequence. *Nature*, **429**: 365-368.

Schneider, T. D. et Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**: 6097-6100.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat.Genet.*, **34**: 166-176.

Shaker, R. et al. (2004). The BioMediator System as a Tool for Integrating Biologic Databases on the Web. *Conference Proceeding: Proceedings of the Workshop on Information Integration on the Web,*

Sharan, R., Maron-Katz, A., et Shamir, R. (2003). CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics.*, **19**: 1787-1799.

Shatkay, H. et Feldman, R. (2003). Mining the biomedical literature in the genomic era: an overview. *J.Comput.Biol.*, **10**: 821-855.

Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, H., Safran, M., Domany, E., et Lancet, D. (2003). GeneNote: whole genome expression profiles in normal human tissues. *C.R.Biol.*, **326**: 1067-1072.

Simon, R. (2003). Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br.J.Cancer*, **89**: 1599-1604.

Simon, R. M. et Dobbin, K. (2003). Experimental design of DNA microarray experiments. *Biotechniques*, **Suppl**: 16-21.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat.Genet.*, **32 Suppl**: 502-508.

Smink, L. J., Helton, E. M., Healy, B. C. et al. (2005). T1DBase, a community web-based resource for type 1 diabetes research. *Nucleic Acids Res.*, **33 Database Issue**: D544-D549.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**:

Smyth, G. K., Yang, Y. H., et Speed, T. (2003) Statistical issues in cDNA microarray data analysis., In *Functional Genomics Methods and Protocols*. **9**:111-136. Humana Press, Totowa.

Sorlie, T., Perou, C. M., Tibshirani, R. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 10869-10874.

Spellman, P. T., Miller, M., Stewart, J. et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**: RESEARCH0046-1-RESEARCH0046-9.

Spellman, P. T. et Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. *J.Biol.*, **1**: 5.1-5.7.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., et Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol.Biol.Cell*, **9**: 3273-3297.

Statnikov, A., Aliferis, C. F., et Tsamardinos, I. (2004). Methods for Multi-category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development. *Medinfo.*, **2004**: 813-817.

Steenman, M., Chen, Y. W., Le Cunff, M., Lamirault, G., Varro, A., Hoffman, E., et Leger, J. J. (2003). Transcriptomal analysis of failing and nonfailing human hearts. *Physiol Genomics*, **12**: 97-112.

Steenman, M., Lamirault, G., Le Meur, N., Le Cunff, M., Escande, D., et Leger, J. J. (2005). Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays. *Eur.J.Heart Fail.*, **7**: 157-165.

Stein, L. (2002). Creating a bioinformatics nation. *Nature*, **417**: 119-120.

Stekel, D. (2003). Microarray Bioinformatics. 1st edition; Cambridge University Press, London.

Stevens, R. D., Robinson, A. J., et Goble, C. A. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics.*, **19 Suppl 1**: i302-i304.

Stoeckert, C. J., Jr., Causton, H. C., et Ball, C. A. (2002). Microarray databases: standards and ontologies. *Nat.Genet.*, **32 Suppl**: 469-473.

Storey, J. D. et Tibshirani, R. (2003a). Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol.Biol.*, **224**: 149-157.

Storey, J. D. et Tibshirani, R. (2003b). Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A*, **100**: 9440-9445.

Stuart, J. M., Segal, E., Koller, D., et Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**: 249-255.

Sturn, A. (2000). Cluster analysis for large Scale gene expression studies

Sturn, A., Quackenbush, J., et Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics.*, **18**: 207-208.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., et Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc.Natl.Acad.Sci.U.S.A*, **96**: 2907-2912.

Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., et Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**: 1210-1217.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., et Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat.Genet.*, **22**: 281-285.

Teusan, R. (2002). MADSense: outil informatique d'aide à la compréhension des gènes humains, de leurs fonctions et inter-relations. Rapport de DEA génomique et Informatique, Rennes.

The Arabidopsis genome initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**: 796-815.

The C.elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. *Science*, **282**: 2012-2018.

The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**: 1425-1433.

The yeast genome directory (1997). The yeast genome directory. *Nature*, **387**: 5-

Theilhaber, J., Connolly, T., Roman-Roman, S., Bushnell, S., Jackson, A., Call, K., Garcia, T., et Baron, R. (2002). Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Res.*, **12**: 165-176.

Thomas, J. G., Olson, J. M., Tapscott, S. J., et Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**: 1227-1236.

Thomas, P. D., Kejariwal, A., Campbell, M. J. et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.*, **31**: 334-341.

Tibshirani, R., Hastie, T., Narasimhan, B., et Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc.Natl.Acad.Sci.U.S.A*, **99**: 6567-6572.

Tkatchenko, A. V., Le Cam, G., Leger, J. J., et Dechesne, C. A. (2000). Large-scale analysis of differential gene expression in the hindlimb muscles and diaphragm of mdx mouse. *Biochim.Biophys.Acta*, **1500**: 17-30.

Toronen, P., Kolehmainen, M., Wong, G., et Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**: 142-146.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., et Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics.*, **17**: 520-525.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., et Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc.Natl.Acad.Sci.U.S.A*, **100**: 8348-8353.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., et Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**: 2549-2557.

Tusher, V. G., Tibshirani, R., et Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc.Natl.Acad.Sci.U.S.A*, **98**: 5116-5121.

Ukkonen, E. (1995). On-line Construction of Suffix-Trees. *Algorithmica*, **14**: 249-260.

van de Vijver, M. J., He, Y. D., van't Veer, L. J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N.Engl.J.Med.*, **347**: 1999-2009.

van Someren, E. P., Wessels, L. F., Backer, E., et Reinders, M. J. (2002). Genetic network modeling. *Pharmacogenomics.*, **3**: 507-525.

Vapnik, V. N. (1998). Statistical Learning Theory. Adaptative and Learning Systems for Signal Processing,Communications,and Control., New York.

Velculescu, V. E., Zhang, L., Vogelstein, B., et Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, **270**: 484-487.

Venter, J. C., Adams, M. D., Myers, E. W. et al. (2001). The sequence of the human genome. *Science*, **291**: 1304-1351.

Vey, N., Mozziconacci, M. J., Groulet-Martinec, A. et al. (2004). Identification of new classes among acute myelogenous leukaemias with normal karyotype using gene expression profiling. *Oncogene*, **23**: 9381-9391.

Vilo, J. et Kivinen, K. (2001). Regulatory sequence analysis: application to the interpretation of gene expression. *Eur.Neuropsychopharmacol.*, **11**: 399-411.

Wang, D., Liu, S., Trummer, B. J., Deng, C., et Wang, A. (2002). Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat.Biotechnol.*, **20**: 275-281.

Wang, D. G., Fan, J. B., Siao, C. J. et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**: 1077-1082.

Wang, J., Hu, L., Hamilton, S. R., Coombes, K. R., et Zhang, W. (2003). RNA amplification strategies for cDNA microarray experiments. *Biotechniques*, **34**: 394-400.

Wettenhall, J. M. et Smyth, G. K. (2004). limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, **20**: 3705-3706.

Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., et Brown, P. O. (2003).

Individuality and variation in gene expression patterns in human blood. *Proc.Natl.Acad.Sci.U.S.A*, **100**: 1896-1901.

Williams, E. J. et Bowles, D. J. (2004). Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res.*, **14**: 1060-1067.

Wingender, E., Chen, X., Fricke, E. et al. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**: 281-283.

Workman, C., Jensen, L. J., Jarmer, H. et al. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology*, **3**: 1-16.

Wroe, C. J., Stevens, R., Goble, C. A., et Ashburner, M. (2003). A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac.Symp.Biocomput.*, 624-635.

Xiang, C. C., Chen, M., Kozhich, O. A., Phan, Q. N., Inman, J. M., Chen, Y., et Brownstein, M. J. (2003). Probe generation directly from small numbers of cells for DNA microarray studies. *Biotechniques*, **34**: 386-393.

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. (2000). Comparison of methods for image analysis on cDNA microarray data.

Yang, Y. H., Buckley, M. J., et Speed, T. P. (2001a). Analysis of cDNA microarray images. *Brief.Bioinform.*, **2**: 341-349.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., et Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**: e15-1-e15-10.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. (2001b). Normalization for cDNA Microarray Data. Technical report #589,

Yang, Y. H. et Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat.Rev.Genet.*, **3**: 579-588.

Yeang, C. H., Ramaswamy, S., Tamayo, P. et al. (2001). Molecular classification of multiple tumor types. *Bioinformatics.*, **17 Suppl 1**: S316-S322.

Yeh, I., Karp, P. D., Noy, N. F., et Altman, R. B. (2003). Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics.*, **19**: 241-248.

Yeung, K. Y. et Bumgarner, R. E. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biol.*, **4**: R83-1-R83-19.

Yeung, K. Y., Haynor, D. R., et Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics.*, **17**: 309-318.

Yeung, K. Y., Medvedovic, M., et Bumgarner, R. E. (2003). Clustering gene-expression data with repeated measurements. *Genome Biol.*, **4**: R34-

Yeung, K. Y. et Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics.*, **17**: 763-774.

Zeeberg, B. R., Feng, W., Wang, G. et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**: R28-1-R28-8.

Références Internet

Références Internet

Dernières consultations le 03/04/2005, premières citations dans le texte

Adresse	Nom	Page
Centre de ressources en bio-informatique		
http://www.ebi.ac.uk/	EBI European Bioinformatic Institute	16
http://www.infobiogen.fr/	Infobiogen	16
http://www.ncbi.nlm.nih.gov/	NCBI	16
Bio-informatique		
http://www.gnu.org	GNU operating system	16
http://www.r-project.org/	R project	20
http://www.omegahat.org	Omegahat	22
http://bioconductor.org	Projet bioconductor	22
http://www.cpan.org/	Comprehensive Perl Archive Network	36
http://sbml.org/index.psp	SBML	124
http://www.mged.org/Workgroups/MAGE/mage-ml.html	MAGE-ML	37
http://www.open-bio.org	Open bioinformatic foundation	17
http://www.bioinformatics.org	The Bioinformatics Organization, Inc.	17
http://psb.stanford.edu/	The Pacific Symposium on Biocomputing	122
Bio-informatique et Puces à ADN		
http://www.mged.org/	Microarray Gene Expression Data Society (MGED)	15
http://cardioserve.nantes.inserm.fr/ptf-puce/index.php	Plate forme puce à ADN de Nantes	25
http://www.madtools.org	MADTOOLS MADSCAN-MADSENSE – MADSTAR/BASE	36
http://base.thep.lu.se/	BASE lund	37
http://base.thep.lu.se/plugins/	BASE plugins	37
http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html	Terry Speed's microarray group	22
http://genomicshome.com/	Y.F. Leung's functional genomic	19
http://genopole.toulouse.inra.fr/bioinfo/microarrays	Genopole Toulouse, INRA	110
Ontologies		
http://obo.sourceforge.net	Open Biomedical ontologies	116
http://www.geneontology.org/GO.current.annotations.shtml	Gene Ontology	112
http://www.ebi.ac.uk/GOA	Gene Ontology Annotation (EBI)	114
http://mged.sourceforge.net/ontologies/index.php	MGED Ontology Working Group	116
http://www.godatabase.org/cgi-bin/amigo/go.cgi	AMIGO	114
http://www.geneontology.org /GO.tools.html	GO Tools	114
http://vortex.cs.wayne.edu/projects.htm	OntoTools (Wayne, Univ.)	115
Outils d'analyse de la littérature		
http://www.pubmed.org	Pubmed	117
http://pubgene.org	PubGene™	120
http://scholar.google.com	Google, spécial science	15
http:www.genomatix.de/products/BiblioSphere/	Bibliosphere (Genomatix)	120
http://www.stratagene.com/products/displayProduct.aspx?pid=559	Pathway Assist (Iobion)	120

Recherche de motifs		
http://www.cbil.upenn.edu/tess	Transcription Element Search System	126
http://www.gene-regulation.com/index.html	TRANSFAC	126
Outils de classification		
http://genome.tugraz.at/Software	Genesis	110
http://www.molmine.com	J-express	110
http://www.cs.tau.ac.il/~rshamir/expander/expander.htm	Expander	98
http://visitor.ics.uci.edu/genex/cybert/	Cyber-T	80
Bases de données généralistes		
http://www.ensembl.org/	Ensembl genome browser	16
http://bioinfo.weizmann.ac.il/cards/index.shtml	GeneCards	119
http://www.expasy.uniprot.org	Uniprot Universal Protein Resource	114
http://www.gene.ucl.ac.uk/nomenclature	HUGO	37
http://www.genome.jp/kegg/kegg2.html	KEGG	124
https://panther.appliedbiosystems.com/pathway/	PANTHER pathways	124
Bases de données d'expression		
http://www.ebi.ac.uk/arrayexpress/	EBI microarray data repository data	19
http://www.ncbi.nlm.nih.gov/geo/	Gene Expression Omnibus (GEO)	19
http://cibex.nig.ac.jp/index.jsp	CIBEX (japon)	122
http://genecards.weizmann.ac.il/cgi-bin/genenote/home_page.pl	GeneNote	123
http://bite-it.helsinki.fi	Gene expression in tooth	122
http://www.informatics.jax.org/	Mouse genome Informatics	122
http://t1dbase.org	Resource for Type 1 Diabetes	123
Sociétés commerciales		
http://www.affymetrix.com/index.affx	Affymetrix	8
http://www.home.agilent.com	Agilent	26
http://www.operon.com/products_main.php	Operon	26
http://www.superarray.com/home.php	SuperArray Bioscience corporation	26

ANNEXE :

Manuel du logiciel MADSCAN


MADSCAN ONLINE version5.0
Tutorial for MicroArray Data Suite of Computed Analysis

*** Nolwenn LE MEUR ***

MADSCAN Online version5.0
MicroArray Data
Suite of Computed Analysis

Tutorial

Contact :

Nolwenn Le Meur - Jean Léger

INSERM U533
Faculté de médecine
1, rue Gaston Veil
44035 Nantes cedex
mailto: nolwenn.lemeur@nantes.inserm.fr; jean.leger@nantes.inserm.fr

Table of Contents

INTRODUCTION

During the last decade, microarray technology has been extensively applied to determine gene expression levels in many tissues, animals and diseases. This high throughput technology allows the monitoring of expression levels of thousands of genes simultaneously. Although attention has been paid continuously to microarray data handling, problems still remain. There are many sources of systematic and random variation in microarray experiments that affect the measurements of gene expression. Data processing is therefore crucial to obtain informative data. However, the overflow of data makes manual handling time consuming and error-prone. Bioinformatic tools have thus become essential to deal with data mining and knowledge extraction but not many easy-to-use tools have been developed to efficiently process raw microarray data. We propose MADSCAN (MicroArray Data Suites of Computed ANalysis), a freely available web server that processes raw microarray data to get a consolidated gene expression data matrix (http://www.madtools.org). This dynamic procedure physically validates the quality of the raw data points and the quality of the microarray, it corrects systematic and random biases by normalizing the filtered data, it detects outliers and it statistically validates the expression level of each reporter. The program can be applied to a single microarray or to a batch of replicated microarrays. MADSCAN is written in R (Ihaka and Gentleman, 1996) and Perl. A user-friendly web-interface is implemented in PhP to allow easy access and rapid handling of data.

The first part of this document deals with some general definitions about microarray experiments and the algorithms used in MADSCAN for the processing of microarray raw data. It includes the filtration of the features, normalization, detection of outliers and statistical validation of gene expression. The second part of this document is an illustration of an online analysis performed by MADSCAN. The uploading of raw microarray data files is explained and the obtained results are described. At the end of the document you will find a glossary containing the definition of technical and popular terms used in the microarray field.

For any comments please contact Nolwenn Le Meur at nolwenn.lemeur@nantes.inserm.fr

Chapitre I. Data processing by MADSCAN

I. Microarray terminology and definition

1. Microarray biological element

On microarrays a biological element (entity) is a coding fragment of a gene. It may be a PCR product or synthesized oligos, which sequence is known. For convenience, we will use the term gene reporter instead of "biological" element in this tutorial.

2. Noise

Data issued from microarray experiments tend to be noisy. Noise is introduced at each step of a microarray experiment: mRNA preparation (variation among tissues, kits and procedures), transcription (inherent variation in the reaction, enzymes), labeling (type), amplification, pin type, surface chemistry …These experimental biases may result in, for example, a high background level, coalescent features or a signal heterogeneity across the array. Therefore one of the challenges in microarray data processing is to correct, or at least minimize the noise, so that the observed gene expression variance is due to gene regulation and not to experimental noise.

3. Data processing

According to the MIAME (Minimum Information About a Microarray Experiment) glossary, data processing means "the set of steps taken to process the data, including: the normalization strategy and the algorithm used to allow comparison of all data" (Brazma *et al.*, 2001). Draghiči (2003) defines the pre-processing as the step that extracts or enhances meaningful data characteristics. In general, processing or pre-processing prepares the data for the application of other data analysis methods like clustering.

Data processing in MADSCAN follows five steps (Fig. 1):

- **Filtration** to flag flawed spots and extract information from borderline features (close to the background level or saturation level),
- **Normalization** to minimize experimental systematic and random biases so that the observed variation arises from biological differences rather than from defects in the microarray technology and experimental steps,
- **Scaling** to bring gene expression ratio of different slides at the same variation, in this case to the same median absolute deviation,
- **Outlier detection** to evaluate the consistency of replicates within one array and between replicated arrays.
- **Data integration** to summarize the data. The replicated data points will be summarized using mean and coefficient of variation values per chip and between replicated chips. This step consolidates the data sets and allows the comparison between them.
- **Significant differentially expressed gene(s) detection** to statiscally quantify the evidence of differentially expressed genes.

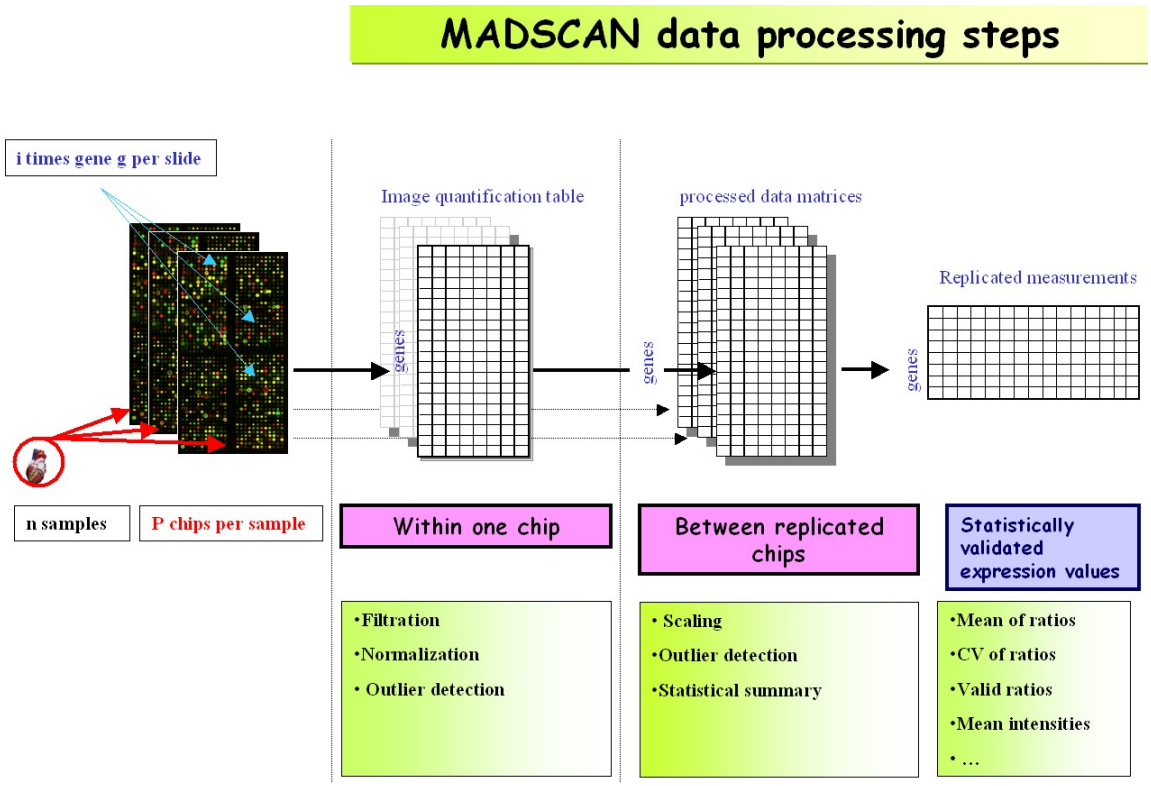


Figure 42. Data processing in microarray experiments. Filtration, normalization and detection of outliers are the required steps to go from raw microarray data to expression data matrices with informative expression level values.

4. Importance of the experimental design

4.1 Experimental layout

The experimental layout is how samples are paired onto arrays and compare to each other. The layout affects the ability to discern and pull apart different sources of variation that could otherwise lead to biased results ((Kerr, 2003; Yang and Speed, 2002)).
The most commonly used experimental layout is the reference design where samples of interest are hybridized to a reference. This design allows indirect comparison of a collection of samples obtained over time (as long the reference sample stay identical). This design also has the advantage to minimize the dye bias, which is one of the greatest sources of variability observed in direct comparison experimental design. In this type of experimental layout, a dye swap approach is crucial to handle dye bias. Other effective designs are "even" designs (Kerr, 2003) such as loop design, where each sample is labeled with both dyes and hybridized in loop. However a loop design is a fragile experimental layout because one bad micoarray and the broken loop avoid easy data analysis.

4.2 Replicates

It is important to design a microarray experiment with replicates (Fig. 2). One needs multiple arrays and multiple spots on each array to have multiple measurements for each reporter under each condition. Multiple measurements from each reporter make it possible to statically assess the quality of the different experimental steps (Lee *et al.*, 2000; Pan *et al.*, 2002).

Most importantly however, one needs biological replicates. They are a means of assessing biological variation of a given condition and of increasing statistical power (Churchill, 2002). Moreover they allow performing sophisticated statistical methods to detect differentially expressed genes. (Dudoit *et al.*, 2000); (Pan, 2002);(Tusher *et al.*, 2001); (Smyth *et al.*, 2003). Replicates should be independent samplings (for example, paired cultures of the mutant (experimental) and the wild type (control) grown on different days). The required number of biological replicates depends on several factors: the desired statistical power to detect differential expression, the desired type I error rate and the statistical method being used to detect change.(Pan *et al.*, 2002). Three biological replicates is the preferred minimum, although sometimes impractical because of the limited supply of precious samples. Technical replicates can improve the precision and the reliability in the measurements at the sample level and thus increase the confidence in the data.

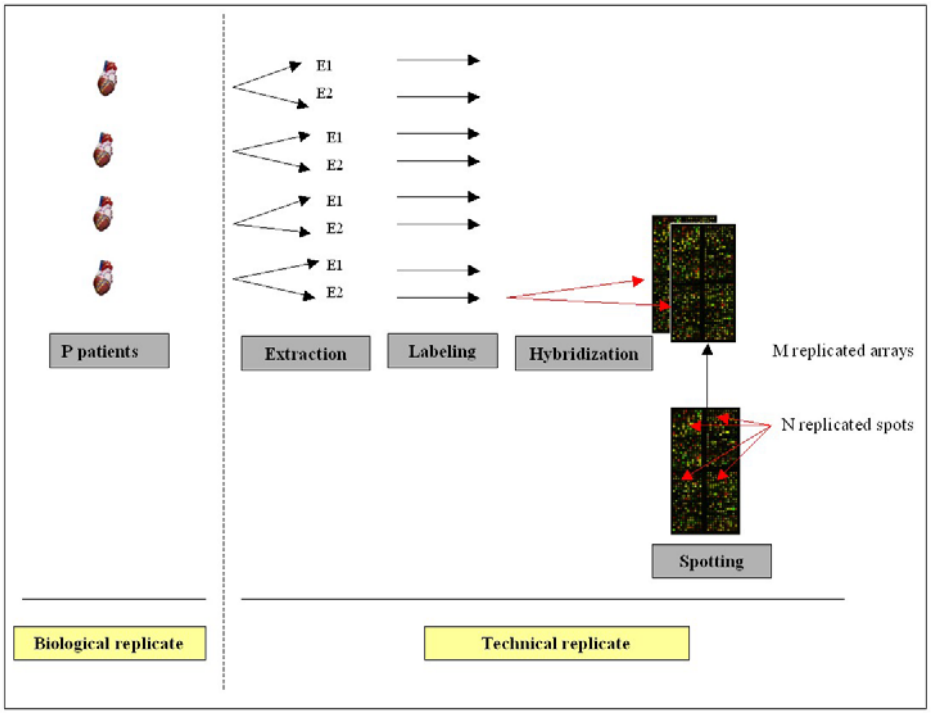


Figure 43. A schematic representation of the different layers of a microarray experimental design.

4.3 Randomization

Randomization is an important concept in statistics. It allows diluting factors that are not of interest but that might influence the outcome of the experiment. Replicates must be printed at random locations throughout the array. If replicates are printed next to each other, a localized defect of the array will affect all replicates making it impossible to distinguish the interesting gene effect from the uninteresting effect of defect. Another example is the use of microarray slides from different batches in experiments comparing a treatment versus a control group. If all control animals are tested using slides from the same batch and all treated animals with a different batch of slides, it will be impossible to distinguish between the uninteresting variability introduced by the slides and the interesting variability introduced by the treatment. These two factors would be confounded.

4.4 Blocking

Blocking is a design technique used to increase the accuracy with which the influence of the various factors is assessed in a given experiment. A block is a subset of experimental conditions that are expected to be more homogeneous than the rest or more under control.

Both blocking and randomizing deal with nuisance factors. However the blocking can only be used when nuisance factors are under control, otherwise randomization remains the only tool available.

5. Microarray layout

Figure 3 defines the different parts, or layout, of a microarray. A microarray is created by an arrayer composed of pins that print the features on the microarray surface. Each pin generates a sub-array. A microarray is therefore constituted of *x* meta-rows and *y* meta-columns of sub-arrays, depending on the number of pins. A sub-array contains *i* rows and *j* columns of features or spots.

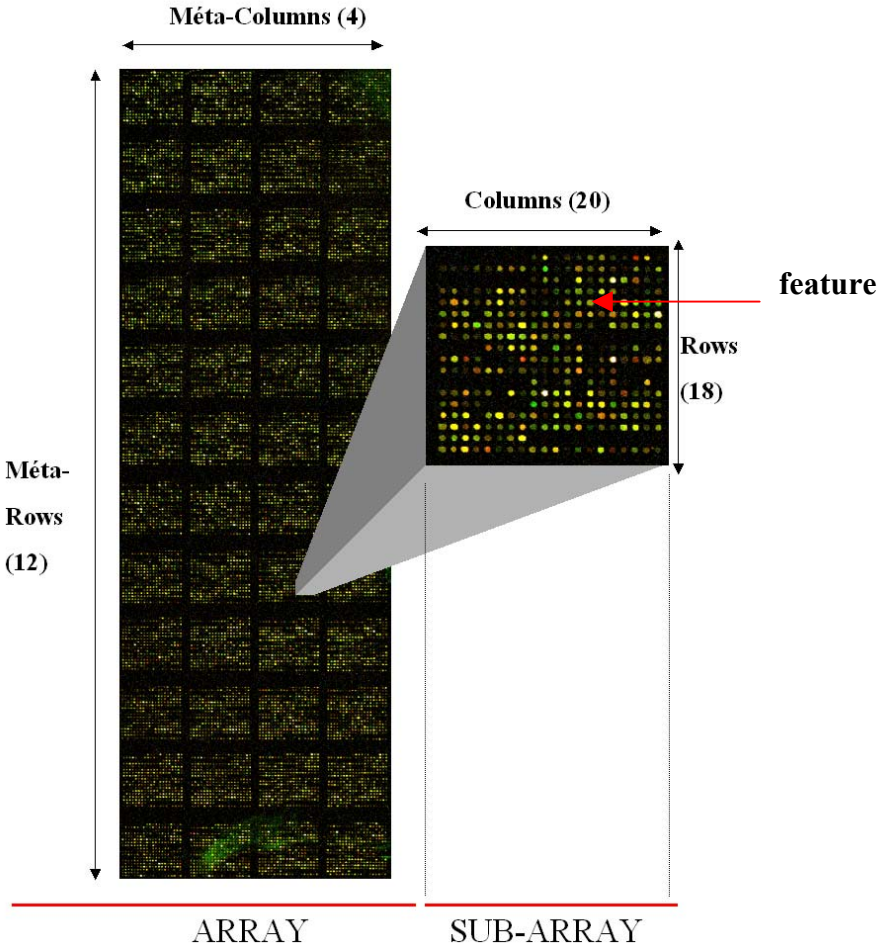


Figure 44. Array layout: a microarray constituted of 12 meta-rows and 4 meta-columns, i.e. 48 sub-arrays. Each sub-array has 18 rows by 20 columns of features, i.e. 360 features. In total this chip contains 17280 features (12*4*20*18).

II. Requirements for MADSCAN analysis

1. Nomenclature and Gene Annotation List (GAL file) for MADSCAN

The creation of a gene annotation list or GAL (Gene Array List) file is a critical point. It is the first source of error in MADSCAN. We recommend that you use the official nomenclature to annotate your genes (http://www.gene.ucl.ac.uk/nomenclature/) or at least that you use gene names of short length without any special characters. The columns “Name” and “ID” must be of the same size, *i.e.* **no empty cells.**

1.1 Example of Nomenclature

We use the official nomenclature to annotate the genes and add characters to identify the different sequences of one gene (1A, 2A, 3A) and the clone of a sequence (1A, 1B). For example the gene coding for the actin, alpha 1, skeletal muscle is labeled **Acta1** and the two different sequences of the gene are labeled in the GAL file as 1A and 2A (Tab. 14)

Table 1. Example of the U533 Nomenclature to label and identify our reporter.

Name	ID
ACTA1_____-1A	R043X01C01
ACTA1_____-2A	R043X01C02
UMPS_____-1A	R043X01G01
PSMB3_____-1A	R043X01K01
PSMD4_____-1A	R043X01O01
RGS11_____-1A	R043X01H01
RGS11_____-1A	R043X01H01
RGS13_____-1A	R043X01P01
MAP4_____-1A	R043X01C13
ZZZTAR043X02D01-1a	R043X02D01
ZZZZZZZZZZZVIDE	ZZZZZZZZZZZVIDE

1.2 Controls and empty "spots"

Negative controls or buffer are features that are known not to be reactive. To evaluate the quality of your negative controls, you must identify them by the Name and ID "**ZZZTA**…" in your GAL file.
If you have *empty features* (incomplete rows of spots in a sub-array), we recommend you to label them as "**ZZZZZZZZZZZVIDE**" to withdraw them from the analysis.

2. Design file format

A tabulated text file is required to define the configuration of your experiment set and your labeling condition. It will help to handle dye-swap experiment. The reference target must be identified as 'ref' and the file must be called **design.txt** (Tab. 15).

Table 2. Design.txt file. The reference target must be called ref.

FileName	Cy3	Cy5
F11.gpr	test	ref
F12.gpr	ref	test
F13.grp	test	ref
F14.gpr	ref	test

3. Data File format

3.1 A-to-Z analysis

As a user of Genepix®, we developed MADSCAN based on Genepix® quality criteria of data measurements. For example, we based the filtration step on the diameter variation and the percentage of satured pixels per feature. Therefore, unless you are using Genepix® for image analysis software you need to prepare a specific file format to enter MADSCAN :

(i) text-tabulated format (.txt),

(ii) use headers as specified in Table 16,

(iii) NAs are accepted in columns 2 to 7,

(iv) for "Block" (column 1) you should indicate the sub-array number.

(v) the "Name" and "ID" columns (2 and 3) must be full, i.e. empty cells are not accepted,

(vi) (vi) for Total and Background Signals (columns 4 to 7) you can use the median intensity given by your favorite image analysis software. In MADSCAN the background intensities are systematically removed from the foreground intensities. In case you prefer not to subtract the background you can put the (G)Rbmed columns at zero (b for background),

(vii) (Vii)Because we are used to take into account the median intensities we labeled the columns (G)Rmed and (G)Rbmed with G and R standing for Green(Cy3) and Red(Cy5) respectively and *med* standing for median. However you can use mean intensities as long as the headers of the columns remain as described below,

(viii) the "Flag" (column 8) is expected to be zero for accepted features and negative for flawed spots.

Table 3. Non-Genepix® file format entry for the A-to-Z analysis.

Block	Name	ID	Rmed	Rbmed	Gmed	Gbmed	Flags
Sub-array number	Gene name	ID	Cy5 intensity	Cy5 background intensity	Cy3 intensity	Cy3 background intensity	Below zero for bad spot and equal zero for good ones

This format also allows performing the filtration and normalization steps in the single step approach.

3.2 Single step file format

Regarding the scaling, the detection of outliers and the data integration in a step by step approach, either you use directly the file resulting from the previous analysis (Tab. 17-5), like the file Normalized_*filename*.txt (Tab. 18), or you use a more general file format (Tab. 19) that contains the Name and ID of the gene reporters along with their different expression values (M in $\log_2$) from the different slides to be analyzed.

Table 4. File formats accepted for the different independent analysis.

Steps	MADSCAN format	Generalized
Filtration	***.gpr**	**Non-Genepix file format *(Tab2)***
Normalization	**Filtration_*filename*.txt**	
scaling	**Normalization_*filename*.txt**	**General file format *(Tab5)***
Outlier detection	**Scaling_*filename*.txt**	
Data integration	**Outlier_*filename*.txt**	
Differentially expressed genes?	**GeneExpression_*filename*.txt** *target.txt (ref 3.3)*	

Table 5. MADSCAN File format for the scaling, detection of outliers and data integration steps

Spot	Block	Name	ID	Rmed	Rbmed	Gmed	Gbmed	Rnorm	Gnorm	A	M	score
1	1	G1	ID1	8500	212	8359	328	7181	9216	12.99	0.36	2
2	1	G2	ID2	867	195	1028	309	683	707	9.44	0.05	2
3	1	G3a	ID3	11176	212	10822	325	9443	12203	13.39	0.37	2

Table 6. General file format for single step analysis (except filtration and normalization and steps)

Name	ID	M1	M2	M3	…	M….
Gene name	ID	Ratio slide #1	Ratio slide #2	Ratio slide #3	…	Ratio slide #n

3.3 Detection of differentially expressed gene module

The module “detection of differentially expressed genes” allows you to evaluate the significance of your gene expression values and estimate the contrast(s) between different factors that characterize your samples and might influence their gene expression level. Therefore, along with the gene expression data file (general file format Tab6.) this module needs a second file describing the

samples (Tab. 20). This file must be named **target.txt**. It contains the information concerning the samples (diagnosis, treatment or not, time course) and/or experimental factors that might influence the expression results (experimenter, date).

Table 7. target.txt file format for the detection of differentially expressed gene analysis

Sample	Factor1	Factor2	Factor3	…	Factor n
S1	drugA	T1	drugB	…	
S2	drugA	T1	nondrugB	…	
S3	drugA	T2	drugB	…	
S4	drugA	T2	nondrugB	…	

III. Data transformation, processing and visualization in MADSCAN

1. Data transformation and processing

1.1 Background subtraction

The background fluorescence signal usually originates from non-specific hybridization of the labeled samples or auto-fluorescence of the glass slide. The unwanted background signal needs to be estimated and removed from foreground signal during image analysis. In MADSCAN, the median of the background intensities of a spot is systematically subtracted from its foreground intensities.

1.2 Log transformation

The log transformation decouples the variance and the mean intensity. Fold changes occurring around small intensity values will be comparable to similar fold changes occurring around large intensity values(Draghici, 2003). A second and very strong argument in favor of the log transformation is related to the distribution of the values. The log transformation makes the distribution symmetrical and almost normal (Fig. 4) (Long *et al.*, 2001; Yang *et al.*, 2001a). Finally, a third argument in favor of using the log transformation is convenience. If the log is taken in base 2, the later analysis and data interpretation are greatly facilitated. For instance, selecting with a 4 fold variation can be done by cutting a ratio histogram at the value log2 (4) = 2. In MADSCAN, the ratio is always calculated and displayed in log base 2.

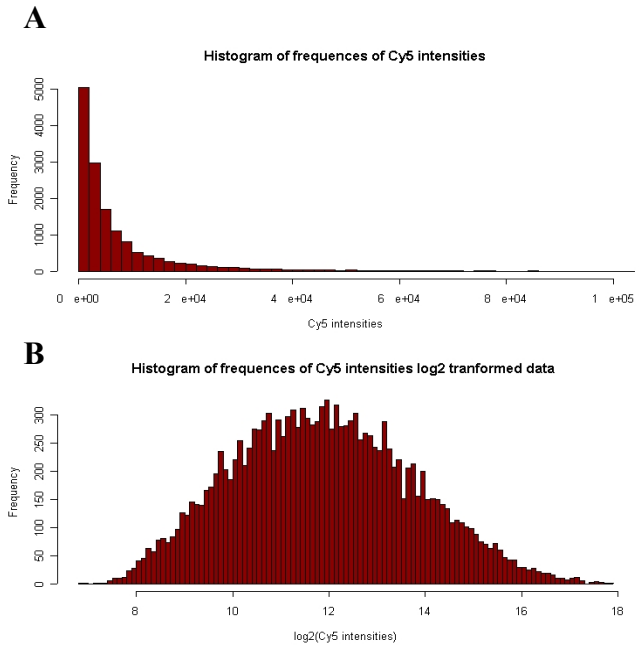


Figure 45 Effect of the log transformation on the distribution of the Cy5 intensity values. (A) Shows the histogram of the background corrected intensity values. The distribution is skewed towards high intensity values. (B) Shows the same distribution after the log transformation.

1.3 Processing methods

Table 21 presents the processing steps available in MADSCAN. These analyses can be performed one by one or all in one step. Table 22 shows the different combinations of analysis available according to your experimental design. Designing experiments with replicates will allow you to test the robustness of your measurements and will greatly increase the confidence in your data. You will be able to perform a complete MADSCAN analysis and go from raw microarray data matrices to consolidated gene expression data matrix.

Table 8. Methods used by MADSCAN for microarray data processing.

Process	Algorithm	Application	Reference
Filtration	Raw data scoring and filtering	within	*In house*
Normalization	Rank invariant Method	within	(Tseng *et al.*, 2001)
	Spatial approach definition	within	*In house*
	Scaled lowess fitness	within	(Yang *et al.*, 2002)
Scaling	MAD scaling	between	(Yang *et al.*, 2001b)
Outlier detection	MAD modified z-scored test Grubbs' test	within/between	(Burke, 2001)
Differentially expressed genes ?	Limma eBayes	between	(Smyth, 2004)

Table 9. Available processing steps according to the data set.

Number of chips	Replicated features within one chip	Available processing: From Filtration to …
1 chip	0	Scaling
1 chip	n ≥ 3	Outlier detection **within** one slide
n different chips	0	Scaling
n different chips	n ≥ 3	Outlier detection **within** one slide
m replicated chips (m ≥ 3)	0	Outlier detection **between** slides and data integration
m replicated chips (m ≥ 3)	n ≥ 3	Outlier detection within and between slides i.e. **both** approaches and data integration

2. Data visualization

2.1 MA plot

Single slide expression data are typically displayed by plotting the log intensity in the red channel (log_2R) versus the log intensity in the green channel (log_2G) (Fig. 5A). Dudoit *et al.* (2000) showed that the log intensity ratio $M = log_2R/G$ vs. the geometric mean log intensity $A = log_2\sqrt{RG}$ [1] is more accurate to reveal spot artifacts and detect intensity dependent patterns in the log ratios (Fig. 5B). An M vs. A plot, also called RI plot for Ratio-Intensity plot, is a 45° counterclockwise rotation of the (log_2G,log_2R)-coordinate system, followed by the scaling of the coordinates (Quackenbush, 2002). MADSCAN proposes to display two MA plots for each analyzed chip, corresponding to the distribution of the data before any data processing and after your last step.

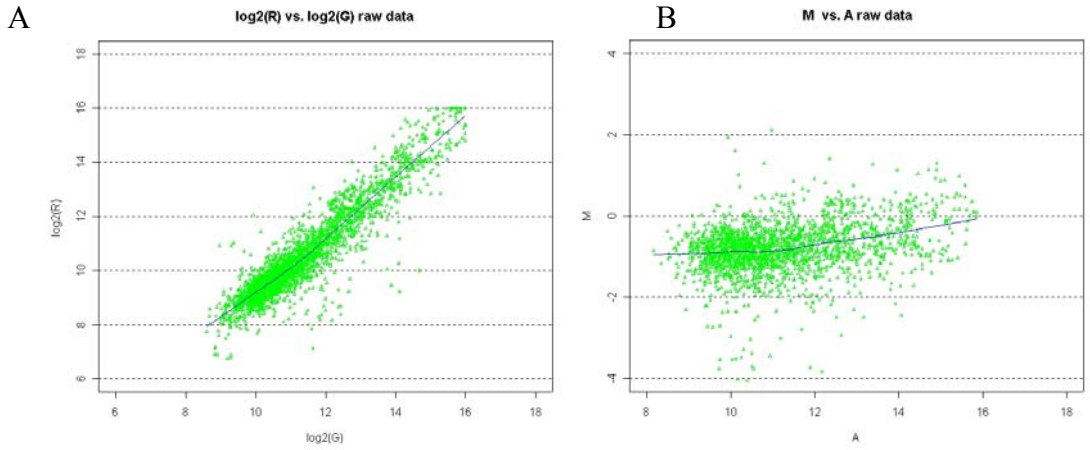


Figure 46. R vs G or MA plot. (A) Represents the typical R vs G plot in log2. (B) Shows the M vs A plot for the same data. The MA plot clearly shows that the data distribution is intensity dependent.

[1] The letter M is for *minus* as $M = log_2R - log_2G$ while A is for *add* as $A = (log_2R + log_2G)/2$.

2.2 Box plot

The box plot, also known as box and whiskers plot, is a graphical display of five statistical descriptors (Fig. 6). The line in the box is the 50th percentile, i.e. the median, 50% of the data are contained below and above this line. The height of the box, i.e. the distance between the 25th and 75th percentiles, is known as the inter-quartile range or inter-quartile distance (IQD). The length of the tails or whiskers is usually 1.5 times the IQD**.** Data points that fall beyond the whiskers are considered outliers (or in the case of microarrays potentially differentially expressed genes).

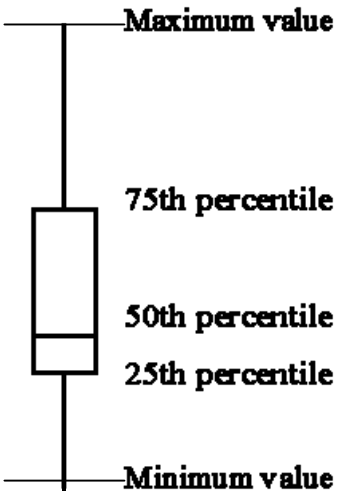


Figure 47. Schema of a box plot. Definition of the symbols.

2.3 Density plot

The density plot presents the distribution of the ratios, which must be close to a normal distribution after the normalization step (Fig. 7).

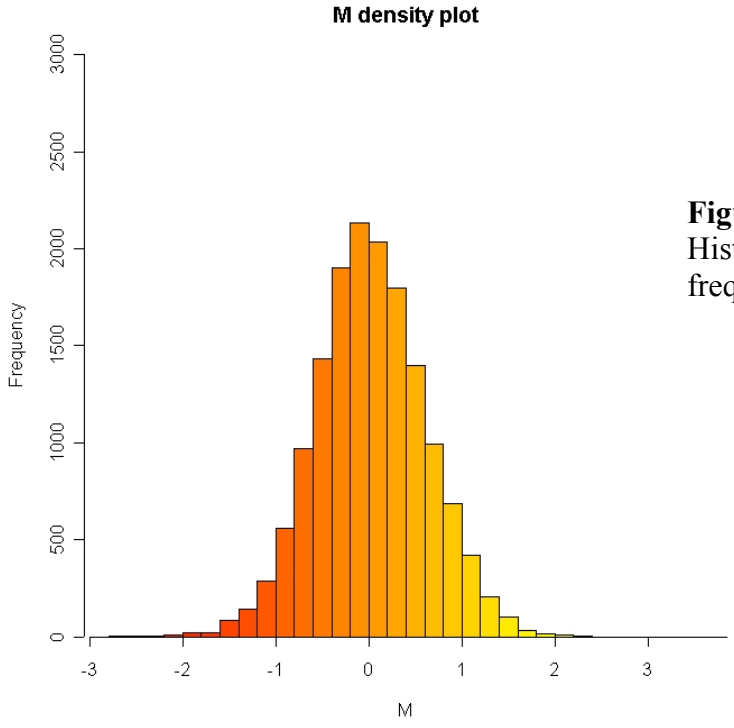


Figure 48. Density plot of ratios (M). Histogram of the distribution of the frequencies of the intensity ratios.

IV. Algorithms in MADSCAN

1. Filtration

The Filtration step aims to flag flawed spots and extract information from borderline features in raw data matrices. The borderline spots are those with very weak expression levels, close to the background level, or those with very high expression levels, *i.e.* close to the saturation level. A suite of algorithms based on quality criteria extracted from Genepix® image analysis software and built like a decision tree allows the scoring of each spot according to its quality. The main criteria are the background level, the signal to noise level, the diameter and the saturation level.
Non-Genepix® user can still perform a complete analysis with MADSCAN as long as they use the file format previously defined (see Chapter I. III Data transformation 2. File format). In that case, the quality criteria taken into account are: the flag given by your favourite image analysis software or your filtration criteria, the background and signal intensity levels.

MADSCAN first estimates the overall quality of the raw data before any filtering and then applies our scoring procedure.

1.1 Estimation of the overall quality of the raw data (before any processing)

For each print-tip group is calculated:

- Spots flagged during image analysis,
- Median background in Cy3 and Cy5 (Bg R(G)),
- Median signal to noise for both channels (R(G) s/n) (Eq. 1),
- Coefficient of variation of the Cy3 and Cy5 signal intensities (CV) (Eq. 2.)
- Mean spot diameter (Dia.)
- Standard deviation of the diameter (Dia. SD)

$$\text{Cy3(5)S/N} \equiv \frac{\text{Cy3(5)Median Intensity - Cy3(5)Median Background Intensity}}{\text{Cy3(5) Background Standard Deviation}} \quad \textbf{Eq 1}.$$

$$\text{CV} \equiv \frac{\text{Standard deviation (Cy3(5) Median Intensity)}}{\text{Mean(Cy3(5) Median Intensity - Cy3(5) Median Background Intensity)}} \quad \textbf{Eq. 2}$$

1.2 Scoring procedure

The filtration is performed step by step following a decision tree with a scoring procedure (Fig. 8). Each feature is tested against a suite of quality criteria (image analysis flags, signal minus background level, signal-to-noise level, diameter variation and saturation level) and gets a score of 0 to 5 according to its quality (Tab. 23). A score of 0 means that the feature failed to pass a quality criterion and is therefore removed from further analysis.

Table 10. Features scores according to their quality.

Score	Quality
0	Poor
2	Validated
3	Close to saturation in Cy3
4	Close to saturation in Cy5
5	Close to saturation in both channels

The first criterion of acceptance for a spot is whether or not it has been flagged by the image analysis software. A marked spot gets a score of 0 and is removed from further analysis.

The second factor is the signal intensity threshold. The estimation of the sub-array local background allows defining signal intensity thresholds. The background corrected spot intensity level must be at least two times higher than its median background intensity (Eq. 3). To prevent highly differential genes removed from the analysis, a feature is kept if one of its intensities is below the threshold but the other has its signal intensity level at least five times greater than its median background intensity.

$$F_{bi}Cy3_{(5)} > 2\ Bg_{bi}Cy3_{(5)} \qquad \textbf{Eq. 3}$$

where i is the i^{th} spot in the b^{th} sub-array.

The third cut-off is the signal-to-noise level. The background corrected spot intensity must be at least three times higher than its median background standard deviation.

$$F_{bi}Cy3_{(5)} > 3\ SD\ Bg_{bi}Cy3_{(5)} \qquad \textbf{Eq. 4}$$

where i is the i^{th} spot in the b^{th} sub-array.

The fourth cut-off is based on the spot diameter. Because each sub-array comes from different print-tips and the print-tips deliver a slightly different amount of probe, the diameters of the spots are heterogeneous. A confidence interval is calculated, by sub-array, around the mean diameter of the spots. A spot is flagged if its diameter is lower than the print-tip mean diameter minus 50 or higher than the print-tip mean diameter plus75.

$$50 < \text{spot diameter} - \text{print-tip mean diameter} < 75$$

Finally the abundantly expressed genes get particular attention (Tab. 24.). Spots with more than 90% of satured pixels in both channels are removed from the analysis (score=0). A feature highly expressed in only one of the two channels gets a score of 3 or 4 (for Cy3 or Cy5 respectively). The estimate of the regression ratio (Rgn) is then used instead of the regular ratio of the medians. A spot approaching saturation in both channels has a score of 5 and its expression ratio is re-calculated. We made the hypothesis that if there are more satured pixels in one channel that means the gene is more expressed in that channel. Its expression level is then the ratio of the percentage of satured pixels in each channel (Eq. 5). If M is the log base 2 intensity ratio for a gene, M_{adjust} is the adjusted expression ratio:

$$Madjust \equiv \log 2\left(\frac{(\%satured\_pixel\_in\_Cy5)}{(\%satured\_pixel\_in\_Cy3)}\right) \qquad \textbf{Eq. 5}$$

Table 11. Handling of saturated spots. Features are labeled as satured if their intensity in at least one of the two channels is above 6500 and contains less than 90% of satured pixels.

Cy5	Cy3	Score	Estimated ratio
> 65000	> 65000	5	%satCy5 / %SatCy3
> 65000	< 65000	4	Rgn Ratio
< 65000	> 65000	3	Rgn Ratio

NB: If score = 5 and %SatCy3(5) = 0 then Ratio =Rgn Ratio

The background corrected intensities are re-evaluated due to the property of the geometrical mean of the intensities which is constant for a given gene in a given experiment (Dudoit *et al.*, 2000). If A is the geometrical mean of the background corrected intensities, $A=\log_2(RG)^{1/2}$ with R=Cy5 and G=Cy3, the new Cy5 and Cy3 (R_{raw} and G_{raw}) are:

$$Graw \equiv (2^{2\times A / 2^{Madjust}})^{1/2} \quad \text{and} \quad R_{Raw} = G_{raw} x 2^{Madjust} \qquad \textbf{Eq. 4}$$

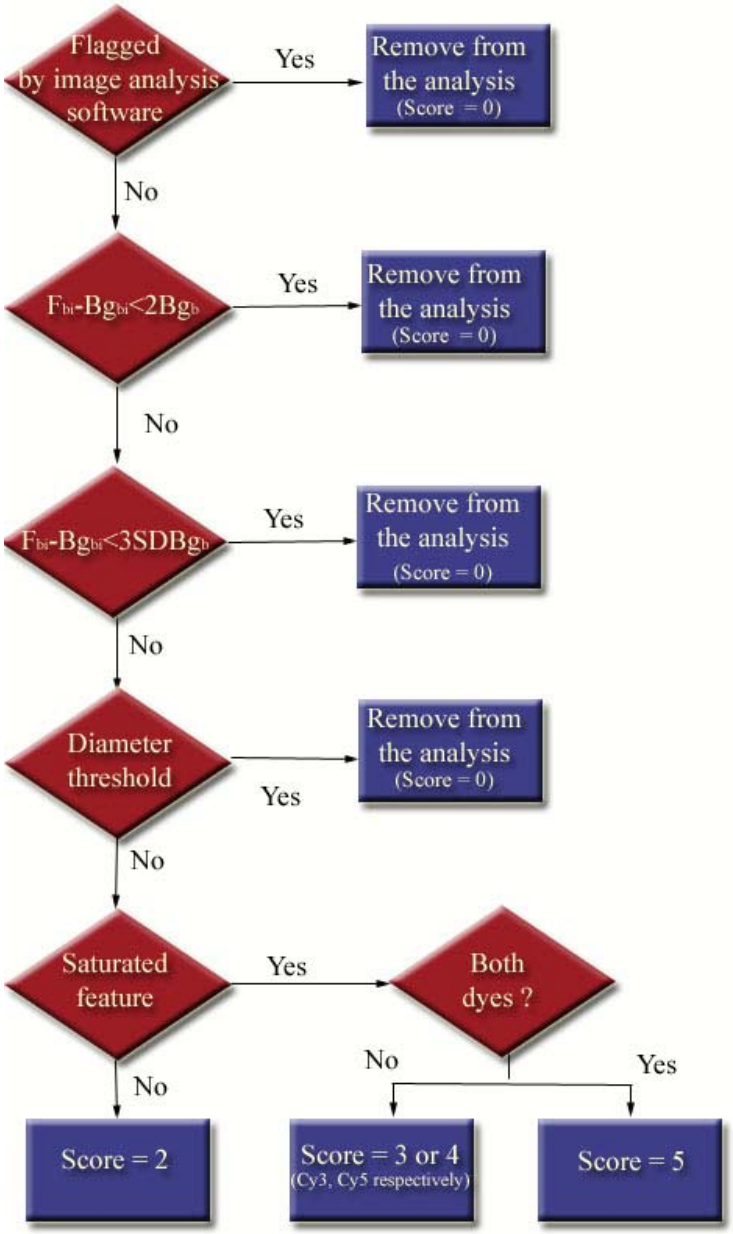


Figure 49. Decision tree for data filtration

2. Normalization

The purpose of normalization is to minimize systematic experimental biases so that the observed variation arises from biological differences rather than from defects in the microarray technology. MADSCAN addresses three main aspects of the normalization process, i.e.:

- selection of a set of reporters as reference to perform the normalization
- spatially dependent biases
- intensity-dependent biases

2.1 The rank invariant method

The selection of a suitable control set is critical to perform the normalization procedure. Traditional methods based on intensity of housekeeping genes often show sample-specific biases (Yang *et al.*, 2002). A method to select *a posteriori* a set of invariant genes seems more efficient. In our program, we have adapted the Rank Invariant Method developed by Tseng *et al.* (2001) to select invariant genes. The set of invariant genes is selected *a posteriori*, i.e. after the raw data filtration. When the number of genes is small (<4000) the rank invariant function will be non-iterative. For a larger number, the estimation of invariant genes can be more sophisticated. The algorithm becomes iterative. A decision tree allows selecting the algorithm in order to maximize the number of invariant genes. This selected set of genes will then be used to achieve the normalization step (Fig. 9).

a) Non-iterative approach

The ranks of Cy3 and Cy5 intensities of each gene on the slide are separately computed. For a given gene if the ranks of Cy3 and Cy5 intensities differ less than a threshold *d* and the rank of averaged intensity is not among the *l* highest ranks or the *l* lowest ranks, this gene is classified as a non-differentially expressed genes (Eq. 4)

$$S \equiv \left\{ g : \left| rank(Cy5_g) - rank(Cy3_g) \right| < d \;\&\; l < rank\left[(Cy5_g + Cy3_g)/2 \right] < G - l \right\} \quad \text{Eq. 4}$$

d=5, *l*=5, G = the highest rank

b) Iterative approach

For a larger number of genes (~>4000), an iterative algorithm will select a more conserved set of genes (Eq. 5 & 6).

For the first loop:

$$S_0 \equiv \left\{ g : \left| rank(Cy5_g) - rank(Cy3_g) \right| < p \times G \;\&\; l < rank\left[(Cy5_g + Cy3_g)/2 \right] < G - l \right\} \quad \text{Eq. 5}$$

For iteration ***i*** to the k^{th}

$$S_i \equiv \left\{ g : g \in S_{i-1} \;\&\; \left| rank_{g \in S_{i-1}}(Cy5_g) - rank_{g \in S_{i-1}}(Cy3_g) \right| < p \times \left| S_{i-1} \right| \right\} \quad \text{Eq. 6}$$

where $\left| S_{i-1} \right|$ is the number of genes in set S_i. The iteration stops at the k^{th} step when $\left| S_k \right| \equiv \left| S_{k-1} \right|$. The set of genes S_k is the rank invariant set.

p=0.05, *l*=10

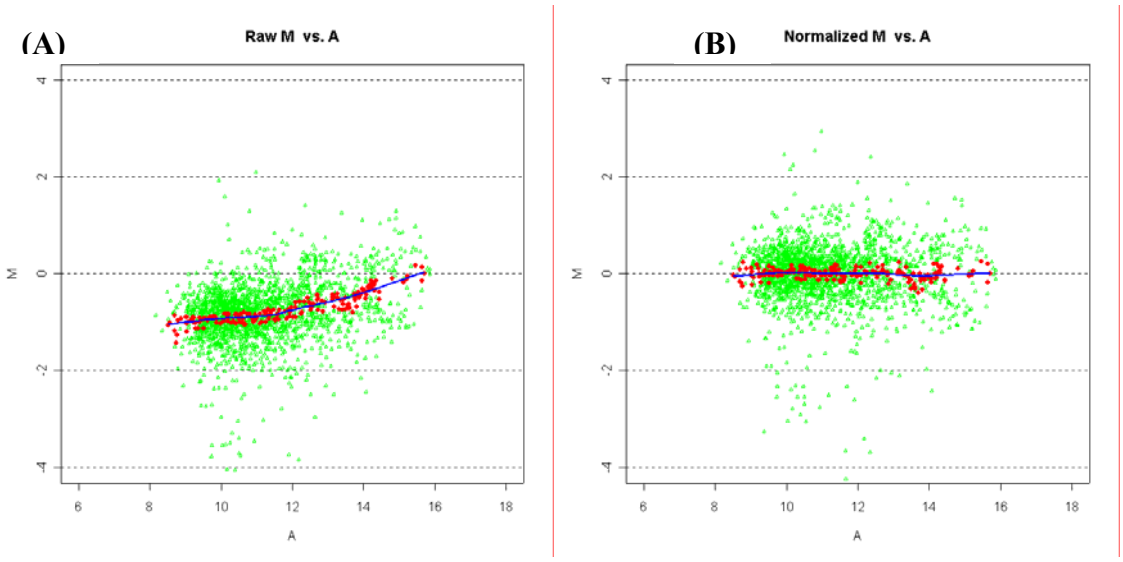


Figure 50. Invariant reporter estimation and Intensity dependent normalization. The invariant genes are estimated *a posteriori* by the Rank Invariant Method. (a) Representation of the selected invariant genes (red dots) among the raw data to be normalized (green dots). The lowess fitness curve (blue line) is built on the invariant genes. (b) Representation of the data normalized based on the invariant genes.

2.2 Spatial normalization

Spatial normalization aims to correct spatially dependent dye biases. Unique normalization across the slide presumes a uniform grading of systematic error all over the array. However the imbalance in the red and green intensities is usually not constant across the slide and can vary according to the overall spot intensities, the location on the array, the plate origin and other geometrical-related variables (Yang *et al.*, 2001b). For a relatively heterogeneous level of signal across the slide the difference between a global (array) and a local (print-tip group) normalization might be significant (Fig. 10A & B).

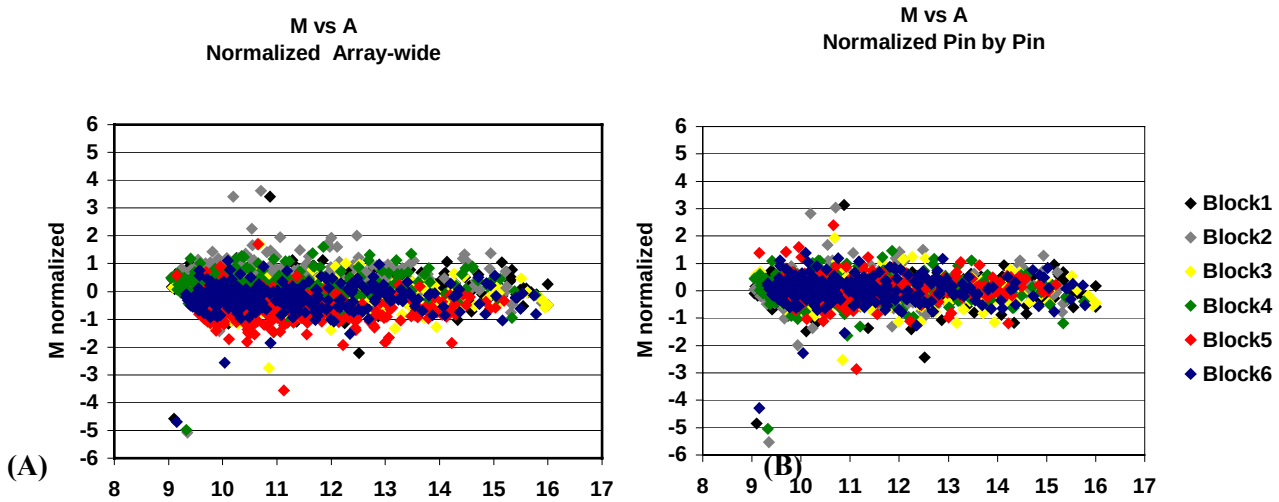


Figure 51. Comparison of a normalization performed on the entire array (A) and a print-tip group normalization (B). The graph of the global normalization (a) shows that one can distinguish the blocks whereas the local normalization (b) allows to correct the print-tip group bias.

In MADSCAN, the normalization is done either globally, pin by pin (print-tip group) or by proximal approach depending on the number of invariant spots. A sufficient number of invariant genes (at least 50%) by sub-array are needed to obtain a robust normalization curve. However the number of invariants in a block can be low or even insufficient for a satisfactory normalization (Yang *et al.*, 2002). In order to correct this defect it is sometimes preferable to include in the algorithm the invariant genes of the blocks adjacent to the one studied (Fig. 11.), i.e. by proximal approach.

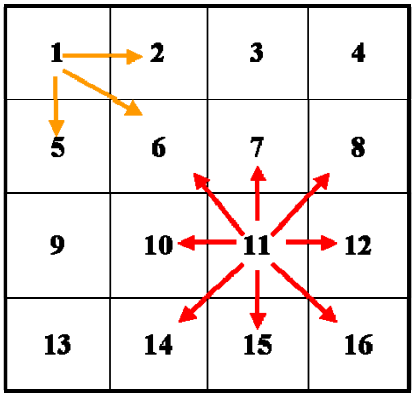


Figure 52 . Illustration of the selection of genes for a proximal approach. If there are an insufficient number of invariant genes in a block, the gene population of the adjacent blocks is used to estimate the invariant genes.

The choice of the spatial method is done according to the pre-estimation of the number of invariant genes by sub-array (Fig. 12). First, for a microarray, the percentage of invariant genes per sub-array is estimated. Then if the sub-array containing the greatest percentage of invariants has less than 20% of invariant genes the analysis will be performed globally. If the greatest percentage of invariant gene is above 50% the analysis will be applied locally, *i.e.* sub-array by sub-array. Otherwise the analysis will be done by the proximal approach (some blocks will be normalized with the invariant genes of neighboring blocks, others will be done independently i.e. sub-array by sub-array). The method applied to normalize is indicated in the table of results under the label "Normalization mode".

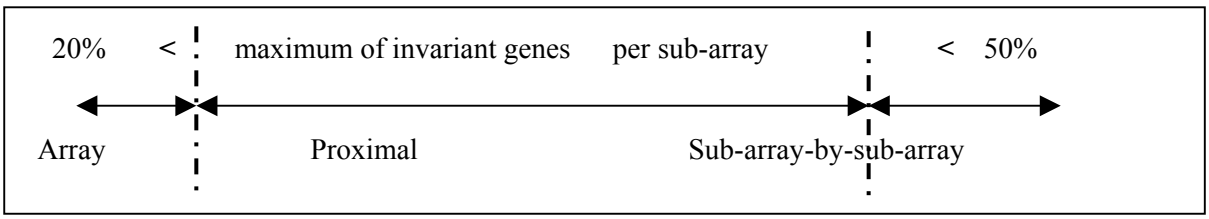


Figure 53. Interval of application for the different spatial approaches of normalization.

2.3 *Lowess* fitness normalization

Normalization has initially been done by correcting the data with a scaling factor or a simple linear regression (Chen *et al.*, 2001). However, the use of linear regression involves many assumptions (normality of the population, linearity of intensity range) to perform accurate fit. Dudoit *et al* (2000) showed that the distribution of the expression ratio is intensity dependent. They proposed to use an intensity-dependent normalization, the Lowess fitness algorithm, which is more appropriate to correct data (Yang *et al.*, 2002); (Kerr *et al.*, 2000).

a) Lowess fitness normalization

Lowess fitness stands for Locally-Weighted Estimated or Locally-Weighted regression (LWR). Historically, local regression is an old data smoothing method proposed back in 1829. In the 1950's it was used as the *Kernel* function and weight function. It was in the 1970's at AT&T that Cleveland and Devlin implemented the *Lowess* function to improve data visualization in time series analysis.

b) Lowess fitness and microarray

The *Lowess fitness* method is applied on the invariant set of genes selected by the rank invariant method (Tseng *et al.*, 2001). Instead of using a global linear normalization (Eq. 7):

$$\mathrm{Log_2R/G} \Leftrightarrow \quad \log_2\mathrm{R/G}-c=\log_2\mathrm{R}/(kG) \qquad \textbf{Eq. 7}$$

where c is the median or the mean of the intensity log ratios M.

The normalization curve function is intensity-dependent, using A, the geometrical mean of the intensities:

$$\mathrm{Log_2R/G} \quad \Leftrightarrow \quad \log_2\mathrm{R/G}-c(A)=\log_2\mathrm{R}/[k(A)G] \qquad \textbf{Eq. 8}$$

Where c(A) is the lowess fit to the M-A plot.

A span f is defined as the fraction of data used to smooth at each data point. We are currently using f =0.4 , i.e. 40% of the data around a point.

For the within-print tip group normalization, *i.e.* spatial approach, the lowess fit simply becomes (print tip+A)-dependent (Eq 9.).

$$\mathrm{Log_2R/G} \qquad \Leftrightarrow \qquad \log_2\mathrm{R/G}-c_i(A)=\log_2\mathrm{R}/[k_i(A)G] \qquad \textbf{Eq. 9}$$

Where i is the i^{th} sub-array.

Finally we normalize according to:

$$\widetilde{M} \equiv \widetilde{f}(A) \qquad \text{between } \min_{g \in S} A_g \text{ and } \max_{g \in S} A_g \qquad \textbf{Eq. 10}$$

An extrapolation is performed based on the 50 genes with the highest and lowest average log intensity ranks selected set from non-differentially expressed genes. If the average log intensity > $\max_{g \in S} A_g$ (or < $\min_{g \in S} A_g$), a linear fitting is performed $M = \alpha + \beta A$ in the subset $T = \{ g:g \in S \ \& \ \mathrm{rank}_{g \in S}(A)_g > |S| - 50\}$

3. Scaling

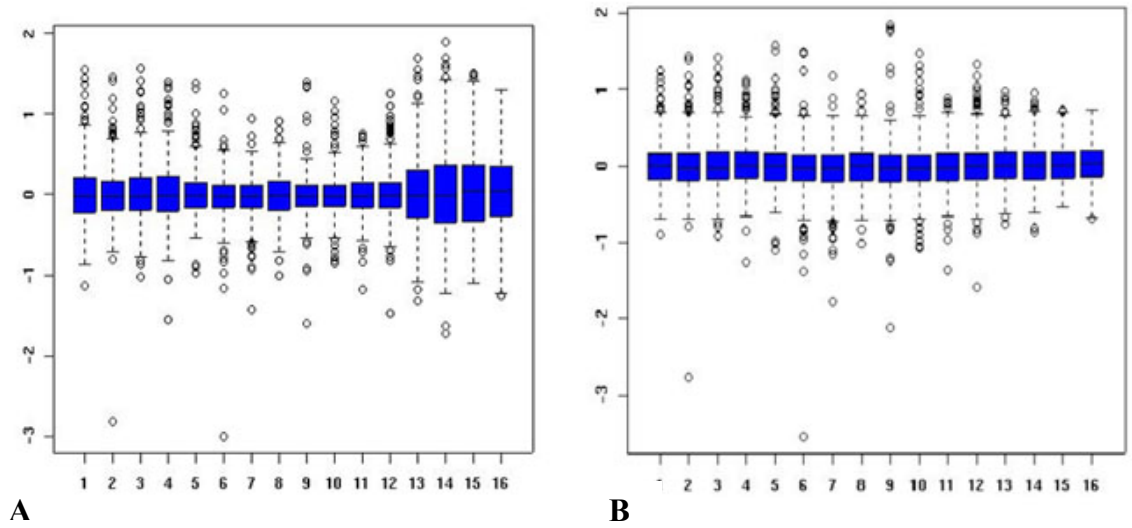
The scaling aims to bring the internal variance within or between slides within the same range. It allows to compare between slides in an experiment. The scaling is also considered as a normalization of the data (Draghici, 2003) . The box plot is an example of graphical display to visualize the spread of the variance before and after scaling (Fig. 13). The use of the median

absolute deviation (MAD) and the geometrical mean (GM) to perform the scaling has been shown to be more efficient than the usual standard deviation (Yang *et al.*, 2001b) (Eq. 11).

$$MAD \equiv median\left\{\left|x_i - \bar{x}_m\right|\right\} \quad \text{where } \overline{x_m} \text{ the median the } x_i \quad \textbf{Eq. 11}$$

$$GM = prod(MADx_{ij})\text{^}(1/n)$$

$$Scaled\_x_{ii} = (x_{ji}/MAD(x)_{ij})*GM$$



A **B**

Figure 54**.** Why scaling data ? **(A) Before any scaling: 16 replicated arrays have been normalized (median centered around zero) but there are different variances. (B) After scaling: the within-variance array is set to the same level, i.e. the median absolute deviation is reduced to the same level.**

4. Outlier detection

A well-thought experimental design has replicated features within a slide and replicated slides. This allows to statistically validate the replicates within and between slides and to label possible outliers in the filtered and normalized quantification data matrices. However, due to the small number of replicates available in microarray experiments, the use of modified statistical tests are required. In MADSCAN we propose two different tests: the MAD Z-test or median absolute deviation modified Z-test and the Grubbs' test that are described hereafter.

4.1 Outliers and Outlier tests

An outlier is a suspect point in term of its relative distance to the mean value (Fig. 14). Outlier tests tell you where you are most likely to have technical errors; they do not tell you that the point is 'wrong'. No matter how extreme a value is in a set of data, the suspect value could nonetheless be a correct piece of information. We propose to label outliers with different tests and compare the results. Some values may be considered as outliers with one test but not with another. An outlier can mask an other thus the detection can be apply iteratively until no outlier is detected but the power of the tests decreases as the number of repetitions increases.

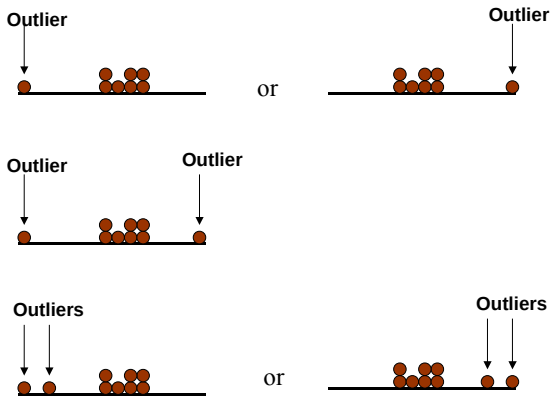


Figure 55. Outliers and masking.

4.2 Median Absolute Deviation (MAD)

The MAD value is an estimate of spread in the data similar to the standard deviation (Burke, 2001) (Müller, 2000). It is an estimator more resistant to outliers than the standard deviation, usually used to perform z-test and detect outliers. The MAD test is a modified z-test where the median absolute deviation is used in place of the standard deviation in the z-score calculation.

$$MAD \equiv median\left\{\left|x_i - \bar{x}\right|\right\}$$

$$z \equiv \frac{0.675\left(x_i - x_m\right)}{MAD}$$

if $\left|z_i\right| \geq 3.5$ then x_i is an outlier

4.3 Grubb's test

This method is also called the ESD or extreme studentized deviate. Grubb's test is recommended by the US Environmental Protection Agency (EPA) as a statistical test for outliers (US EPA, 1992). Grubb's test will take on one of two forms depending on whether the largest value in the dataset is suspected or if the smallest value is suspected.

If the largest value is suspected:

$$T_n \equiv \frac{\left(x_n - \bar{x}\right)}{\sigma}$$

If the smallest value is suspected:

$$T_1 \equiv \frac{\left(\bar{x} - x_1\right)}{\sigma}$$

where x_n is the largest value,

x_1 is the smallest value

$\bar{x}$ is the mean of all the n values

σ is the sample standard deviation of all n values

The critical value for the test performed depends on the sample size n and the selected significance level (Annex 2). A value (ratio) is labeled as outlier if its value T_i is higher than the tabled critical value.

5. Data integration

The data integration step summarizes the data in a consolidated data matrix. It prepares the data to allow comparison of data sets. This will allow the application of further data analysis methods such as statistical tests to detect differentially expressed genes or clustering approaches to group and visualize patterns of gene expression.
MADSCAN creates a file that contains the name of the genes and their different value rows: the within slide median of ratios in $\log_2$ (M), the within slide coefficient of variation (cv), the between slide median ratio and CV. The same data for the geometric mean of the intensities (A).

6. Detection of differentially expressed genes

Fold change calculation is the simplest and most intuitive approach to find the genes that are differentially regulated between control and experiment (Draghici, 2003). Typically, an arbitrary threshold such as 2 fold is chosen and the difference is considered as significant if it is larger than the threshold. This method is often used because it is simple and intuitive. However, arbitrarily choosing the threshold may often be inappropriate. For example, a low signal to noise ratio for genes with low expression levels often generates a distribution of gene expression log ratio in a funnel shape. The variance of the genes with low expression values will be higher than those with high expression level. Genes of low expression values will tend to be less reliable. Moreover, the use of a constant threshold will reduce the sensitivity of detection (more false positives and false negatives at low level and high gene expression level respectively).
An other method is to calculate **unusual ratio**, i.e. select the genes for which the ratio experiment/control is far from the mean experiment/control ration (~ +/- 2σ). However this approach is sensitive to noise and will also generates a important number of false positives and false negatives.
Thus methods based on statistics are needed to evaluate the false positives and false negatives among the called genes. Moreover, due to the fact that in microarray experiment we have more variables (genes) than conditions (patients) one need to perform multiple testing adjustment. Actually numerous statistical tools already exist. Among them, the most well known software are *SAM* (Tusher *et al.*, 2001), *PAM* (Tibshirani *et al.*, 2002), *multtest* (Dudoit *et al.*, 2002) or *Limma* (Smyth, 2004). Because the first two are already implemented in R and as an Excel add-ins, and because the *multtest* package is less efficient than *Limma* for the analysis of multi-factorial design experiments, we choose to implement the Gordon Smyth ‘s algorithm in MADSCAN.

6.1 Limma

Limma is a Bioconductor (http://bioconductor.org) library for the analysis of gene expression microarray data (Smyth *et al.*, 2003). It especially use linear models for the analysis of designed experiment and the assessment of differential expression. Limma provides the ability to compare between many RNA targets simultaneously. A graphical user interface, Limma GUI, is also available (Wettenhall and Smyth, 2004). However this graphical interface tool is not really flexible if one only want to use the algorithm to detect differentially express gene.

Limma use linear models and contrasts to accommodate complex microarray experiments involving multiple RNA sources. Empirical Bayes shrinkage of the gene-wise residual variances is provided to ensure stable results even when the number of arrays is small (Smyth, 2004)l. The sample standard deviations of the implemented moderated t-statistics is shrunken towards a pooled standard deviation value. For more details on the method please read Smyth,G.K.’s paper at http://www.bepress.com/sagmb/vol3/iss1/art3/

6.2 Multiple testing correction

To detected differentially expressed genes in microarray, thousand of comparisons are made that raise the chance of committing at least one Type I Error, i.e. false positive detection (Tab. 25). Individual p-values of 0.01 no longer correspond to significant findings. Correction for multiple testing are thus needed to adjust p-values and control family wise Type I error rate.

Table 12

Number of genes	Gene significant level			
	p-values<0.01	0.05	0.1	0.15
10	<1	<1	1	1.5
20	<1	1	2	3
50	<1	2.5	5	7.5
100	1	5	10	15
500	5	25	50	75
1000	10	50	100	150
5000	50	250	500	750
10000	100	500	1000	1500

D'après Drăghici (Chapman &Hall 2003)

The proposed methods to adjust p-values are:

Bonferroni The p-values are multiplied by the number of comparisons. It is a quite simple but very conservative correction, for thousands of genes, these methods are not pratical.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561–576. (An excellent review of the area.)

Sarkar, S. (1998). Some probability inequalities for ordered MTP2 random variables: a proof of Simes conjecture. *Annals of Statistics*, **26**, 494–504.

Sarkar, S., and Chang, C. K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, **92**, 1601–1608.

FDR *(Benjamini & Hochberg* The method of Benjamini and Hochberg (1995) controls the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family wise error rate.

Holm
Strong control of the family wise error rate, valid under arbitrary assumptions and less conservative than Bonferroni approach.

Hocheberg or **Hommel**
Hochberg's and Hommel's methods are valid when the hypothesis tests are independent or when they are non-negatively associated (Sarkar, 1998; Sarkar and Chang, 1997). Hommel's method is more powerful than Hochberg's, but the difference is usually small and the Hochberg p-values are faster to compute.

none Pass-through option

Chapitre II. MADSCAN online

MADSCAN analysis can be done either step by step or from A to Z, i.e. one can apply one test at a time or ask for running all the steps in a single but complete procedure. The "A to Z" approach was preferentially developed based on experiments where replicated chips had a common reference, i.e. on a 'reference design', with possible 'dye-swap'. Nevertheless, other experimental designs, such as 'time series' or 'loop' designs, can be analyzed by the A to Z approach up to the normalization step (included) or outlier testing within slide. The following steps (outlier detection 'between' and data consolidation) can be achieved step by step with reformatted normalized data files (see Chap I - II Requirements for MADSCAN analysis - Table 2).

You have access to the different analysis (forms) thought the MADSCAN menu on the left frame of your Internet navigator. You can also access the MADSCAN tutorial, the references for the method used, how to cite MADSCAN and the different options to restart analysis or logout (Fig. 15). In this chapter we will describe how to fill out the A to Z analysis form and describe with some files of results coming from an online data analysis. The other forms, the "single step" approaches, are close to the "A to Z analysis". Regarding online help, we invite you to perform an online demo **and use the online help button for more details**.

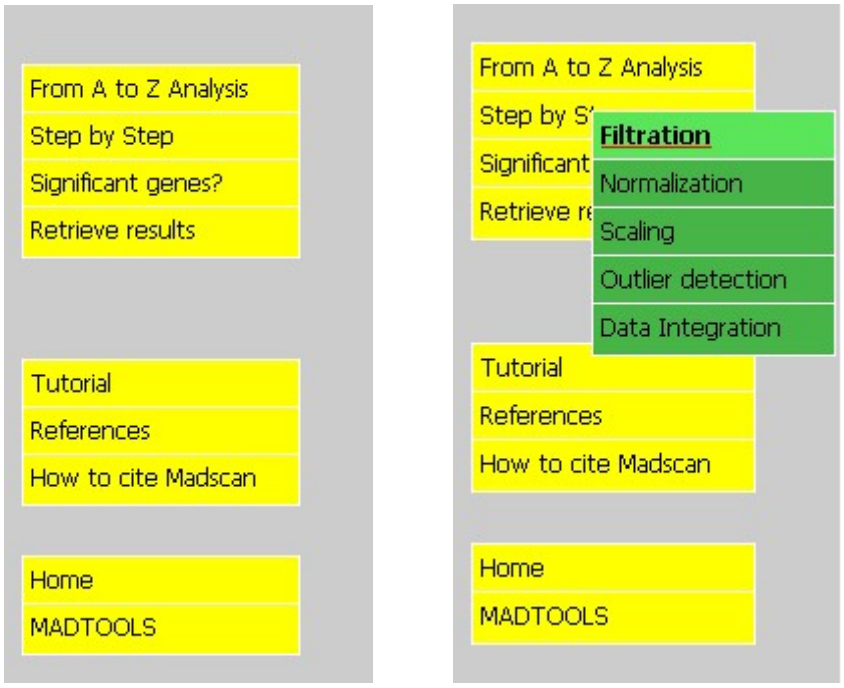


Figure 56. MADSCAN Menu**: By clicking on the "Step by step Analysis" link you will have access to the different individual steps.**

I. Data uploading

MADSCAN was developed with **Microsoft Internet Explorer 6.0 ®.** We recommend you to use IE 6.0 or higher. Functioning under other web browsers is not guaranteed.

1. Submission format

In addition to the format of the microarray data matrix resulting from your favorite image analysis software (*.gpr or tabulated text files –see Chapter III §II Requirements), some formatting is required to upload the data in MADSCAN. The input files must be compressed in a directory containing either *.gpr files or text files. The compressing format must be **zip**. This way the data can be analyzed slide by slide or in a batch of slides (replicated slides or not).
The output file will be a ***.tgz** directory containing text files. It is a highly compressed format. We recommend **WinRar ®** for the compressing and uncompressing of the different directories.

2. User information and warnings

During any dynamical data processing an animated hourglass will advice you that the process is currently running (Fig. 16). Please be patient, it might take up to 30 minutes to complete a "From A to Z" analysis for a batch of 3 slides with 20000 spots each, depending on the traffic.

Normalization of AML1-2_normalization.zip:
Process is running, please wait ...

Figure 57. Hourglass, displays of a currently running analysis.

Some warning messages will appear in case of error in the format data entry such as filename starting with a numeric and/or containing spaces, missing information *etc* (Fig. 17A and B).

ERROR

Missing filename !

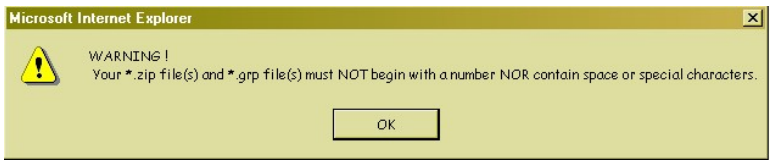


Figure 58. Warnings for misfiling the data input form.

II. From A to Z analysis

1. Data input

Most of the forms present the same fields to be filled out (Fig. 18; Tab. 26). You can still look at the help button to remind you how to fill out the form. The resulting files will depend on the analysis performed. The help button of the page on results also gives an explanation of the results displayed on the result screen.

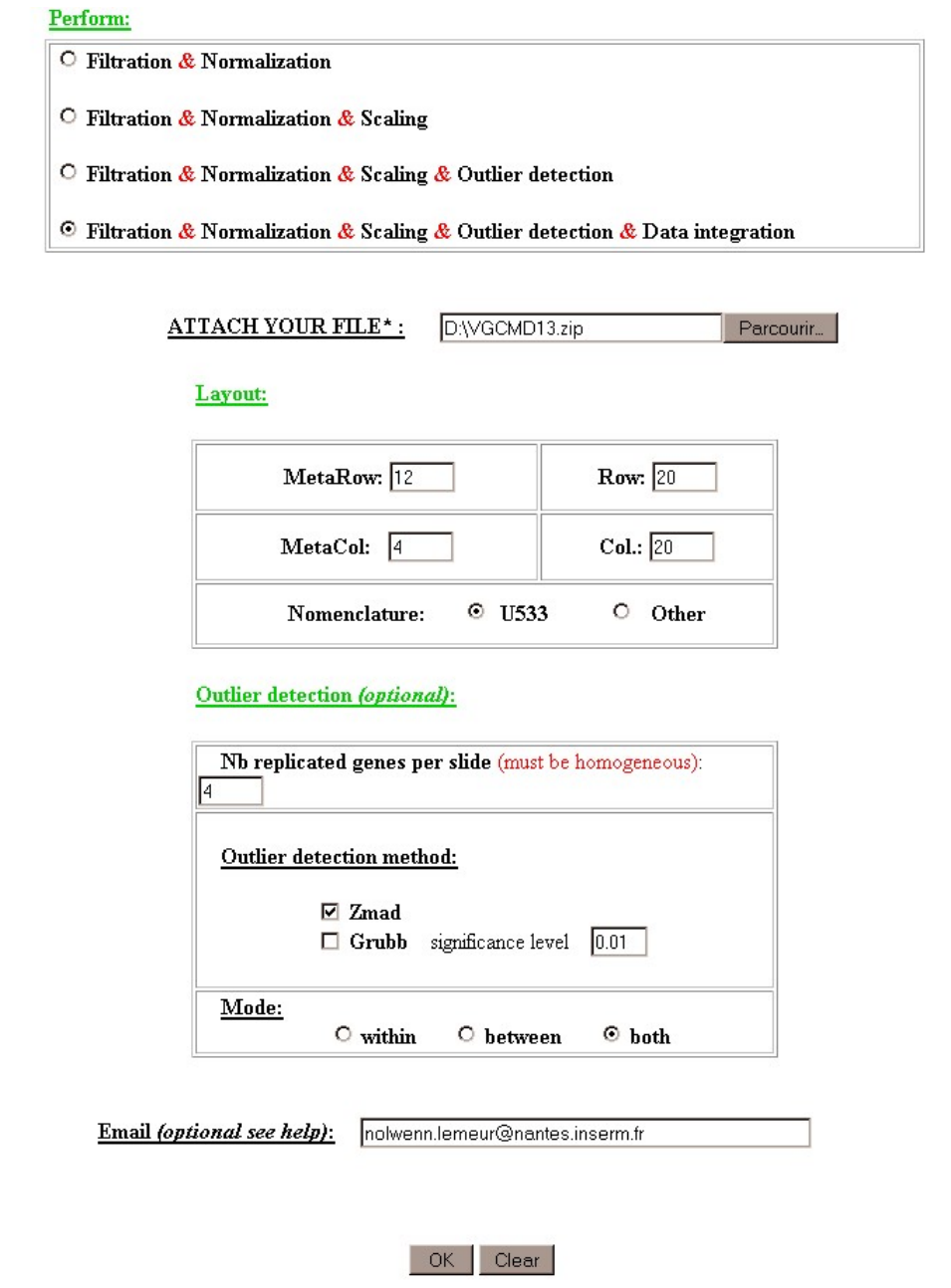


Figure 59. Form for the "A to Z" analysis.

Table 13. Fields of the different forms to be filled out to perform an analysis.

MetaRow	Number of rows of sub-arrays
MetaCol	Number of columns of sub-arrays
Row	Number of rows per sub-array
Col	Number of columns per sub-array
Format	**Filtered by MADSCAN**, i.e. coming directly from the first step, or **raw data file** that at least contains the columns ("Spot", "Block", "Name", "ID", "Rmed", Rbmed","Gmed", "Gbmed"). **Normalized by MADSCAN**, i.e. coming directly from the first step, or **raw data file** that at least contains the columns ("Spot", "Name", "ID", "Rnorm","Gnorm","A","M").
Number of replicated genes per slide	Number of times one gene is replicated within a slide. The number of replicates must be identical for each gene.
Zmad	Estimation of outliers by a modified z-test (use of the MAD, **M**edian **A**bsolute **D**eviation, instead of the mean).
Grubbs'	**Grubbs'** test for detecting outliers or ESD method (extreme studentized deviate) determines how far the outliers are from the others. The Z-test is used to calculate the differences and a modified table of critical values allows estimating the outliers.
Mode (outlier detection and scaling)	**Within:** one array **Between:** arrays Both: first within one array then between arrays
Reference (outlier test)	Column used as reference for the identity of the genes. It can either be the Name or the ID column. The column chosen as reference must be "clean", i.e. preferably official symbols.
Attach your file	Only Genepix or tabulated text files are accepted. Your filename should not start with a number nor contain spaces. Concerning GAL file. Avoid special character such as '#','/','?' in your gene name. NA is accepted.
Email	Entering your email address will allow you to retrieve your results in case of network falling.

2. Results

MADSCAN generates different results available under different formats such as online tables that summarize the performed analysis; downloadable text files and graphs. The downloadable text files can directly be accessed after the analysis run-time or ,if you requested it by entering an e- mail in the form, can be retrieved later on (within the week of the analysis) with the job identifier obtained by e-mail notification.

2.1 Main window

Table 27A & B are an example of the data output. They summarize the overall quality of the data before and after processing. These two tables, along with other files of results, are downloadable from this same main results window via the button "Download Results" (see 2.3 Download results).

The upper table on the results page resumes the data quality before any processing step. In our example, four slides have been analyzed in a batch. The percentage of flagged spots give an idea of the number of spots available for further analysis. You can find results on your buffer or negative controls if you have followed our nomenclature ("ZZTA…", "ZZZ…"). You can get the number of buffer, controls detected and the median intensity level of the buffer, which must be close to the channel background intensity. The signal to noise intensities, the median diameter of the spots and the coefficient of variation for Cy3 and Cy5 are all global raw quality criteria.
The lower table resumes the data quality after normalization and outlier detection. The percentage of spots of score 2 and the percentage of saturated and flagged spots give an idea of the quality of the data after Filtration. In the file of results, the spots of good quality have a score of 2, the spots close to saturation in Cy3 or Cy5 have a value of 3 and 4 respectively and those close to saturation in both channels have a score of 5. Flagged features have a score of zero. The median, standard deviation and median absolute deviation of the ratio are calculated for each slide and give the quality of the normalization step. Finally the approach used to normalize is specified as well as the number of normalized spots.

For your guidance, the last columns of the tables contain thresholds of quality. These data were estimated empirically and must be compared to your data with caution.

Table 14. Output of a MADSCAN analysis. **(A) Table summary about the condition in which the analysis was performed. (B) Table summary of the raw microarray data before any analysis. (c) Table summary of the data pre-processed and normalized.**

A

MADSCAN Analysis of VGCMD13	
Slide(s)	Imc00035.gpr - Imc00137.gpr - Imc00288.gpr - Imc00378.gpr -
Layout (MetaRow;MetaCol;Row;Col)	12:4:20:20
Analysis	A to Z analysis, i.e. from the data filtration to the data integration (matrix transposition)
Outlier detection method	zmad / both slide(s)

B

*** Quality Control of raw data ***

	Slide Imc00035	Slide Imc00137	Slide Imc00288	Slide Imc00378	Threshold
Flagged features	5 %	4 %	1 %	3 %	< 35 %
Blank					
Detected	90	104	154	186	0
Cy5 background level	392	224	184	195	-
Cy3 background level	532	329	470	265	-
Cy5:					
background level	357	200	188	186	< 500
Signal-to-noise	105	292	386	258	> 30
Coefficient variation	147	190	174	173	< 200
Cy3:					
background level	532	303	556	271	< 500
Signal-to-noise	72	164	138	199	> 30
Coefficient variation	140	197	146	161	< 200
Diameter:					
Median	91	84	97	101	-
SD	17	15	15	14	< 50

C

*** Summary of data processing ***

	Slide Imc00035	Slide Imc00137	Slide Imc00288	Slide Imc00378	Threshold
Features:					
Initial number of features	16868	16868	16868	16868	-
Validated	92 %	92 %	95 %	95 %	> 65
Saturated	0.1 %	0 %	0.1 %	0 %	< 3
Flagged	8 %	8 %	5 %	5 %	< 35
Statistically validated features	14357	14636	15255	15228	-
Ratio:					
Mean	0	-0.01	-0.02	0	< 0.06
Median	0	-0.01	-0.02	0	< 0.06
Standard Deviation	0.31	0.27	0.28	0.32	< 0.50
Median Absolute Deviation	0.16	0.16	0.16	0.16	< 0.30
Normalization mode	Proximal	Proximal	Proximal	Proximal	Pin

2.2 Detailed results

Details of the processing steps can be accessed via the "Detailed Results" button (Tab. 28- 17). The details are given slide by slide. The first table presents raw data quality sub-array by sub-array. This may be useful information in case of heterogeneous background level. In case of poor quality slides some values may be highlighted in red to warn the user (Tab 15B). The percentage of the different quality scores obtained after Filtration are detailed in table 29. The effect of normalization and scaling on the medians and the variances of sub-array ratios are shown in table 30.

Table 15A. Sub-array details for raw data quality

Block	% Flags	Buffer detected	Bg R Buffer	Bg G Buffer	Bg R	Bg G	Rs/n	Gs/n	CV R	CV G	Dia.	Dia SD.
1	12	0	321	275	349	327	237	21	127	124	100	12
2	7	0	307	278	329	286	307	27	139	137	98	13
3	4	0	295	264	304	264	250	188	163	159	95	12
4	3	0	266	236	282	241	307	288	165	162	96	13
5	2	0	303	263	324	261	356	449	139	139	98	11
6	2	0	295	232	310	243	355	374	140	143	97	11

Table 15B Sub-array details of raw data of poor quality**. The values highlighted in red indicate a slide of poor quality. The percentage of flagged spots in a slide should not be much higher than 35%.**

Block	% Flags	Bg R	Bg G	Rs/n	Gs/n	CV R	CV G	Dia.	Dia SD.
1	**63**	0	0	1000	1520	120	120	89	7
2	**64**	0	0	965	1113	136	136	95	9
3	**30**	0	0	2563	2689	132	132	95	9
4	**67**	0	0	1246	1627	118	108	87	6
5	**28**	0	0	2978	3121	130	128	95	7
6	**32**	0	0	2385	2867	134	143	95	11

Table 16. Percentages of the different quality scores obtained in the different sub-arrays. **Scores of 0 indicate poor quality spots, scores of 2 are for spots of good quality and scores of 3 to 5 indicate spots close to the saturation level.**

Block	% Poor spots score 0	% score 2	% Sat.Cy3 score 3	% Sat.Cy5 score 4	% Sat.both score 5
1	18	82	0	0	0
2	13	87	0	0.3	0
3	8	91	0	0.7	0
4	8	92	0	0	0
5	5	95	0	0	0

Table 17. Effects of normalization and scaling on the sub-array medians of ratios, standard deviation and median absolute deviation.

Block	RAW DATA			NORMALIZED DATA		
	Median	SD	MAD	Median	SD	MAD
1	-0.05	0.51	0.29	-0.06	0.46	0.24
2	0	0.52	0.25	-0.03	0.54	0.24
3	0.03	0.38	0.24	0.01	0.44	0.24
4	0.06	0.36	0.24	0.05	0.45	0.24
5	0.18	0.31	0.19	0.18	0.45	0.24
6	0.17	0.33	0.22	0.16	0.43	0.25
7	0.11	0.37	0.23	0.07	0.38	0.24
8	0.11	0.36	0.23	0.07	0.43	0.24
9	0.02	0.31	0.22	0.03	0.41	0.24
10	0.03	0.33	0.24	0.02	0.42	0.24

2.3 Download results

The results can be downloaded locally via the 

button. You can download a *.tgz directory that contains several files of results:

- Normalized_*filename*.txt
- Tendency_Variation_*filename*.txt
- Filtered_filename.txt
- Summary_Raw_Data.txt
- Summary_Filtration.txt
- Summary_Processed_Data.txt
- Summary_Outliers.txt
- ZMAD_Outliers.txt

(or Grubb_Outliers.txt)

- Consolidated_Matrix_*filename*.txt
- Control_*filename*.txt

All the files can be opened under tabular spreadsheets such as Excel ®. In most of the files the header is simplified compared to the *.gpr or text file you entered. Usually the date and the name of the file will be stored. The column names are also simplified. For example, the letters Rmed (Red) and Gmed (Green) are used in place of F635 median and F532 median labels respectively. The Rbmed and Gbmed are the background intensities. If you are performing a single step analysis, you will find at least one of those files for results and maybe one or two with a different title (see online help for more details). For example if you only perform normalization you will get the files Normalized_*filename*.txt, Tendency_Variation_*filename*.txt and Summary_Processed_Data.txt.

a) Summary_Raw_Data.txt

The *Summary_Raw_Data* text file contains the data of the first table presented on your main results page of the web browser (Tab. 31). In the text file, the date of the analysis is saved.

Table 18. Overall raw data quality.

Slide	Flagged_spots	DetectedBuffer	RbBuffer	GbBuffer	Rbmed	Gbmed	Rsn	Gsn	Dia	SDDia	Rcv	Gcv
Threshold	<35 %	0	< Cy5 background level	< Cy3 background level	<500	<500	>30	>30	spotting_parameter	50	< 200	< 200
Imc00035.gpr	5	81	389	504	357	532	105	72	91	17	147	140
Imc00137.gpr	4	89	230	328	200	303	292	164	84	15	190	197
Imc00288.gpr	1	68	185	457	188	556	386	138	97	15	174	146
Imc00378.gpr	3	79	196	267	186	271	258	199	101	14	173	161

b) RawQuality_*filename*.txt

This file contains the first table presented under "detailed results" (Tab. 32).

Table 19. Raw data quality per block before filtration. **Example of the first nine blocks**

Block	Flags Bad	Flags Buffer	Rb Buffer	Gb Buffer	Rb med	Gb med	Rsn	Gsn	Dia	SDDia	Rcv	Gcv
1	4	4	334	418	344	425	240	212	96	13	14814	14438
2	5	3	384	541	442	585	172	132	95	15	14113	13513
3	7	3	389	556	429	594	166	114	94	18	13996	13197
4	2	3	312	454	357	549	171	139	94	15	14308	13891
5	3	2	332	415	348	451	173	111	92	14	13977	13542
6	5	3	399	515	417	563	161	108	92	16	12865	12468
7	5	1	379	527	445	610	161	117	92	18	11569	11098
8	4	1	417	568	441	600	255	171	92	16	13047	12865
9	2	3	347	474	339	465	211	182	92	16	12843	13032

c) Summary_Filtration_filename.txt

This file contains the percentage of the different scores obtained after Filtration, also presented as table under the button "detailed results" (Tab. 33). The date and the name of the slide analyzed are also written.

Table 20. Scores per block. **Percentages of features with different scores in the different blocks. Example of the first six blocks**

Block	% Score 0	% Score 2	% Score 3	% Score 4	% Score 5
1	18	82	0	0	0
2	13	87	0	0.3	0
3	8	91	0	0.7	0
4	8	92	0	0	0
5	5	95	0	0	0
6	6	95	0	0	0

d) Normalized_filename.txt

The score corresponds to the score attributed after the filtration step. A and M are the geometric mean of the intensities and the normalized ratio in base 2 respectively.

Table 21. Example of the *Normalized_filename.txt* file.

Spot	Block	Name	ID	Rmed	Rbmed	Gmed	Gbmed	Rnorm	Gnorm	A	M	score
686	2	OAZ1	R044X10H01	65529	351	52931	281	57649	59682	15.84	-0.05	4
1019	3	ALPHACOLL-1	R044X08I13	65218	338	29485	222	59475	31872	15.41	0.9	4
1064	3	CLONE-1A	R044X11N01	65507	378	43297	276	57649	48477	15.69	0.25	4
3711	11	FZD7	R043X07J15	65364	342	48431	292	57449	54350	15.77	0.08	4
1	1	HPIP	R043X01C01	10762	299	10469	252	10016	10735	13.34	-0.1	2
2	1	UMPS	R043X01G01	2228	304	1813	235	1958	1557	10.77	0.33	2
3	1	PSMB3	R043X01K01	21726	309	16450	273	19962	17258	14.18	0.21	2

e) Tendency_Variation_*filename*.txt

The median ratio, the standard deviation and the absolute deviation of the median are calculated sub-array by sub-array before (RawRatio…) and after (NormRatio) normalization (Tab. 34):

Table 22. Example of the Tendency_Variation_*filename.txt* file.

Block	RawRatio	RawSD	RawMAD	NormRatio	NormSD	NormMAD
1	0.02	0.48	0.26	0.14	0.48	0.24
2	-0.03	0.48	0.26	0.15	0.5	0.26
3	0.08	0.37	0.22	0.26	0.39	0.2
4	0.18	0.31	0.19	0.31	0.33	0.18
5	0.15	0.35	0.23	0.28	0.38	0.25
6	0.1	0.37	0.23	0.26	0.38	0.22
7	0.03	0.33	0.24	0.2	0.34	0.2

f) Summary_Processed_Data_filename.txt

The Summary text file is the one you see on your results screen (Tab. 35). In the text file, the date of the analysis is also saved

Table 23. Example of the Summary_Processed_Data_*filename.txt* file.

Slide	Spots_score2	Satured spots	Flagged spots	Mean	SD	Median	MAD	Mode	Nb spots	Nb Norm	Pass_outlier_test
Threshold	>65 %	<3 %	<35 %	0.06	0.05	0.06	0.3	Pin	-	-	-
Imc00035	83	0.05	17	0	0.3	0	0.15	Proximal	16868	13976	12932
Imc00137	72	0.02	28	0	0.25	-0.01	0.15	Proximal	16868	12078	11359
Imc00288	92	0.07	8	0	0.26	-0.01	0.15	Proximal	16868	15496	14701
Imc00378	80	0.03	20	0	0.29	0	0.15	Proximal	16868	13479	12652

g) Outlier detection results

·**Zmad** Estimation of the outliers by a modified z-test (use of the MAD, Median Absolute Deviation, instead of the mean).

Grubbs' Grubbs' test for detecting outliers or ESD method (extreme studentized deviate) determines how far the outliers are from the other values. The Z-test is used to calculate the differences and a modified table of critical values allows estimating the outliers.

Grubbs' test can be applied starting from *n*=3 (one gene replicated 3 times within a slide or between slides). The available significance levels for Grubbs' test are (0.05; 0. 025; 0.01).

Table 24 .Results after a Zmad test.

Name	ID	Slide1	Slide2	Slide3	Slide4	Zscore1	Zscore2	Zscore3	Zscore4
GENE1004	R043X06O02	-0.16	-0.05	-0.01	OutMad1_0.02	0.36	0.91	1.47	NA
GENE1004	R043X06O02	-0.04	-0.05	-0.03	-0.18	1.09	0.89	1.14	0.68
GENE1004	R044X06O08	-0.18	-0.02	-0.14	-0.14	0.59	1.25	0.14	0.21
GENE1004	R044X06O08	-0.3	-0.13	-0.11	-0.17	2.12	0	0.21	0.52
GENE1005	R049X01H14	-0.2	-0.17	**OutMad1**_0.23	0.1	1.42	1.17	**3.5**	1.06
GENE1005	R049X01H14	-0.1	-0.07	-0.03	-0.09	0.57	0.34	0.02	0.5
GENE1005	R050X01H20	0.12	NA	-0.03	0.01	1.21	NA	0.02	0.32
GENE1005	R050X01H20	0.29	-0.21	-0.02	0.06	2.69	1.5	0.11	0.78
GENE1006	R044X10K15	-0.39	0.3	0.34	-0.06	0.68	0.79	0.88	0.03
GENE1006	R044X10K15	-0.28	0.37	0.3	-0.38	0.44	0.94	0.79	0.67
GENE1006	R043X10K21	-0.13	**OutMad1**_0.02	0.36	-0.37	0.13	**3.62**	0.91	0.64

The **z1** for the second value of GENE3 in the first slide is higher than the cut-off 3.5. Therefore it is labeled as an outlier.”OutMad1_” means that the value is an outlier within the slide (n=4). .”OutMad2_” indicates that the value is an outlier between the slides (n=8). Afterwards, it is up to you to keep or remove this value for further analysis.

Table 25. Table of results of the Grubb outlier test.

Name	ID	M 1	M 2	zmax	zmin
GENE5	R044X03C06	NA	NA	-	-
GENE5	R044X03C06	NA	NA	-	-
GENE5	R043X03C12	NA	-0.29	-	0.69
GENE5	R043X03C12	-0.21	0.31	1.14	-
GENE6	R044X01G04	**-0.04**	-0.05	-	-
GENE6	R044X01G04	-0.19	-0.09	-	-
GENE6	R043X01G10	-0.08	**OutGrubb1 _ -0.50**	-	**2.33**
GENE6	R043X01G10	**-0.04**	-0.06	**0.64**	-
GENE7	R043X07N05	0.01	0.08	-	-
GENE7	R043X07N05	0.26	-0.01	-	-
GENE7	R044X07N11	0.60	0.01	1.99	-
GENE7	R044X07N11	-0.29	0.06	-	1.49

M maximum for the 8 values and for which zmax is calculated

M minimum for the 8 values and for which zmin is calculated

In this example, **n=8**, **z-table=2.22** (Annex 1). The zmin for GENE6 in the second slide is above the z-table of 2.22. Therefore it is labeled as an outlier. Like in the previous test you have the choice to keep or remove this value.

In most cases the Zmad test is more stringent and identifies more genes as outlier than the Grubb's and Dixon tests.

h) Data integration or consolidated gene expression matrix

Figure 19 presents the form that will transpose your dataset and summarize it with one line per gene. The resulting file is the Consolidated_Matrix_*yourfilename*.txt. It contains the names of the genes and the different values (Tab. 36) : the within slide median of ratio in $\log_2$ (M), the within slide coefficient of variation (CV), the between slide median ratio, CV and the same data for the geometric mean of the intensities (A) .

Table 26 Example of an output file from the data integration process.

Name	Slide1 M1	Slide1 M2	Slide 1M3	M4	M5	M6	M7	M8	Median _M1	Median _M2	cv1	cv2	Median	cv
54TM_____ ___-1A	0.16	0.28	0.28	NA	0.25	0.34	0.25	0.33	0.28	0.29	28.87	16.84	0.28	22.22
A:04413_____ __-1A	0.03	0.12	0.14	0.36	0.23	0.12	0.1	0.3	0.13	0.17	86.21	50.29	0.13	63.63
A1BG_____ ___-1A	-0.07	NA	-0.08	0.08	0.06	0.05	-0.04	-0.06	-0.07	0.00	384.12	2452.21	-0.04	803.87
A2BP1_____ ___-1A	0.04	0.06	0.09	NA	NA	NA	0.15	0.44	0.06	0.29	39.74	69.51	0.09	105.19
A2M_____ ___-1A	0.89	0.8	0.52	NA	0.7	0.88	1.15	0.93	0.8	0.90	26.19	20.24	0.88	23.44
AAMP_____ ____-1A	-0.59	NA	-0.65	-0.55	-0.67	-0.38	-0.41	-0.75	-0.59	-0.54	8.44	33.52	-0.59	23.83

2.4 Graphs

Three types of graphs are displayed by clicking on 

.

a) The Box and whisker plot

The Box and whisker plot is a descriptive visualization of the data dispersion in the different slides analyzed. This representation is particularly interesting to visualize the effect of the "scaling step" (Fig. 20).

The "box" in a box plot shows the median ratio as a line and the first (25th percentile) and third quartile (75th percentile) of the ratio distribution as the lower and upper parts of the box. The median is the ratio at the 50% percentile: half of all genes get a ratio higher than the median, and 50% get a ratio lower. It is the middle point in the distribution of ratios. The 25th percentile is the point at which 25% of the genes ratio are lower (and 75% score higher) than the median. The 75th percentile is the point at which 75% of the genes ratio score lower (and 25% score higher) than the median. Thus, the area in the "box" represents the middle 50% of the genes.
The "whiskers" shown above and below the boxes technically represent the largest and smallest observed ratios that are less than 1.5 box lengths from the end of the box. In practice, these ratios are about the lowest and highest values one is likely to observe.
The open circles "o" are ones that are, respectively "very rare" and "exceedingly rare." Such scores may represent very differential genes or outliers.

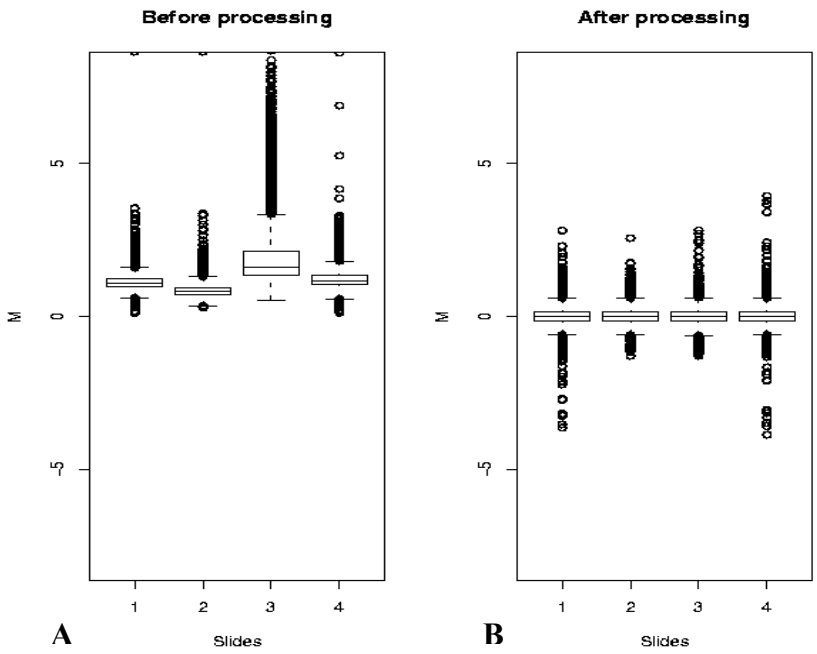


Figure 60. Box plots before (A) and after scaling (B).

The M vs. A scatter plot shows the normalized ratios ($M=\log_2 R-\log_2 G$) versus the geometric mean of the intensities ($A=1/2(\log_2 R+\log_2 G)$) sub-array by sub-array (Fig. 21).

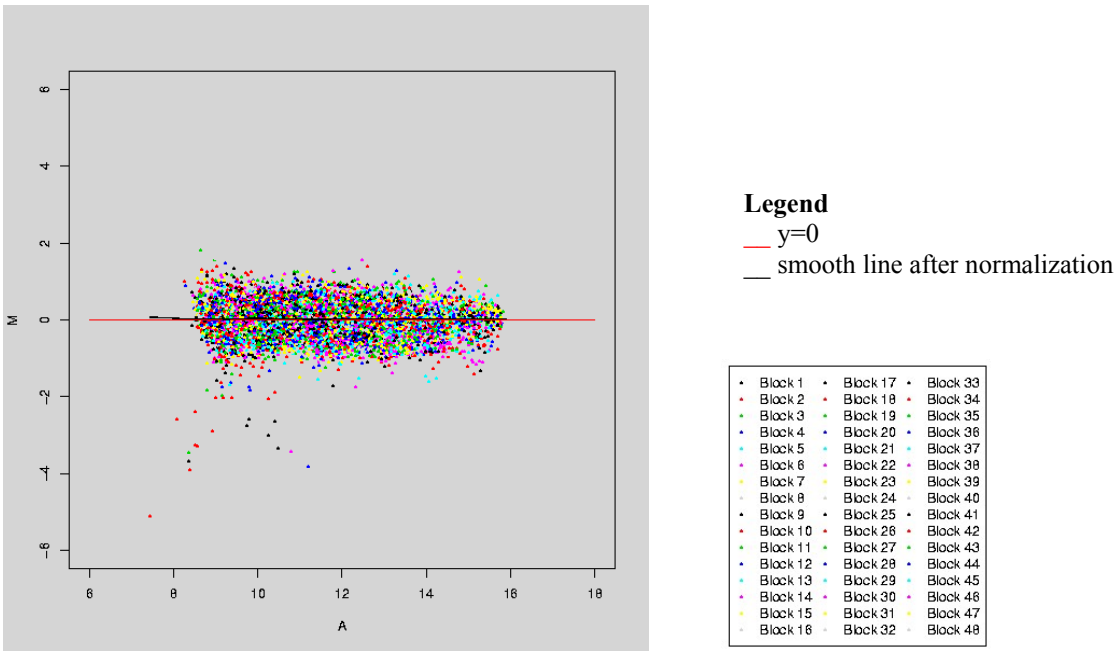


Figure 61. M vs. A scatter plot in log2 of one of the normalized AMLSpA024-L1 slide.

The density plot presents the dispersion of the ratios (Fig. 22).

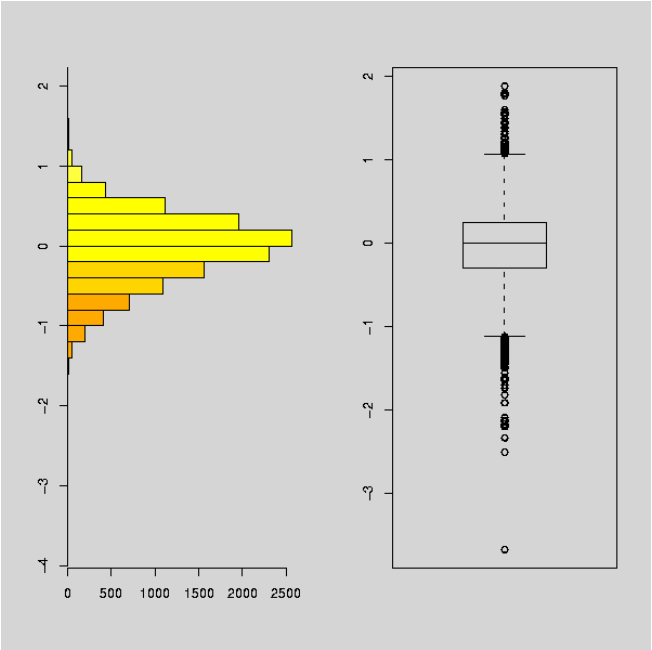


Figure 62. Density plot of the normalized ratios (in log2) and the corresponding box plot for slide AMLSpA024-L2.

III. Estimation of differentially expressed genes: Limma analysis

1. Data input

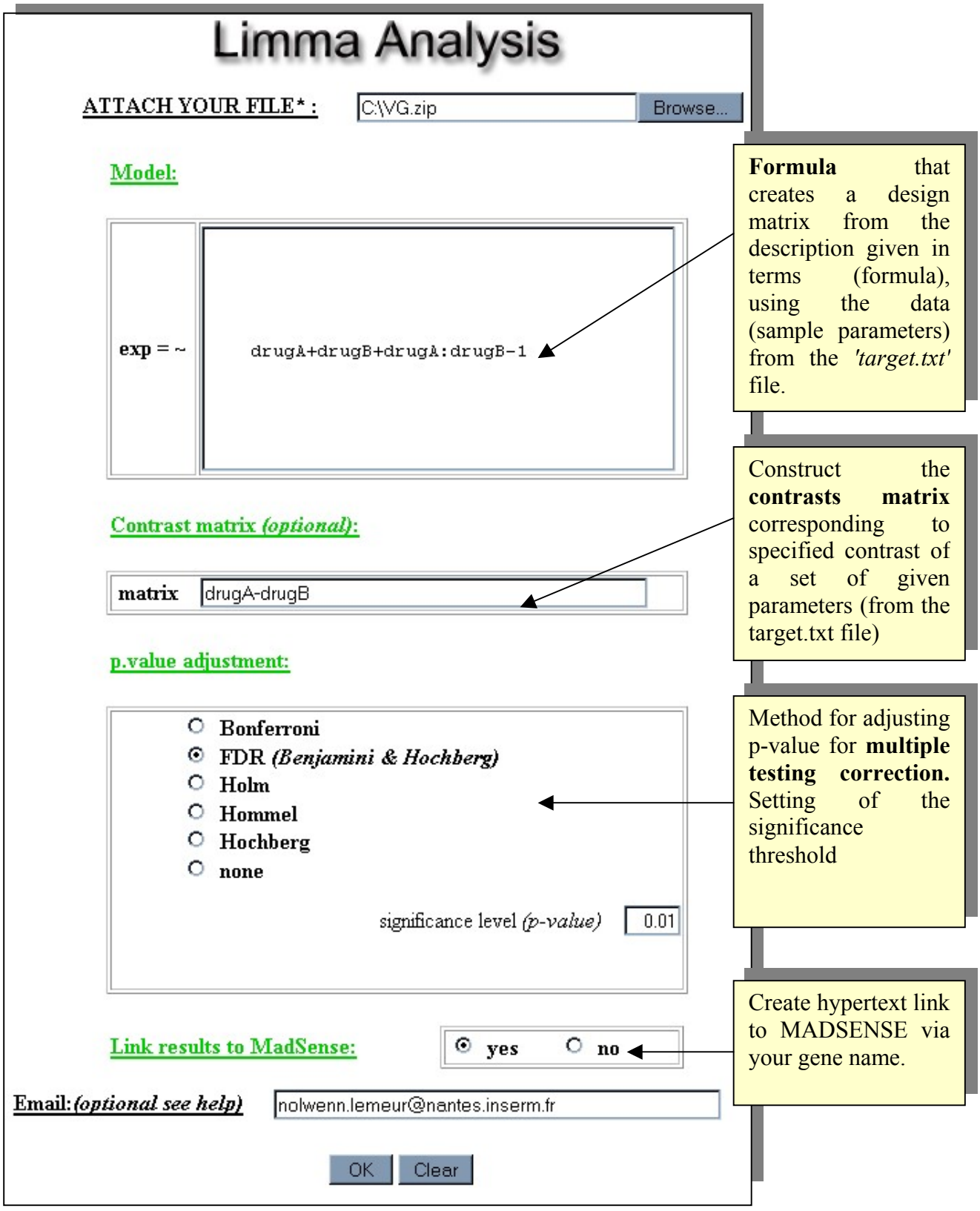


Figure 63. Limma analysis input form.

1.1 Files

Two files are needed: a gene expression data matrix and a file with sample (targets) descriptors. Both files have to compressed in an unique *.zip file to be uploaded by our tool.

Gene expression data matrix. You must enter one text-tabulated file containing gene expression data for your different sample (see example). Be careful, the first source of error is the gene name. Avoid special character such as '#','/','?' in your gene name. NA is accepted.

NAME	ID	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
GENE1	ID1	0.10	0.20	0.10	0.09	1.2	1.11
GENE2	ID2	0.92	1.30	1.19	0.98	0.91	0.91
GENE3	ID3	-0.42	-0.35	0.71	0.89	0.91	0.91
GENE4	ID4	1.20	1.51	1.32	1.25	1.33	1.33
GENE5	ID5	0.83	0.69	0.82	0.79	-0.63	-0.73

target.txt file. A tabulated text file is required to describe your sample parameters. The file must be named 'target.txt' and the sample descriptors must be discrete values, e.g. 0 or 1.

Sample	drugA	drugB
Sample1	1	0
Sample2	1	0
Sample3	0	1
Sample4	0	1
Sample5	1	1
Sample6	1	1

1.2 Model

Formula that creates a design matrix from the description given in terms (formula), using the data (sample parameters) from the *'target.txt'* file.

Ex: *expr*= ~ drugA + drugB + drugA:drugB – 1 (in short *expr*= ~ drugA*drugB – 1)

1.3 Contrast Matrix:

Construct the contrasts matrix corresponding to specified contrast of a set of given parameters (from the target.txt file)

Ex: contrast matrix = pathology – drugA

1.4 p-value adjustment:

Bonferroni The p-values are multiplied by the number of comparisons

Holm Strong control of the family wise error rate, valid under arbitrary assumptions and less conservative than Bonferroni approach.

Hocheberg or **Hommel** Hochberg's and Hommel's methods are valid when the hypothesis tests are independent or when they are non-negatively associated (Sarkar, 1998; Sarkar and Chang, 1997).

Hommel's method is more powerful than Hochberg's, but the difference is usually small and the Hochberg p-values are faster to compute.

FDR *(Benjamini & Hochberg)* The method of Benjamini and Hochberg (1995) controls the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family wise error rate.

none Pass-through option

2. Results

2.1 Files

*a) limma*_padjust-*result*_factor.txt

As an example, we tested the effect of drug A on gene expression (drugA *vs.* Reference). The adjustment used to correct the *p-value* is a FDR. The file of results, *limma_fdr-result_DrugA.txt,* gives you the mean effect, the t-statistics, the adjusted *p.value* and the Bayes log odds score for every genes.

	Model	drugA	drugA	drugA	drugA	drugA	drugA				
	Design	1	1	0	0	1	1				
NAME	ID	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	M	t	P.Value	B
GENE4	ID4	1.2	1.51	1.32	1.25	1.33	1.33	1.36	17.05	0.001	137.37
GENE2	ID2	0.92	1.3	1.19	0.98	0.91	0.91	1.11	13.97	0.002	89.61
GENE5	ID5	0.83	0.69	0.82	0.79	-0.63	-0.73	0.76	9.57	0.002	37.85
GENE3	ID3	-0.42	-0.35	0.71	0.89	0.91	0.91	-0.39	-4.85	0.003	3.90
GENE1	ID1	0.1	0.2	0.1	0.09	1.2	1.11	0.15	1.89	0.059	-6.05

mean

- M

Difference in gene expression ratio between sample treated and non treated

- Moderated t-statistics

t is the emprirical Bayes moderated t-statistics. It is the ratio of the M-value to its standard error. This has the same interpretation as an ordinary t-statistic except that the standard errors have been moderated across genes, effectively borrowing information from the ensemble of genes to aid with inference about each individual gene.

- p-values

The p-value (p-value) is obtained from the moderated t-statistic, usually after some form of adjustment for multiple testing. The most popular form of adjustment is "fdr" which is Benjamini and Hochberg's method to control the false discovery rate. The meaning of the adjusted p-value is as follows. If you select all genes with p-value below a given value, say 0.05, as differentially expression, then the expected proportion of false discoveries in the selected group should be less than that value, in this case less than 5%.

- Bayes log odds

B (B-statistics or lods) is the empirical Bayes log-odds of differential expression. B-statistic probabilities depend on various sorts of mathematical assumptions which are never exactly true for microarray data. The B-statistics also depend on a prior guess for the proportion of differentially expressed genes. Therefore they are intended to be taken as a guide rather than as a strict measure of the probability of differential expression.

For example, suppose that B=1.75. The odds of differential expression is exp(1.5)=5.75, i.e, about five and three quater to one. The probability that the gene is differentially expressed is 5.75/(1+5.75)=0.85, i.e., the probability is about 85% that this gene is differentially expressed. A B-statistic of zero corresponds to a 50-50 chance that the gene is differentially expressed. The B-statistic is automatically adjusted for multiple testing by assuming that 1% of the genes, or some other percentage specified by the user, are expected to be differentially expressed.

If there are no missing values in your data, then the moderated t and B statistics will rank the genes in exactly the same order. Even you do have spot weights or missing data, the p-values and B-statistics will usually provide a very similar ranking of the genes.

b) Classification test

In a complex experiment with many contrasts, it may be desirable to select genes firstly on the basis of their moderated F-statistics, and subsequently to decide which of the individual contrasts are significant for those genes. This cuts down on the number of tests which need to be conducted and therefore on the amount of adjustment for multiple testing (Smyth, 2004).

As an example, we tested the difference (contrast) of effects between drug A and B on gene expression (drugA-drugB). The file of results, limma_*padjust*-result_c FcontrastMatrix.txt, gives you the result of the F-statistics.

		DrugA	DrugA	DrugB	DrugB	DrugA+B	DrugA+B	
NAME	ID	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	DrugA-drugB
GENE1	ID1	0.83	0.69	0.82	0.79	-0.63	-0.73	0
GENE2	ID2	1.2	1.51	1.32	1.25	1.33	1.33	0
GENE3	ID3	0.92	1.3	1.19	0.98	0.91	0.91	-1
GENE4	ID4	0.1	0.2	0.1	0.09	1.2	1.11	0
GENE5	ID5	-0.42	-0.35	0.71	0.89	0.91	0.91	0

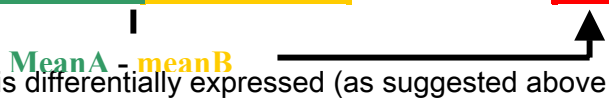
Here the gene 3 is possibly (not statistically) differentially expressed, down-regulated by the drug A.

c) limma_fdr-result_contrastMatrix.txt

We tested the difference of effects between drug A and drug B on gene expression (drugA-drugB). The adjustment used to correct the *p-value* is a FDR. The file of results, limma_*padjust*-result_contrast_drugA-drugB.txt, gives you the mean effect, the t-statistics, the adjusted *p.value* and the Bayes log odds score for every genes.

	Model	drugA - drugB	drugA - drugB	drugA - drugB							
	Design	1	-1	0							
NAME	ID	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	M	t	P.Value	B
GENE3	ID3	-0.42	-0.35	0.71	0.89	0.91	0.91	-1.18	-10.55	0.001	48.04

GENE4	ID4	1.2	1.51	1.32	1.25	1.33	1.33	0.07	0.62	0.820	-7.15
GENE1	ID1	0.1	0.2	0.1	0.09	1.2	1.11	0.06	0.49	0.820	-7.23
GENE5	ID5	0.83	0.69	0.82	0.79	-0.63	-0.73	-0.04	-0.4	0.820	-7.26
GENE2	ID2	0.92	1.3	1.19	0.98	0.91	0.91	0.02	0.22	0.820	-7.32



Here, only Gene 3 is differentially expressed (as suggested above by the classify test) and statistically significant at a p.value= 0.001.

IV. Frequently asked questions

Why do I get R error?
1. I uploaded a *.rar file instead of a *.zip file
2. I have incomplete name or ID columns in the gene list.
3. I try to integrate data (last MADSCAN step) with different number of replicates per gene.

Why does the program fail to upload my *.zip file?
Verify that you are working with IE 6.0 or higher

Why are some files missing in the results folder?
Your *.gpr and *.txt must begin with a letter and must not contain space.

Most frequent errors in GAL files:
-incomplete name or ID columns: the resulting table contains spaces and the reading fails. NAs are accepted if data are missing.
-use of special character such as "#" "." "@"

What do I do when I get the web browser message "Connection TIME OUT"?
This error message doesn't mean the MADSCAN analysis failed. It is only a network problem. If you have specified your email address in the submission form you will receive a message as soon as the analysis will be achieved. This message contains your analysis ID and you will be able to retrieve your results by going back to MADSCAN and entering your analysis ID into the "Retrieve result" page.

Can I analyze microarray with only one block ?
Yes, if you have array with only one block such (as Agilent oligo arrays) you can still analyze them with MADSCAN. Just define the layout as such: metaRow: 1, metaCol: 1, Row: *nspot*, Col: 1

CITING MADSCAN

Le Meur, N., Lamirault, G., Bihouée, A., Steenman1, M., Bédrine-Ferran, H., Teusan, R., Ramstein, G., J. Léger, J.J. (2004) **A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values. Importance of replication.** *Nucleic Acids Research,* **32**(18): *5349-5358*

Related publications

Steenman, M., Lamirault, G., Le Meur, N., Le Cunff, M., Escande, D., Léger, J.J. (2004) **Distinct molecular portraits of human failing hearts identified by dedicated cDNA microarrays.** *European Journal of Heart Failure (In press). http://dx.doi.org/10.1016/j.ejheart.2004.05.008*

Bedrine-Ferran, H., Le Meur, N., Gicquel I, Le Cunff, M., Soriano, N., Guisle, I., Mottier, S., Monnier, A., Teusan, R., Fergelot, P., Le Gall, J.Y., Leger, J., Mosser, J. (2004). **Transcriptome variations in human Caco-2 cells : a model for enterocyte differentiation and its link to iron absorption.** *Genomics*, **83**(5):747-950

Le Bouter, S., El Harchi, A., Marionneau, C. *et al.* (2004). **Long-term amiodarone administration remodels expression of ion channel transcripts in the mouse heart.** Circulation, **110**: 3028-3035.

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series* B, **57**, 289–300.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat.Genet., **29**, 365-371.

Burke, S. (2001) Missing Values, Outliers, Robust Statistics & Non-parametric Methods. Statistics and data analysis. Statistics and data analysis.LC.GC Europe online Supplement., **59**, 19-24.

Chen, Y., Dougherty, E., and Bittner, M. (2001) Ratio-based decisions and the quantitative analysis of cDNA microarray images. J.Biom.Opt., **2**, 364-374.

Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. Nat.Genet., **32** Suppl, 490-495.

Draghici, S. (2003). Data Analysis Tools for DNA Microarrays. First Edition; Chapman & HallBoca Raton, Florida.

Dudoit, S., Yang, Y. H., Callow, M., and Speed, T. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.Technical report #578,

Dudoit, S., Yang, Y. H., Callow, M., and Speed, T. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica sinica., **12**, 111-139.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. J.Comput.Graph.Statist., **5**, 299-314.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383–386.

Kerr, M. K. (2003) Experimental design to make the most of microarray studies. Methods Mol.Biol., **224**, 137-147.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. J.Comput.Biol., **7**, 819-837.

Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000) Importance of replication in

microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc.Natl.Acad.Sci.U.S.A., **97**, 9834-9839.

Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W., and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. J.Biol.Chem., **276**, 19937-19944.

Müller, J. W. (2000) Possible Advantages of a Robust Evaluation of Comparisons. Journel of Research of the National Institute of Standards and Technology., **105**, 551-555.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics., **18**, 546-554.

Pan, W., Lin, J., and Le, C. T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol., **3**, RESEARCH0022

Quackenbush, J. (2002) Microarray data normalization and transformation. Nat.Genet., **32** Suppl, 496-501.

Smyth, G. K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology., **3**(1).

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc.Natl.Acad.Sci.U.S.A., **99**, 6567-6572.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res., **29**, 2549-2557.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc.Natl.Acad.Sci.U.S.A., **98**, 5116-5121.

Wettenhall, J. M. and Smyth, G. K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. Bioinformatics.,

Yang, Y. H., Buckley, M. J., and Speed, T. P. (2001a) Analysis of cDNA microarray images. Brief.Bioinform., **2**, 341-349.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., **30**, e15

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. (2001b) Normalization for cDNA Microarray Data.1-11.

Yang, Y. H. and Speed, T. (2002) Design issues for cDNA microarray experiments. Nat.Rev.Genet., **3**, 579-588.

Illustration Table

ANNEXES

Annex 1: GLOSSARY

A
The letter A stands for the geometrical mean of the intensities (Cy3 and Cy5) of a feature, i.e. *add* as $A=(log_2R+log_2G)/2$

Array design
(Synonym: layout)
The layout or conceptual description of arrays that can be implemented as one or more physical arrays. The array design specification consists of the description of the common features of the array as a whole, and the description of each array design element (e.g., each spot). MIAME distinguishes between three levels of array design elements: feature (the location on the array), reporter (the nucleotide sequence present in a particular location on the array), and composite sequence (a set of reporters used collectively to measure an expression of a particular gene)

Arrayer
(Synonym: spotter)
Robot to print elements on the surface of an array

Batch
Collection of microarrays with the same probe layout.

Biases
Random (experimental) and systematic variation such as differences in the labeling (i.e., dye biases), sample preparation, the hybridization, scanner settings, auto fluorescence.

Biological Element
A Biological Element represents a coding fragment of a gene.

Block
(Synonym: grid; print-tip-group; sub-array)
Identified by its Meta-Column and Meta-Row coordinates

CGI
Abbreviation of ***C****ommon* ***G****ateway* ***I****nterface*, a specification for transferring information between a World Wide Web server and a CGI program. CGI programs are the most common way for Web servers to interact dynamically with users. For example, many HTML pages that contain forms use a CGI program to process the form's data once it is submitted. The program could be written in any programming language, including C, Perl, Java, or Visual Basic.

Coefficient of variation
(Abbreviation: CV) The coefficient of variation is an attribute of a distribution: its standard deviation divided by its mean.

Cy5(Cy3)
Fluorophore Cyanine 5 and 3, which fluoresce at a wavelength of 635mn and 533mn respectively.

Dye
Fluorescent label such as the cyanine 5 and 3.

Dye swap
(Synonym: dye flip)
To label every sample with both dye.

Even design
Every sample is labeled with both dyes and is equally often hybridized (ex. Loop design).

Experimental layout
The experimental layout is how samples are paired onto arrays and compare to each other. The layout affect the ability to discern and pull apart different sources of variation that could otherwise lead to biased results.

Feature
(Synonym: spot; element)
A feature refers to a specific instance of a reporter positioned upon an array

Filtration
(Synonym: Physical validation)
Process that aims to flag flawed spots and extract information from borderline features (close to the background level or saturation level),
Filtering
Process of removing flawed spots based on a decision of quality criteria.

Flag
Indicates a flawed and unreliable feature.

GAL
Gene Array List

Gbmed
Median background of the pixel intensities in Cy3 with *G* standing for green, *b* for background and *med* for median.

Gmed
Median of the pixel intensities in Cy3 with *G* standing for green and *med* for median.

Grid
(synonym: Block ;sub-array)

Gs/n
Signal to noise coefficient in Cy3

ID
(synonym: Reporter Identifier)
Since identifiers are used for cross-referencing, they must be unique to ensure unambiguous access to reporters. However, in a tabular presentation, features may anyway have the same reporter identifier if they are replicates.

Layout
(Synonym: Array design)
represents a general positioning for printing elements on an array.

M
The ratio of the intensities in log base 2, i.e. *minus* as $M= log_2R-log_2G$

MA plot
(Synonym: RI plot)
The scatter plot of the Ratio of the intensities versus the mean of the Intensities.

Name
(synonym: Reporter Name)
The information present in this field is meant for being displayed and should therefore be human-readable. Official gene name should be used for creating a Reporter name (http://www.gene.ucl.ac.uk/nomenclature/).
In case of complex arrays containing different sequences of a same gene, Reporter name could be used to identify the order of their position in the gene.

Noise
Non-specific signal.

Normalization
Process to minimize experimental systematic biases so that the observed variation arises from biological differences rather than from defects in the microarray technology.

Outliers
Inconsistent measures of replicated data points. The outlier detection allows evaluating consistency of replicates within an array and between replicated arrays.

PCR
Polymerase Chain Reaction is an *in vitro* technique for the amplification of a region of DNA.

Perl
PERL or Practical extraction and report language is a high-level programming language with an eclectic heritage written by Larry Wall and a cast of thousands. It derives from the ubiquitous C programming language and to a lesser extent from sed, awk, the Unix shell, and at least a dozen other tools and languages. Perl's process, file, and text manipulation facilities make it particularly well-suited for tasks involving quick prototyping, system utilities, software tools, system management tasks, database access, graphical programming, networking, and world wide web programming (http://www.perldoc.com/).

PHP
(Recursive acronym: Hypertext Pre-processor or Personal Home Page)
PHP is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML.

Physical validation
(Synonym: filtration)
Process that aims to flag flawed spots and extract information from borderline features (close to the background level or saturation level),

Probe
(Synonym: reporter)
A probe represents the content of a feature. A reporter may be related to a biological entity, but not necessarily, for instance when a

reporter is a negative control corresponding to printing buffer. When a reporter relates to biological elements, it is associated to a bio-sequence.

Print-tip
(Synonym: pin)

Print-tip-group
(Synonym: block; grid; sub-array)

R
Language and environment for statistical computing and graphics. It is a GNU project that is similar to the S language and environment, which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.
R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form (http://www.r-project.org). It compiles and runs out of the box on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). It also compiles and runs on Windows 9x/NT/2000 and MacOS.

Rbmed
Median background of the pixel intensities in Cy5 with *R* standing for red, *b* for background and *med* for median.

Reference design
samples of interest are hybridized to a reference

Replicate
To replicate means duplicate, repeat, or perform the same task more than once. Replications allow the experimenter to obtain estimates of the experimental errors.

Reporter
(Synonym: Probe)
A reporter represents the content of a feature. A reporter may be related to a biological entity, but not necessarily, for instance when a reporter is a negative control corresponding to printing buffer. When reporter relates to biological elements, it is associated to a bio-sequence.

Reporter Identifier
(synonym : ID)
Since identifiers are used for cross-referencing, they must be unique to ensure unambiguous access to reporters. However, in a tabular presentation, features may have the same reporter identifier if they are replicates.

Reporter name(synonym: Name)
The information present in this field is meant for being displayed and should therefore be human-readable. Official gene name should be used for creating a Reporter name. In case of complex array containing different sequence of a same gene, Reporter name could be devised to identify the order of their position in the gene.

Rmed
Median of the pixel intensities in Cy5 with *R* standing for red and *med* for median.

Rs/n
Signal to noise coefficient in Cy5

Scaling
To bring gene expression ratios of different slides at the same median absolute deviation.

Slide
Glass slide not different from the one used under a microscope.

Spot
(Synonym: feature)
Element (clone, oligonucleotide, buffer, blank) printed on a slide

Spotter
(Synonym: arrayer)

Sub-array
(Synonym: block; grid; print-tip-group)

Target
The entity that is labeled and hybridized to a particular chip

To spot
To precisely apply tiny droplets containing functional DNA to glass slides by the means of a robot.

To hybridize
To allow the labeled probes to bind to complementary DNA strands on the slides.

To scan
The slides are put into a scanning microscope that can quantify the brightness of each fluorescent dot; brightness reveals how much of a specific spotted DNA fragment is present within targets, an indicator of how actively it is transcribed

Annex 2

Table of significance levels for the Grubb's test

n	Significance level		
	5%	2.5%	1%
3	1.15	1.15	1.15
4	1.46	1.48	1.49
5	1.67	1.71	1.75
6	1.82	1.89	1.94
7	1.94	2.02	2.10
8	2.03	2.13	2.22
9	2.11	2.21	2.32
10	2.18	2.29	2.41
11	2.23	2.36	2.48
12	2.29	2.41	2.55
13	2.33	2.46	2.61
14	2.37	2.51	2.66
15	2.41	2.55	2.71
16	2.44	2.59	2.75
17	2.47	2.62	2.79
18	2.50	2.65	2.82
19	2.53	2.68	2.85
20	2.56	2.71	2.88
21	2.58	2.73	2.91
22	2.60	2.76	2.94
23	2.62	2.78	2.96
24	2.64	2.80	2.99
25	2.66	2.82	3.01