JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ALAIN BACCINI
PHILIPPE BESSE
SÉBASTIEN DÉJEAN
PASCAL G. P. MARTIN
CHRISTÈLE ROBERT-GRANIÉ
MAGALI SAN CRISTOBAL

Stratégies pour l'analyse statistique de données transcriptomiques

Journal de la société française de statistique, tome 146, n° 1-2 (2005), p. 5-44

<http://www.numdam.org/item?id=JSFS_2005__146_1-2_5_0>

STRATÉGIES POUR L'ANALYSE STATISTIQUE DE DONNÉES TRANSCRIPTOMIQUES

Alain BACCINI *, Philippe BESSE*, Sébastien DÉJEAN*,
Pascal G.P. MARTIN **, Christèle ROBERT-GRANIÉ ***,
Magali SAN CRISTOBAL ****

RÉSUMÉ

Afin d'illustrer la diversité des stratégies applicables à l'analyse de données transcriptomiques, nous mettons d'abord en œuvre des méthodes issues de la statistique exploratoire (ACP, positionnement multidimensionnel, classification), de la modélisation (analyse de variance, modèles mixtes, tests) ou de l'apprentissage (forêts aléatoires), sur un jeu de données provenant d'une étude de nutrition chez la souris. Dans un second temps, les résultats obtenus sont mis en relation avec des paramètres cliniques mesurés sur les mêmes animaux, en utilisant cette fois l'analyse canonique. La plupart des méthodes fournissent des résultats biologiquement pertinents sur ces données.

De cette expérience, nous tirons quelques enseignements élémentaires : il n'y a pas, *a priori*, de meilleure approche ; il faut trouver la « bonne » stratégie associant exploration et modélisation, adaptée tant aux données qu'à l'objectif recherché. Dans cette optique, une collaboration étroite entre statisticien et biologiste est indispensable.

Mots clés : biopuces, analyse en composantes principales, classification, modèle linéaire, forêts aléatoires, analyse canonique.

ABSTRACT

In order to illustrate the variety of strategies applicable to transcriptomic data analysis, we first implement methods of exploratory statistics (PCA, multidimensional scaling, clustering), modelling (ANOVA, mixed models, tests) or learning (random forests), on a dataset coming from a nutrition study for mice. In a second stage, relationships between the previous results and clinical measures are studied through canonical correlation analysis. Most of the methods provide biological relevant results on these data.

* Laboratoire de Statistique et Probabilités - UMR CNRS 5583, Université Paul Sabatier - 31062 Toulouse Cedex 9.
{alain.baccini,philippe.besse,sebastien.dejean}@math.ups-tlse.fr

** Laboratoire de Pharmacologie et Toxicologie - UR 66 - INRA, 180, chemin de Tournefeuille - B.P. 3 - 31931 Toulouse Cedex.

*** Station d'Amélioration Génétique des Animaux - INRA

**** Laboratoire de Génétique Cellulaire - INRA, Auzeville B.P. 27 - 31326 Castanet-Tolosan Cedex.
{pmartin,robert,msc}@toulouse.inra.fr

From this experience we conclude that there is not one best approach; we have to find the "good" strategy combining exploration and modelling to fit the data as well as to achieve the biological purpose. From this point of view, a strong collaboration between statistician and biologist is essential.

Keywords : microarray, principal component analysis, clustering, linear model, random forests, canonical correlation analysis.

1. Introduction

Les données transcriptomiques, ou *données d'expression des gènes*, sont issues de technologies variées (*Polymerase Chain Reaction* – PCR – en temps réel, biopuces constituées d'acide désoxyribonucléique – ADN – complémentaire ou d'oligonucléotides déposés sur membrane de nylon ou sur lame de verre), chacune nécessitant des prétraitements spécifiques : analyse d'image, quantification et normalisation. Ces derniers ne sont pas abordés dans cet article, même si, déterminants dans beaucoup de résultats, ils doivent être envisagés avec prudence et évalués en termes d'impact sur les conclusions retenues.

Les problématiques rencontrées, les conditions expérimentales, ainsi que les objectifs visés varient considérablement : présélection de gènes, prise en compte d'une cinétique, importance de covariables, présence de variables phénotypiques... Néanmoins, la spécificité majeure de ces données, celle qui remet le plus en cause le bon usage et le savoir faire du statisticien, est la très haute dimensionnalité du nombre de gènes par rapport au nombre comparativement très restreint d'échantillons biologiques sur lesquels leur expression est mesurée. De façon formelle, le problème se pose comme l'observation d'une variable, l'*expression* (ou quantité d'acide ribonucléique messager – ARNm – produite), dans des situations expérimentales croisant au moins deux facteurs : le gène et le type d'échantillon biologique (tissus sain ou pathologique, cellule sauvage ou modifiée...). Le premier facteur peut présenter de quelques centaines à quelques dizaines de milliers de niveaux, tandis que le second, pour des raisons évidentes de coût, ne présente en général que quelques dizaines de niveaux au maximum. L'objet de l'analyse statistique est alors d'extraire une information pertinente concernant l'effet de différents facteurs sur l'état fonctionnel de la cellule.

L'objectif de cet article est de montrer, sur un exemple réel, quels sont, dans la vaste panoplie disponible, les outils qui se montrent les plus efficaces pour apporter des solutions pertinentes aux questions du biologiste. Nous verrons que chaque outil et, plus précisément, chaque option (pondération, distance, transformation...) apporte un point de vue différent sur les données. Le défi consiste alors à conduire ces choix en conscience, compte tenu de leurs limites et de l'interprétation biologique qui en est faite, afin de déterminer la technique, ou l'optique, qui offre la représentation la plus éclairante au biologiste. Ainsi, nous apporterons, pour les différentes techniques, quelques commentaires sur l'interprétation biologique des résultats; ceux-ci feront l'objet d'une publication ultérieure en cours de préparation (Martin *et al.*, 2005b).

Dans le paragraphe suivant, nous présentons le jeu de données à l'origine de cet article ainsi que la problématique biologique. Le troisième paragraphe est consacré à la démarche exploratoire se focalisant sur l'analyse d'un tableau de données pour lequel on cherche des représentations graphiques pertinentes. Dans le paragraphe 4, nous présentons diverses méthodes de modélisation qui ont permis de confirmer et de compléter la connaissance sur l'expression des gènes acquise avec la démarche exploratoire. L'un des objectifs recherchés, à travers l'expérimentation biologique mise en œuvre, étant de mettre en relation l'expression de certains gènes avec les quantités d'acides gras hépatiques, le paragraphe 5 précise comment l'analyse canonique permet de répondre à cette question. Enfin, conclusions et perspectives sont présentées dans le dernier paragraphe.

2. Les données

Les données ont été fournies par l'équipe de pharmacologie moléculaire de l'Unité de Recherche 66 de l'INRA de Toulouse. Elles proviennent d'une étude de nutrition chez la souris. Pour 40 souris, nous disposons :

- des données d'expression de 120 gènes sélectionnés parmi les 30000 de la souris comme étant susceptibles d'être régulés par les conditions nutritionnelles de cette étude; elles ont été recueillies sur membrane nylon avec marquage radioactif;
- des concentrations de 21 acides gras hépatiques obtenues par chromatographie en phase gazeuse au laboratoire de Biochimie ENSAR-INRA de Rennes (pour chaque souris, la somme des 21 mesures est égale à 100).

Les 120 gènes analysés constituent une puce dédiée (*INRArray 01.2*). Ces gènes ont été sélectionnés, d'une part, sur la base d'une étude bibliographique, d'autre part, à partir d'une étude préliminaire réalisée sur 10000 gènes dont l'expression a été mesurée par *microarray* dans des conditions susceptibles de mettre en évidence les gènes cibles de PPARα (Martin *et al.*, 2005a). Cette biopuce, en développement permanent, compte aujourd'hui (juin 2005) environ 300 gènes.

Par ailleurs, les 40 souris sont réparties selon deux facteurs croisés dans un plan complet, équilibré, à quatre répétitions. Les deux facteurs sont décrits ci-dessous.

- Génotype (2 niveaux). Les souris sont soit de type sauvage (`WT`) soit de type PPARα-déficientes (`PPAR`), avec 20 souris dans chaque cas (Lee *et al.*, 1995).
- Régime (5 niveaux). Les cinq régimes alimentaires sont notés :
  - `dha` : régime enrichi en acides gras de la famille *Oméga 3* et particulièrement en acide docosahexaénoïque (DHA), à base d'huile de poisson;
  - `efad` (*Essential Fatty Acid Deficient*) : régime constitué uniquement d'acides gras saturés, à base d'huile de coco hydrogénée;
  - `lin` : régime riche en *Oméga 3*, à base d'huile de lin;

- `ref` : régime dont l'apport en *Oméga 6* et en *Oméga 3* est adapté des Apports Nutritionnels Conseillés pour la population française, sept fois plus d'*Oméga 6* que d'*Oméga 3*;
- `tsol` : riche en *Oméga 6*, à base d'huile de tournesol.

Quatre souris de chaque génotype sont soumises à chaque régime alimentaire.

La question centrale de l'analyse des données d'expression consiste à détecter des gènes qui ont un comportement différent selon les conditions auxquelles ils sont soumis. Dans notre cas, nous cherchons à déterminer les gènes :

- caractéristiques de l'effet génotype,
- caractéristiques de l'effet régime,
- en relation avec des taux élevés d'acides gras hépatiques.

Sans revenir sur la phase de normalisation des données, précisons toutefois qu'elle a été abordée en centrant les valeurs sur l'intensité moyenne de 13 ARN exogènes ajoutés à l'échantillon avant marquage (*spike*). Elle a ensuite été validée par la présence de 8 autres ARN exogènes introduits à des doses parfaitement connues dans la solution initiale (Martin *et al.*, 2005b). Nous vérifions dans la première partie de l'analyse exploratoire qu'aucune membrane ne présente un comportement atypique (voir Fig. 1).

3. Démarche exploratoire

Dans ce paragraphe, nous passons en revue diverses techniques exploratoires dans le but, d'une part, de nous familiariser avec les données, d'autre part, de commencer à identifier certains gènes ayant un comportement particulier. Le lecteur souhaitant des développements sur les techniques évoquées ci-dessous pourra les trouver dans Saporta (1990), dans Lebart *et al.* (1995) ou dans Baccini *et al.* (2005). Signalons également l'ouvrage récent de McLachlan *et al.* (2004), plus spécifiquement orienté vers l'analyse des données d'expression.

3.1. Stratégie

Dans une étude statistique sophistiquée, avant d'utiliser des méthodes de modélisation ou des techniques d'apprentissage, il est toujours prudent de commencer par une étude *exploratoire* à l'aide d'outils, certes élémentaires mais robustes, en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec les données et, surtout, de dépister les sources de problèmes comme les valeurs manquantes, erronées ou atypiques, les distributions « anormales » (dissymétrie, multimodalité, épaisseur des queues de distributions), les liaisons non linéaires...

Ensuite, au vu des résultats précédents, on peut être conduit à mettre en œuvre divers prétraitements des données afin de rendre ces dernières conformes aux hypothèses des techniques de modélisation ou d'apprentissage qu'il sera nécessaire d'utiliser pour atteindre les objectifs fixés. Ces prétraitements peuvent être les suivants :

- transformation des variables (logarithme, puissance...), centrage, réduction, passage aux rangs;
- codage en classes ou recodage de classes;
- imputation ou non des données manquantes;
- réduction de dimension, classification et premier choix de variables;
- classification ou typologie des observations.

Attention, le côté élémentaire des techniques exploratoires ne doit pas conduire à les négliger au profit d'une mise en œuvre immédiate de méthodes beaucoup plus sophistiquées, donc beaucoup plus sensibles aux problèmes cités ci-dessus. Si ces problèmes ne sont pas pris en compte au début de l'analyse, ils réapparaîtront ensuite comme autant d'*artefacts* susceptibles de dénaturer, voire de fausser, toute tentative de modélisation.

Pour l'analyse exploratoire, au-delà des techniques unidimensionnelles, deux grandes familles de méthodes sont utilisées :

- les méthodes factorielles;
- la classification (*clustering*), ou apprentissage non supervisé.

Dans les deux cas, de nombreux choix sont laissés à l'utilisateur qui doit soit les effectuer en connaissance de cause, soit les tester pour arriver à une représentation satisfaisante compte tenu de ses a priori et des conditions expérimentales. Ces choix doivent bien sûr être connectés à ceux relatifs à la normalisation des données.

3.2. Choix méthodologiques

Nous donnons ici une présentation synthétique des choix qui doivent être explicités. Cette liste n'est sans doute pas exhaustive et devra être complétée avec l'approfondissement de l'expertise du traitement de ce type de données. Nous pouvons d'ores et déjà insister sur l'indispensable dialogue entre biologiste et statisticien pour opérer ces choix en connaissance de cause, tant sur les aspects techniques que sur leurs implications biologiques. Ces choix ne pourront bien sûr pas tous être discutés en détail dans le cadre restreint de cet article.

3.2.1 Transformations

Les données traitées sont issues des procédures de normalisation spécifiques à la technologie utilisée pour les produire. Néanmoins, elles peuvent encore subir des transformations dont nous précisons ci-dessous les plus courantes.

Passage au logarithme. Cette transformation corrige une distribution trop dissymétrique (*skewness*) et réduit l'influence des grandes valeurs (éventuellement atypiques). Elle est justifiée dans la mesure où, dans certains systèmes naturels, divers effets peuvent être modélisés par des facteurs multiplicatifs plutôt qu'additifs.

Centrage. Les données se présentant sous la forme d'une matrice, il est habituel, par exemple dans une analyse en composantes principales (ACP),

de centrer les colonnes. L'information liée à la moyenne peut être utile en soi, mais est rarement très informative. Dans les données transcriptomiques, le rôle des lignes et des colonnes, c'est-à-dire la distinction entre individus et variables, n'est pas toujours explicite. Le choix doit être fait en fonction de la nécessité de centrer les données d'expression par gène, par biopuce, voire par gène et par biopuce, ce qui correspond au double centrage (lignes et colonnes).

Réduction. Dans les méthodes descriptives multidimensionnelles, il est courant de faire une réduction des variables soit lorsque leurs unités de mesure sont différentes soit, lorsqu'avec la même unité de mesure, on observe des variances très hétérogènes. Dans l'analyse de données transcriptomiques, on peut être conduit à faire une réduction des variables si celles-ci sont bien identifiées. Toutefois, il convient d'être très prudent compte tenu des problèmes suivants. En ramenant à un la variance des gènes, les effets de surexpression ou sous-expression de certains d'entre eux sont éliminés. Cette transformation a donc surtout un sens lors de l'étude d'un sous-ensemble de gènes déjà sélectionnés car différentiellement exprimés. D'autre part, dans une ACP, la réduction peut conduire à un effet taille artificiel sur l'axe 1 dû à un grand nombre de gènes d'expressions voisines et sans particularité. Nous l'avons observé dans l'ACP présentée en 3.4 où le phénomène le plus structurant, la séparation des deux génotypes, apparaît sur l'axe 1 dans l'ACP non réduite et sur l'axe 2 dans l'ACP réduite, le premier axe correspondant à un effet taille dans ce second cas.

Marges unitaires. Une autre façon d'éliminer l'influence des unités de mesure consiste à diviser les lignes (ou les colonnes) d'un tableau par ses marges en lignes (ou en colonnes). C'est la pratique courante lorsque le tableau contient des effectifs (table de contingence) et cela conduit à l'analyse des correspondances. Pour les raisons évoquées ci-dessus (surexpression et sous-expression), cette approche ne semble pas bien adaptée à l'analyse des données d'expression.

Passage aux rangs. Lorsque les données sont parsemées de valeurs atypiques sans qu'aucune transformation fonctionnelle (logarithme, puissance...) ne puisse en atténuer efficacement les effets, une façon « brutale », ou robuste, de procéder consiste à remplacer une valeur par son rang dans la séquence ordonnée. En pratique, cela revient à calculer le coefficient de Spearman plutôt que celui de Bravais-Pearson.

3.2.2 Pondérations et distances

Il arrive que l'on introduise des pondérations sur les lignes ou sur les colonnes du tableau des données. Cette pratique permet de donner plus ou moins d'importance à certains éléments ; cela permet de redresser un échantillon. Un autre exemple simple consiste à affecter des poids nuls à certaines lignes ou à certaines colonnes, alors dites supplémentaires : elles n'interviennent pas dans les calculs mais restent représentées dans les graphiques. Par défaut, les poids sont égaux pour les lignes et égaux pour les colonnes. Sauf cas très particulier, on utilise des poids égaux dans l'analyse des données d'expression.

D'un point de vue mathématique, chaque ligne (chaque colonne) est considérée comme un vecteur d'un espace vectoriel muni d'un produit scalaire induisant une norme euclidienne et une distance (entre ces vecteurs). Par défaut, cette distance est celle, classique, dont le carré est la somme des carrés des écarts entre les coordonnées de deux vecteurs. Introduire des pondérations sur les lignes (sur les colonnes) conduit à pondérer le calcul de cette distance. La matrice associée au produit scalaire est alors une matrice diagonale comportant le carré des pondérations sur la diagonale. Elle remplace la matrice identité associée à la distance classique.

D'autres matrices carrées symétriques et définies positives sont également utilisables de façon plus générale. Citons l'inverse de la matrice des variances-covariances résiduelles (ou intra-classes) en analyse discriminante (métrique de Mahalanobis) ou la matrice diagonale des inverses des fréquences marginales en analyse des correspondances (métrique du khi-deux). Citons encore les matrices largement répandues dans le contexte transcriptomique faisant intervenir la corrélation linéaire usuelle, ou la corrélation par rangs, entre deux variables X et Y, et dont le terme général peut être défini par $1 - \mathrm{cor}(X, Y)$ ou $\sqrt{1 - \mathrm{cor}(X, Y)^2}$. Le problème du choix de la distance sera discuté en 3.5.

3.2.3 Factorisation et projections

Beaucoup de méthodes proposées recherchent des nouvelles variables non corrélées (*factor scores*) obtenues par combinaisons linéaires des variables initiales et optimisant un critère. Il s'agit de la maximisation de la variance dans le cadre de l'ACP qui conduit à la construction des variables principales. La décomposition ainsi obtenue a-t-elle un sens pour les données considérées ? Combien de composantes sont nécessaires pour résumer l'information et fournir des représentations graphiques pertinentes des nuages de points (individus et variables) ? Seule la discussion entre biologiste et statisticien peut éclairer ces questions.

Sur le plan mathématique, lorsqu'on utilise la métrique identité dans l'espace vectoriel des individus, les variables principales sont obtenues à partir des facteurs principaux, vecteurs propres associés aux plus grandes valeurs propres d'une matrice carrée, symétrique et définie non négative (variances-covariances, corrélations, produits scalaires...). De manière plus générale, les métriques définies dans l'espace des individus et dans celui des variables interviennent dans la définition de la matrice diagonalisée.

3.2.4 Classification

Une approche classique dans toute discipline scientifique consiste à faire de la taxinomie, c'est-à-dire à rechercher des classes homogènes des objets étudiés (gènes ou échantillons biologiques dans notre cas) au sens d'un critère défini par une matrice de distances ou de dissemblances. Le choix de ce critère, qui sera précisé au 3.6, est évidemment prépondérant pour la signification et l'interprétation des résultats obtenus.

3.3. Techniques unidimensionnelles

Parmi les outils et indicateurs recensés, les diagrammes en boîtes présentent un grand intérêt pour l'analyse des données d'expression. Eux seuls permettent de représenter simultanément un grand nombre de distributions (quelques centaines) sur un même graphique et d'en analyser rapidement les spécificités en termes de médiane, de dispersion et de valeurs atypiques. Il serait bien entendu impossible de considérer simultanément autant d'histogrammes.

Ainsi, la représentation des diagrammes en boîtes pour les souris, ordonnées selon le génotype et le régime suivi (Fig. 1), ne donne *a priori* aucune tendance spécifique sur le comportement de l'ensemble des gènes. Cette représentation atteste de la qualité de la production et du prétraitement des données. En effet, celles-ci ont été recueillies en utilisant une membrane par souris; ainsi, une quelconque anomalie sur un support, affectant l'ensemble des mesures relatives à une souris particulière, apparaîtrait nécessairement sur cette représentation. Notons seulement que quelques gènes atypiques, facilement repérables sur la figure 2 comme les plus surexprimés, se retrouvent dans les valeurs extrêmes pour chaque souris sur la figure 1.

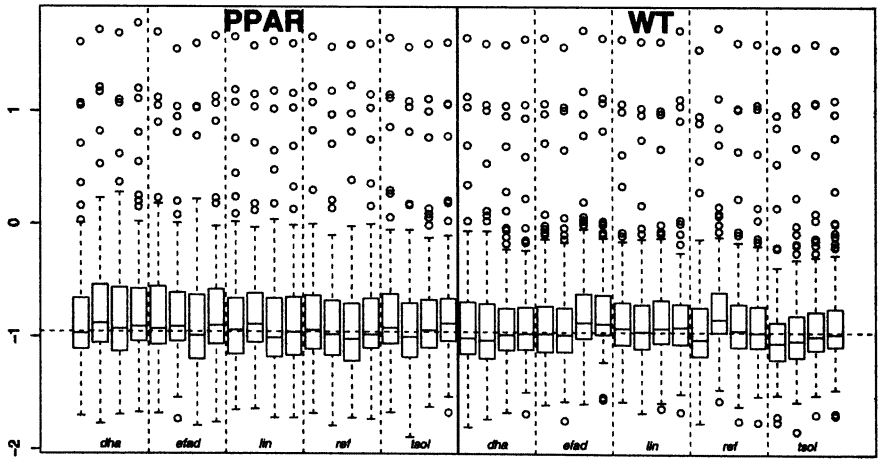

FIG 1. — Diagrammes en boîtes pour les 40 souris. La ligne verticale pleine sépare les souris selon leur génotype. Les lignes verticales en pointillés séparent les souris selon le régime qu'elles ont suivi. La ligne horizontale en pointillés représente la médiane de l'ensemble des valeurs

Les diagrammes en boîtes pour chaque gène (Fig. 2) révèlent des gènes dont l'expression est, sur l'ensemble des souris, nettement différentes des autres (par exemple, `16SR`, `apoA.I`, `apoE`). Les gènes des ARN ribosomiques, comme le `16SR` (ARN 16s ribosomique mitochondrial), présentent, dans toutes les cellules de l'organisme, des niveaux d'expression plus élevés que tous les gènes codant des ARN messagers. Ces ARN servent en effet à la traduction des ARN messagers en protéines. Par ailleurs, on peut constater que les expressions de

certains gènes varient beaucoup plus que d'autres sur l'ensemble des souris (par exemple, `FAS`, `S14` et `THIOL`). Pour ces derniers gènes, on peut supposer qu'une part de cette variabilité est due aux facteurs considérés, ce que nous essaierons de confirmer par la suite au moyen de techniques de modélisation.

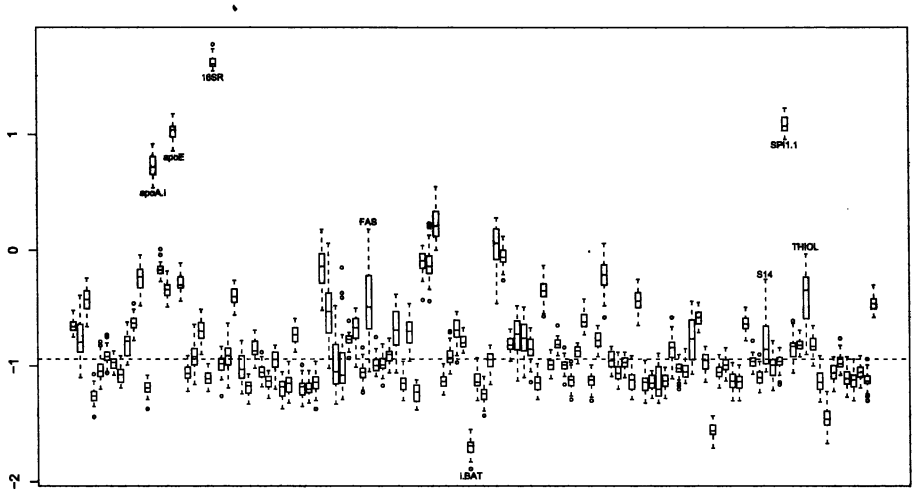

FIG 2. — Diagrammes en boîtes pour les 120 gènes. Quelques gènes particuliers ont été étiquetés

L'intérêt de ces représentations réside davantage dans la vision synthétique qu'elles offrent que dans l'information biologique que l'on peut en extraire. Elles nous orientent également dans les premiers choix méthodologiques à établir avant de poursuivre l'analyse. En effet, les boîtes relatives à la distribution des gènes mettent clairement en évidence un certain nombre de gènes dont l'expression est systématiquement supérieure à celle des autres, quelles que soient les conditions expérimentales. De plus, la variabilité de ces expressions est, le plus souvent, très faible. Ce constat nous conduit à effectuer un centrage des gènes (en colonnes), afin d'éviter un effet taille lors de la mise en œuvre de techniques factorielles. En revanche, rien dans ces représentations ne nous pousse à centrer les échantillons (en lignes), ce qui, par ailleurs, ne se justifierait pas sur le plan biologique. En effet, nous travaillons sur des données acquises via des puces dédiées sur lesquelles les gènes considérés ont été présélectionnés et sont donc, *a priori*, potentiellement différentiellement exprimés dans les conditions étudiées. Un centrage des échantillons serait susceptible de cacher des phénomènes biologiques. Ce raisonnement ne tiendrait pas pour une expérimentation pangénomique, où l'on pourrait supposer que globalement les gènes s'expriment de la même façon et que les surexprimés compensent les sous-exprimés.

3.4. Analyse en Composantes Principales

Pour la mise en œuvre de l'ACP, nous travaillons sur le tableau croisant les souris en lignes (en tant qu'individus), et les gènes en colonnes (en tant que variables). Comme nous l'avons précisé dans le point précédent, ce tableau est préalablement centré en colonnes. Le centrage évite un effet taille sans grande signification quant à la discrimination des phénomènes observés.

D'un point de vue très technique, dans les logiciels statistiques R et S-PLUS, deux fonctions sont disponibles pour réaliser des ACP :

- `princomp` ne réalise que des ACP centrées et extrait les valeurs propres et vecteurs propres de la matrice des covariances (par défaut) ou des corrélations; le nombre de colonnes de la matrice des données doit être inférieur ou égal au nombre de lignes;
- `prcomp` calcule directement la décomposition en valeurs singulières (SVD) de la matrice des données centrées (par défaut, mais avec la possibilité de ne pas centrer) et non réduite (par défaut, mais avec possibilité de réduire); cette fonction accepte un nombre de colonnes supérieur au nombre de lignes.

Compte tenu de la particularité des données transcriptomiques, nous avons systématiquement utilisé la procédure `prcomp`.

Nous donnons ci-après le tableau (tableau 1) et le graphique (figure 3) des premières valeurs propres qui nous ont conduit à considérer trois dimensions représentant environ les deux tiers de l'inertie globale.

TABLEAU 1. – Parts de variance expliquée et cumuls pour les six premiers axes principaux

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Parts de variance expliquée	0.350	0.196	0.124	0.0608	0.0447	0.0289
Cumuls	0.350	0.546	0.670	0.7311	0.7757	0.8046

Les figures 4 et 5 donnent la représentation des souris et celle des gènes, d'abord dans le premier plan principal, ensuite dans celui correspondant aux dimensions 1 et 3. Dans le cadre de cette ACP, il est cohérent de rechercher quels sont les gènes contribuant le plus à la définition des trois premiers axes. Une adaptation de l'expression classique de la contribution d'un individu fournit, pour la $j^{\text{ème}}$ variable, la quantité :

$$\frac{\sum_{k=1}^{3} \lambda_k (u_j^k)^2}{\sum_{k=1}^{3} \lambda_k}, \tag{1}$$

où les λ_k sont les valeurs propres de l'ACP et les u_j^k les coordonnées des vecteurs propres normés correspondants.

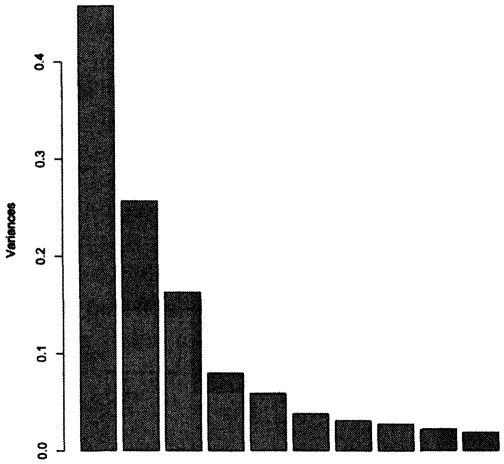

FIG 3. — Éboulis des dix premières valeurs propres de l'ACP

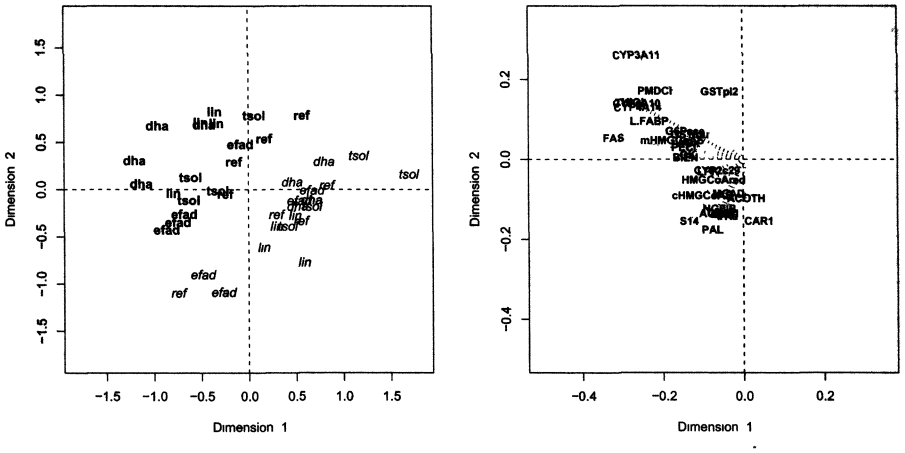

FIG 4. — Représentations sur le premier plan principal de l'ACP. À gauche : individus-souris identifiés par leur génotype (**WT** en gras, *PPAR* en italique) et leur régime (`dha`, `efad`, `lin` , `ref`, `tsol`). À droite : les 30 variables-gènes qui contribuent le plus aux trois premiers axes

Ces quantités permettent de rechercher, par exemple, les 25 % des gènes contribuant le plus à la définition de l'espace propre à trois dimensions jugé pertinent. Avec cette sélection, la représentation des variables ainsi restreinte à 30 gènes est plus facilement lisible sur les figures 4 et 5. Toutefois, dans le cas d'une puce pangénomique, avec potentiellement plusieurs milliers de gènes, une telle représentation ne serait pas exploitable.

Le premier plan (Fig. 4) doit être interprété globalement puisque sa première bissectrice sépare exactement les souris **WT** des souris *PPAR*. Les gènes à coordonnées négatives sur l'axe 1 et positives sur l'axe 2 sont sensiblement

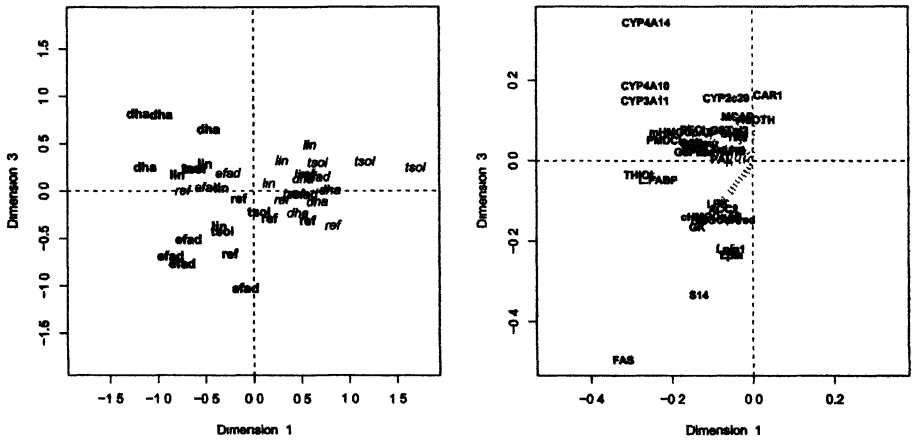

FIG 5. — Représentations sur le plan composé des axes principaux 1 et 3 avec les mêmes conventions que pour la figure 4

plus exprimés chez les souris WT, en particulier CYP3A11, CYP4A10, CYP4A14, THIOL, PMDCI, GSTpi2, L.FABP et FAS. À l'inverse, les gènes à forte coordonnée négative sur l'axe 2 s'expriment davantage chez les souris PPAR, par exemple, S14, PAL et CAR1. Ceci est en partie connu des biologistes (Aoyama *et al.*, 1998).

Le phénomène le plus marquant concernant l'axe 3 (Fig. 5) est l'opposition, chez les souris WT, entre les régimes dha, dont les coordonnées sont toutes positives, et efad, dont les coordonnées sont toutes négatives. Les gènes les plus exprimés dans le premier cas (régime dha chez les souris WT) sont CYP3A11, CYP4A10, CYP4A14, CYP2c29 et CAR1 ; dans le second cas (régime efad chez les mêmes souris), il s'agit des gènes FAS, S14, Lpin et Lpin1. Parmi ces régulations, on note une opposition entre les gènes CYP4A, connus pour être impliqués dans le catabolisme des acides gras, et les gènes FAS et S14, impliqués eux dans la synthèse des lipides. Par ailleurs, la régulation du gène CYP3A11 par l'acide DHA a déjà été décrite dans Berger *et al.* (2002).

3.5. Positionnement multidimensionnel

L'ACP considère implicitement des distances euclidiennes entre les lignes et entre les colonnes du tableau, relativement aux métriques associées à des matrices particulières : la matrice identité pour les lignes, la matrice diagonale des poids pour les colonnes. Il peut être intéressant, sur un plan biologique, de choisir d'autres distances. Ce choix est souvent proposé par les logiciels de traitement statistique de données transcriptomiques destinés aux biologistes. Le positionnement multidimensionnel (*multidimensional scaling*, ou MDS), également appelé ACP d'un tableau de distances, permet d'illustrer les effets des différents choix.

L'objectif du MDS (Mardia *et al.*, 1979) est de rechercher, dans un espace euclidien de dimension réduite q fixée *a priori*, la représentation euclidienne

la plus proche, c'est-à-dire respectant au mieux les distances (pas nécessairement euclidiennes) entre les individus décrits dans une matrice symétrique D, $(n \times n)$. Plus précisément, il s'agit de construire une matrice $(n \times q)$, dite des coordonnées principales, de sorte que les distances euclidiennes calculées entre les vecteurs lignes de cette matrice soient globalement les plus proches des distances initiales. Les coordonnées principales fournissent alors la représentation graphique recherchée. Dans le cas très particulier où la matrice des distances D est euclidienne, calculée à partir de la matrice $X(n \times p)$ des coordonnées des individus, l'ACP de X coïncide à un facteur près au MDS de D.

Nous avons procédé au positionnement multidimensionnel des données d'expression génique à partir d'une matrice de distance inter-gènes calculée selon différentes formules :

- distance euclidienne, $d_1(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$, positive ou nulle ;
- distance associée à la corrélation carrée, $d_2(X,Y) = \sqrt{1 - \mathrm{cor}(X,Y)^2}$, comprise entre 0 et 1 ;
- distance associée à la corrélation, $d_3(X,Y) = 1 - \mathrm{cor}(X,Y)$, comprise entre 0 et 2.

Remarquons tout d'abord que, dans les trois cas, plus la valeur est petite plus les gènes dont on mesure l'éloignement sont proches. Ensuite, pour d_2 et d_3, une valeur proche de 1 caractérise deux gènes non corrélés, ce qui n'est pas nécessairement le cas de la distance euclidienne. Enfin, il est important de noter qu'une corrélation forte et négative entre deux gènes conduit à deux résultats opposés selon d_2 (valeur proche de 0) et d_3 (valeur proche de 2), d_2 ne dépendant pas du signe de la corrélation contrairement à d_3.

La figure 6 illustre les trois possibilités avec le positionnement multidimensionnel des gènes. L'analyse conjointe de ces trois graphiques conduit à de nombreuses interprétations sur le plan biologique. Sans entrer dans les détails, nous noterons que ces trois graphiques tendent à séparer deux groupes de gènes qui interviennent dans deux fonctions biologiques opposées : les `CYP4A`, `PMDCI`, `PECI`, `AOX`, `BIEN`, `THIOL`, `CPT2`, `mHMGCoAS`, `Tpalpha` et `Tpbeta` sont impliqués dans le catabolisme des lipides et la cétogénèse alors que les gènes `FAS`, `S14`, `ACC2`, `cHMGCoAS`, `HMGCoAred` et, plus indirectement, `GK` et `LPK` sont impliqués dans la synthèse de lipides au niveau hépatique. On observera qu'aucun des trois graphiques de la figure 6, analysé individuellement, ne conduit à la totalité de cette interprétation mais que c'est bien l'analyse conjointe de ces représentations qui permet d'affiner la connaissance du biologiste sur ces données. Succintement, notons également que d'autres gènes tendent à participer à ces groupes. Par exemple, le gène `Lpin1` est proche des gènes impliqués dans la lipogénèse. Bien que sa fonction soit actuellement inconnue, Peterfy *et al.* (2001) ont observé que la lignée de souris déficiente pour `Lpin1` présente des altérations du métabolisme des lipides.

Les gènes dont la position sur le graphique sera la plus modifiée en passant de la distance d_2 à la distance d_3 seront ceux présentant des corrélations négatives et importantes avec de nombreux autres gènes. Un cas typique dans notre exemple est celui de `CAR1` dont l'ACP (et d'abord la matrice des corrélations)

a montré qu'il était négativement corrélé avec des gènes tels que `GSTpi2`, `CYP3A11`, `FAS`... La position relative des couples de gènes ainsi obtenus change de façon importante entre les deux graphiques. On observera en particulier le couple `CAR1-GSTpi2` totalement opposé sur l'axe 1 selon d_3 et relativement proche selon d_2 (tandis qu'il présente une opposition moins marquée selon d_1). La surexpression du gène `CAR1` et la sous-expression du gène `GSTpi2` chez les souris déficientes en récepteur PPARα n'ont pas été décrites et constituent l'un des résultats originaux de ce travail. L'étude d'un lien potentiel entre ces deux modifications d'expression nécessitera la mise en œuvre d'expériences complémentaires.

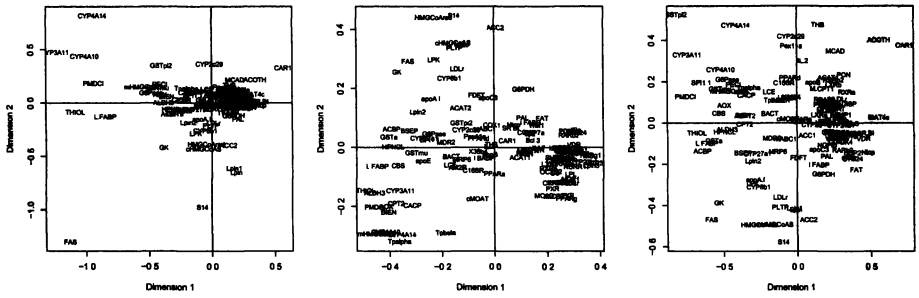

FIG 6. — Positionnement multidimensionnel des gènes sur les axes 1 et 2 selon 3 distances différentes : distance euclidienne (d_1 à gauche), corrélation carrée (d_2 au centre), corrélation (d_3 à droite)

TABLEAU 2. – Comparaison de la distance euclidienne (d_1), de la distance basée sur la corrélation carrée (d_2) et de la distance basée sur la corrélation (d_3) sur un exemple élémentaire. Les calculs sont effectués pour chacune des variables Yi ($i = 1, ..., 4$) par rapport à la variable X

	X	$Y1$	$Y2$	$Y3$	$Y4$
	1	2	-2	10	2
	2	3	-3	20	3
	3	4	-4	30	5
	4	5	-5	40	4
	5	6	-6	50	1
$d_1(X, Y_i)$	0	2,2	16,9	66,8	4,7
$d_2(X, Y_i)$	0	0	0	0	0,99
$d_3(X, Y_i)$	0	0	2	0	1,1

Ces tendances générales sont illustrées dans le tableau 2 sur la base d'un exemple élémentaire. Ce tableau a été construit de manière à illustrer les comportements différents des trois distances envisagées. Hormis dans le cas de $Y1$ où d_1, d_2 et d_3 donnent la même indication (faible éloignement), les conclusions divergent selon la distance envisagée. Pour $Y2$, corrélée négativement avec X, d_1 et d_3 sont importantes alors que d_2 est nulle, comme entre

X et $Y1$. Pour $Y3$, d_1 montre deux variables très éloignées alors que d_2 et d_3 sont nulles. Pour $Y4$, non corrélée avec X, d_1 est relativement faible alors que d_2 et d_3 sont proches de 1.

D'une manière générale, on peut retenir que l'utilisation de la distance euclidienne tend à rapprocher des gènes dont les expressions sont proches. En revanche, les deux autres indicateurs considèrent que deux gènes sont proches si leur expression varie dans le même sens selon les conditions expérimentales. La corrélation (d_3) distingue les gènes corrélés négativement, ce que ne permet pas la corrélation carrée (d_2) qui doit donc être utilisée en connaissance de cause.

Notons que la distance d_1 est plus courante en statistique alors que d_3 l'est davantage dans les études relatives aux biopuces. Autant que possible, une comparaison des trois approches est recommandée. On se référera à Draghici (2003, chapitre 11) pour une discussion plus détaillée sur le sujet.

3.6. Classification

La mise en œuvre d'une méthode de classification non supervisée vise à obtenir une partition des individus (ou des variables) en classes, sans *a priori* sur le nombre de classes. Toute méthode de ce type est basée sur deux critères de distance : l'un entre individus (ou variables), l'autre entre groupe d'individus (ou de variables). Le premier choix est guidé par les mêmes considérations que dans le cas du MDS : les mêmes distances entre gènes peuvent être considérées avec les mêmes implications biologiques. Depuis les travaux de Eisen *et al.* (1998), la distance la plus fréquemment rencontrée pour l'étude du transcriptome est du type de d_3, basée sur la corrélation ; il nous semble cependant pertinent d'utiliser les trois types de distances et d'en apprécier la complémentarité quant à l'interprétation des résultats, ce que nous avons fait ci-dessus pour le MDS. En revanche, nous nous contenterons ici de présenter une classification basée sur la distance euclidienne d_1. Le deuxième choix intervenant en classification concerne le critère d'agglomération, c'est-à-dire la façon dont est définie la distance entre deux groupes, et n'a pas d'interprétation biologique simple. Ce choix a plus une implication géométrique, sur la forme des classes obtenues. Nous avons utilisé le critère de Ward (regroupement des deux classes minimisant la perte d'inertie inter-classes, voir Lebart *et al.*, 1995) parce qu'il favorise la construction de classes relativement « sphériques » et qu'on peut lui associer des critères guidant la détermination du nombre de classes (Baccini *et al.*, 2005).

Une représentation très répandue en analyse des données de biopuces consiste à fournir simultanément les résultats de deux classifications ascendantes hiérarchiques menées indépendamment sur les lignes (individus-souris) et sur les colonnes (variables-gènes) du tableau de données.

L'interprétation de la figure 7 présente des analogies avec celle de l'ACP sur le premier plan principal. Si l'on s'intéresse aux individus-souris, on peut constater que les deux génotypes sont différenciés en deux groupes, à l'exception de trois souris de type `PPAR` ayant suivi les régimes **`efad`** (pour deux d'entre elles) et **`ref`**. Ce sont ces trois mêmes individus que l'on retrouve

projetés dans la partie négative du premier axe de l'ACP (Fig. 4). Pour les variables-gènes, un groupe attire particulièrement l'attention sur l'image : sur une bande verticale (encadrée en pointillés sur la figure) correspondant à 14 gènes, les niveaux de gris sont nettement plus variables que sur le reste de l'image. Il s'agit des gènes

CYP4A10, CYP4A14, CYP3A11, L.FABP, THIOL, PMDCI, S14, Lpin1, Lpin, FAS, GSTmu, GSTpi2, CYP2c29, G6Pase

qui apparaissent tous parmi les gènes les plus corrélés aux trois premiers axes principaux de l'ACP (Fig. 4 et 5)

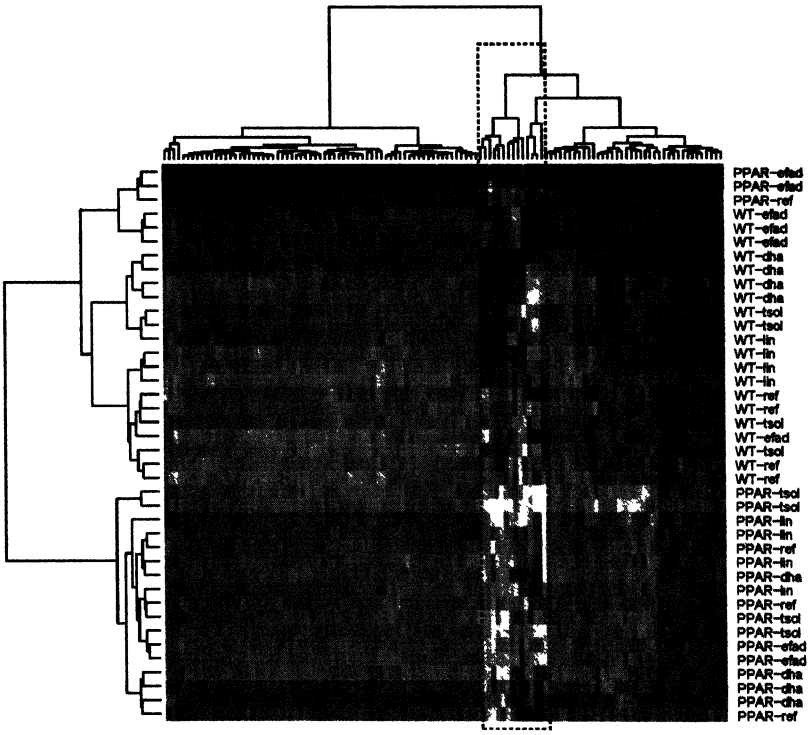

FIG 7. — Représentation simultanée des résultats des classifications ascendantes hiérarchiques des individus-souris et des variables-gènes selon la méthode de Ward, avec la distance euclidienne. Sur l'image du tableau des données, les intensités sont croissantes du blanc vers le noir. Le rectangle en pointillés met en évidence 14 gènes au comportement particulier

MDS et classification apparaissent donc comme des techniques complémentaires, mais elles ne sont pas sensibles de la même façon aux perturbations. La perturbation d'une donnée peut fortement influencer la structure d'un dendrogramme alors qu'en MDS, la prise en compte conjointe de toutes les distances deux à deux assure une certaine robustesse pour le calcul des coordonnées

principales. Pour cette raison, il est utile de représenter les classes dans une projection sur les axes factoriels obtenus soit par MDS soit par ACP. L'éboulis des valeurs propres (Fig. 8) nous oriente vers une représentation du MDS en deux dimensions.

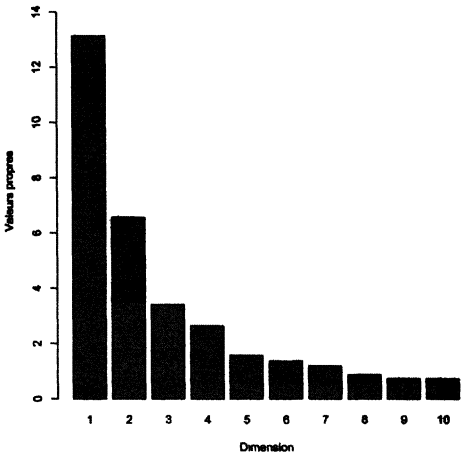

FIG 8. — Éboulis des valeurs propres pour le MDS de la matrice des distances euclidiennes inter-gènes

La représentation de la figure 9 est analogue à celle déjà présentée (Fig. 6, graphique de gauche). Elle est complétée par un codage graphique des gènes, selon leur appartenance à un groupe issu de la classification hiérarchique. Pour cela, nous avons coupé l'arbre afin d'en extraire 5 groupes.

Brièvement, on peut noter que l'axe 1 met en évidence l'opposition précédemment évoquée entre `CAR1` (surexprimé chez les souris `PPAR`) et un groupe de gènes (`CYP3A10`, `CYP4A10`, `CYP4A14`, `PMDCI`, `THIOL` et `L.FABP`) qui est surexprimé chez les souris `WT`. De manière similaire, l'axe 2 oppose les gènes induits par le régime `dha` (valeurs positives, gènes impliqués dans le catabolisme des lipides et dans le métabolisme des xénobiotiques) aux gènes induits par le régime `efad` (valeurs négatives, gènes principalement impliqués dans la synthèse des lipides). En remontant vers les feuilles de l'arbre de classification, on notera que le groupe des gènes représentés au centre et en italique est séparé en deux sous-groupes qui conservent une cohérence vis-à-vis des fonctions biologiques de catabolisme et de synthèse des lipides respectivement. Une observation des données individuelles révèle que ces régulations opérées par les régimes semblent plus marquées chez les souris `WT`. Nous verrons ultérieurement que d'autres techniques (forêts aléatoires par exemple) permettent de confirmer ces observations de manière plus objective.

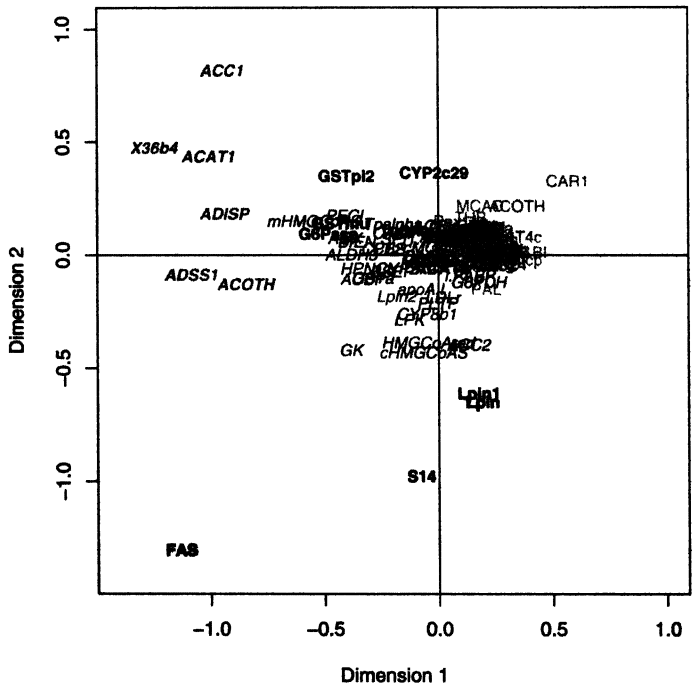

FIG 9. — Représentation par positionnement multidimensionnel (distance euclidienne) des 5 groupes issus de la classification hiérarchique des gènes

4. Modélisations

4.1. Analyses de variance et modèle mixte

Avant de détailler les modèles utilisés, précisons tout de suite que les tests mis en œuvre en ANOVA ou dans les modèles mixtes supposent des observations indépendantes. Or, pour un génotype donné, un régime donné et une souris donnée, les expressions des différents gènes de cette souris ne sont pas indépendantes, à cause des régulations entre gènes. Dans une expérience comme celle considérée ici, il n'est pourtant pas possible de prendre en compte cette structure de dépendance entre gènes. En effet, la prise en compte, par exemple, d'un modèle mixte avec le facteur gène en effet aléatoire et une structure de covariance « pleine » entraînerait l'estimation de 7260 paramètres $\left(\frac{120(120+1)}{2}\right)$. Toutefois, certaines simulations ont montré que les tests considérés restent en général acceptables, en dépit de la dépendance des données (voir, par exemple, Nguyen, 2004 et Dudoit *et al.*, 2003).

4.1.1 Principe des analyses de variance

L'analyse de variance (ANOVA) permet d'apprécier l'effet d'une ou de plusieurs variables qualitatives (les facteurs) sur une variable quantitative (la variable réponse, ici le niveau d'expression des gènes). Dans le domaine de l'analyse transcriptomique, cette approche a été largement développée, en particulier par Kerr *et al.* (2000). Pour l'analyse de nos données, un modèle d'ANOVA à trois facteurs (génotype, régime, gène) permet de mettre en évidence des effets de l'interaction d'ordre 3 très significatifs à l'aide du test de Fisher. Cela signifie qu'il existe des gènes régulés simultanément par le régime et le génotype, les effets du régime et du génotype étant non additifs. Le modèle d'ANOVA considéré s'écrit

$$y_{ijkl} = g_i + r_j + G_k + gr_{ij} + gG_{ik} + rG_{jk} + grG_{ijk} + e_{ijkl} \qquad (2)$$

où y_{ijkl} représente le logarithme du niveau d'expression du gène k ($k = 1, ..., 120$), pour le régime j ($j = 1, ..., 5$) et le génotype i ($i = 1, 2$), mesuré chez la souris l ($l = 1, ..., 4$) ; g_i représente l'effet du génotype i, r_j celui du régime j, G_k celui du gène k, gr_{ij} représente l'effet de l'interaction du génotype i et du régime k, gG_{ik} l'effet de l'interaction du génotype i et du gène k, rG_{jk} l'effet de l'interaction du regime j et du gène k et grG_{ijk} représente l'interaction d'ordre 3 combinant le génotype i, le régime j et le gène k. On suppose que les résidus e_{ijkl} du modèle sont indépendants et identiquement distribués suivant une loi normale de moyenne nulle et de variance σ^2. L'écriture d'un tel modèle suppose que les gènes sont tous de même variabilité. Cette hypothèse est discutable (en effet, la figure 2 montre clairement quelques gènes à forte variabilité) ; nous verrons par la suite comment lever cette difficulté.

Une autre approche consiste à écrire un modèle d'ANOVA par gène, sous la forme

$$y_{ijl} = g_i + r_j + gr_{ij} + e_{ijl} \qquad (3)$$

où les notations utilisées ici sont identiques à celles du modèle (2). Ici, il est nécessaire de faire autant d'analyses de variance que de gènes étudiés (soit 120 dans notre exemple) mais nous disposerons d'une variance estimée par gène. Toutefois, une telle analyse n'est pas toujours recommandée car, en règle générale, le nombre d'observations par gène est très faible, ce qui conduit à des estimations de variance très peu précises. Notons cependant que ces 120 analyses conduisent à 120 estimations des 10 effets $genotype_i \times regime_j$. Un modèle équivalent, mais utilisant simultanément l'ensemble des données pour estimer les paramètres, s'écrit comme le modèle (2) en posant

$$\mathrm{var}(e_{ijkl}) = \sigma_k^2 \qquad (k = 1, \ldots, 120). \qquad (4)$$

D'autre part, entre le modèle (2), supposant toutes les variances des gènes égales, et le modèle (4), supposant une variance différente pour chaque gène, il est possible d'ajuster un modèle intermédiaire prenant en compte les hétérogénéités de variances de l'expression des gènes, en définissant simplement des groupes de gènes de variabilité homogène (Robert-Granié *et al.*, 1999 ; Foulley *et al.*, 2000 ; San Cristobal *et al.*, 2002 ; Delmar *et al.*, 2005).

Ainsi, sur les 120 gènes analysés, un histogramme des variances nous a conduit à définir trois groupes de gènes ayant des variabilités très différentes : un groupe contenant les gènes FAS, G6Pase, PAL et S14, présentant des variabilités résiduelles importantes (variances supérieures à 0.02) ; un deuxième groupe à variabilité modérée (variances comprises entre 0.009 et 0.02), comprenant les gènes CYP2c29, CYP3A11, CYP4A10, CYP4A14, CYP8b1, GSTmu, GSTpi2, L.FABP, Lpin, Lpin1, TRa et cHMGCoAS ; enfin un dernier groupe à faible variabilité (variances inférieures à 0.009), contenant l'ensemble des autres gènes. À partir de ces trois groupes de gènes, nous pouvons construire un modèle dont la variance dépend de cette nouvelle variable à trois classes. Le modèle s'écrit encore comme les modèles (2) et (4) en posant cette fois

$$\mathrm{var}(e_{ijkl}) = \sigma_h^2, \tag{5}$$

où $h = \{1, 2, 3\}$ représente l'indice d'hétérogénéité de variance.

Nous pouvons ainsi comparer les gènes différentiellement exprimés selon les 3 modèles :

- Modèle (2), modèle d'ANOVA avec une unique variance pour l'ensemble des gènes ;
- Modèle (4), modèle d'ANOVA avec une variance différente par gène ;
- Modèle (5), modèle d'ANOVA avec trois groupes de variances différentes.

Notons que le modèle (4) implique l'estimation de 120 variances différentes, alors que le modèle (5) ne nécessite l'estimation que de trois paramètres de variances ; ce dernier est donc beaucoup plus économe en nombre de paramètres à estimer. Enfin, d'un point de vue technique et opérationnel, la mise en oeuvre de ces modèles peut être réalisée en utilisant la fonction `lme` du logiciel statistique R ou la procédure `mixed` du logiciel SAS.

4.1.2 Problème des tests multiples

L'objectif de l'analyse statistique est de déterminer quels sont les gènes différentiellement exprimés entre les 2 génotypes et les 5 régimes. Quelle que soit la méthode statistique utilisée, il existe une probabilité non nulle (risque de première espèce α) de détecter des faux positifs (gènes déclarés différentiellement exprimés alors qu'ils ne le sont pas) et une autre probabilité non nulle (risque de deuxième espèce β) de ne pas être capable de détecter des gènes réellement différentiellement exprimés (faux négatifs). Il est bien entendu souhaitable de minimiser ces deux probabilités d'erreur sachant que, toutes choses égales par ailleurs, la seconde augmente quand la première diminue et réciproquement. Le test de Student est couramment utilisé pour tester l'égalité de deux moyennes (l'hypothèse nulle consistant à considérer que les moyennes des intensités des signaux d'un gène donné dans chacune des deux conditions sont égales). Ainsi, quand la statistique de Student excède un certain seuil (dépendant du risque de première espèce α choisi, généralement 5 %), les niveaux d'expression du gène étudié entre les deux populations testées sont considérés comme significativement différents. Lorsque l'on souhaite tester plus de deux conditions, le test de Fisher, qui est une extension du

test de Student, est utilisé. L'hypothèse nulle consiste à supposer l'absence d'expression différentielle d'un gène entre les diverses conditions et l'hypothèse alternative à supposer une différence d'expression.

Bien sûr, prendre un risque de 5 % dans une expérimentation où 10 000 gènes, par exemple, sont étudiés simultanément peut conduire à obtenir autour de 500 faux positifs, ce qui est parfaitement inacceptable. C'est pourquoi ont été proposées des modifications du test de Student adaptées à l'analyse du transcriptome (méthodes de Bonferroni, FWER, FDR...). Le lecteur souhaitant des détails sur ces approches peut se référer, entre autres, à Benjamini & Hochberg (1995), Bland & Altman (1995), Dudoit *et al.* (2002) ou Speed (2003) et plusieurs articles de ce même volume.

La méthode de Bonferroni est une méthode qui ne permet pas un strict contrôle de α, mais qui en donne une majoration. Pour avoir un risque global α, il faut que chacune des p comparaisons soit effectuée avec un risque $\alpha' = \alpha/p$. En pratique, Bonferroni fournit une liste de gènes différentiellemnt exprimés dans laquelle on contrôle le nombre de faux positifs. Mais, lorsque le nombre des gènes est grand, cette liste est souvent vide.

En général, on présente ces taux d'erreurs dans le tableau suivant où m tests sont effectués :

Réalité	Décision		
	H_0 vraie	H_1 vraie	Total
H_0 vraie	U	V	m_0
H_1 vraie	T	S	m_1
	W	R	m

Pour une analyse de biopuces dans laquelle on teste les effets différentiels de m gènes, R est le nombre de gènes déclarés différentiellement exprimés, alors que m_1 est le nombre réel (mais inconnu) de gènes différentiellement exprimés. Diverses méthodes sont proposées pour contrôler ces divers taux d'erreurs. Nous en précisons deux ci-dessous.

Le *Family Wise Error Rate* (FWER) représente la probabilité d'effectuer au moins une erreur de première espèce sur l'ensemble des comparaisons :

$$P[V \geqslant 1] \leqslant m\alpha.$$

On prend donc un seuil nominal de $\alpha' = \alpha/m$. Au même titre que Bonferroni, plus il y a de tests (c'est-à-dire de gènes à tester), moins on rejette H_0 (moins il y a de gènes déclarés différentiellement exprimés). La notion suivante est très utile pour pallier cet inconvénient.

Le *False Discovery Rate* (FDR) contrôle l'espérance du taux de faux positifs, ou le nombre de faux positifs parmi les différences déclarées significatives. Pratiquement, on range par ordre croissant les m *p-values* des m tests (les gènes), on recherche le plus grand rang k des *p-values* tel que

$$p - value\ (k) < \alpha k/m$$

et on sélectionne les gènes de rang plus petit que k.

Il existe d'autres approches récentes, ou en cours de développement, pour contrôler le FDR, le nombre moyen d'erreurs... (voir, par exemple, Bar-Hen et Robin, 2003, Aubert *et al.*, 2004, Broët *et al.*, 2004, Dalmasso *et al.*, 2005, Delmar *et al.*, 2005, ou encore plusieurs articles de ce même volume).

Pour revenir à notre étude, à partir de chaque modèle proposé dans le paragraphe précédent, nous pouvons rechercher les gènes différentiellement exprimés entre les deux génotypes à régime fixé (120 comparaisons pour chacun des 5 régimes) ou entre régimes à génotype fixé (1 200 comparaisons par génotype), ce qui conduit à effectuer 3 000 comparaisons. Le tableau 3 présente le nombre de gènes sélectionnés selon les trois modèles considérés et selon le test ou l'ajustement utilisé (Student, Bonferroni et Benjamini-Hochberg qui correspond à l'approche FDR).

TABLEAU 3. – Nombre de gènes sélectionnés selon le modèle et le test utilisés

Tests	Modèle (2)	Modèle (4)	Modèle (5)
Student à 5 %	85	103	97
Student à 1 %	55	65	67
Benjamini-Hochberg à 5 %	44	56	57
Benjamini-Hochberg à 1 %	35	40	38
Bonferroni à 5 %	26	25	26
Bonferroni à 1 %	20	22	24

On peut remarquer que, pour un mode d'ajustement donné, le nombre de gènes sélectionnés est peu différent selon le modèle utilisé et que, globalement, les trois modèles sélectionnent le même groupe de gènes. Les petites différences sont principalement liées à l'ordre de sélection de ces gènes.

D'autre part, on peut, à partir de critères de sélection de modèle tels que le critère d'Akaike (AIC; Akaike, 1974) ou le critère de Schwarz (BIC; Schwarz, 1978), ou encore en effectuant un test du rapport de vraisemblance, choisir le modèle le plus adéquat.

Le tableau 4 présente les valeurs des critères AIC et BIC pour les trois modèles mis en compétition.

TABLEAU 4. – Valeurs des critères AIC et BIC

Modèles	-2AIC	-2BIC
(2)	-6576.9	-6570.7
(4)	-6946.6	-6612.1
(5)	-7044.5	-7036.2

Le meilleur modèle est celui pour lequel les valeurs des critères -2AIC ou -2BIC sont les plus petites. Dans les deux cas, il s'agit du modèle (5).

Le test du rapport de vraisemblance consiste, quant à lui, à comparer deux modèles emboîtés (par exemple, (2) *vs* (4) : l'hypothèse nulle considérée suppose alors que toutes les variances sont égales). La statistique du rapport de vraisemblance nécessite de calculer la différence entre les logarithmes des vraisemblances sous chacun des deux modèles. Sous l'hypothèse nulle, cette statistique suit asymptotiquement une loi de khi-deux dont le nombre de degrés de liberté est égal à la différence des nombres de paramètres à estimer sous chacun des deux modèles considérés. Si nous effectuons ces différents tests du rapport de vraisemblance ((2) *vs* (4), (2) *vs* (5), (4) *vs* (5)), il en ressort que le modèle (5), avec trois groupes de variances, est encore le meilleur.

À partir de ce modèle (5), on peut estimer les différents effets du modèle, et s'intéresser aux différences d'expression des gènes entre génotypes à régime fixé ou encore aux différences d'expression des gènes entre régimes à génotype fixé.

En raison de la multiplicité des tests, la correction proposée par Benjami & Hochberg (1995), souvent considérée comme la plus appropriée, a été utilisée. Lorsque nous considérons les différences d'expression des gènes entre génotypes à régime fixé, l'hypothèse nulle représente l'absence d'expression différentielle d'un gène entre les deux génotypes. On peut visualiser l'ensemble des résulats des *p-values* de ces différents tests en effectuant une ACP centrée sur le tableau contenant, en lignes, les 57 gènes différentiellement exprimés (selon le modèle (5) et la correction de Benjamini & Hochberg à 5 % (Tab. 3)) et, en colonnes, l'opposé du logarithme de la *p-value* associée au différentiel d'expression entre les deux génotypes pour les cinq régimes.

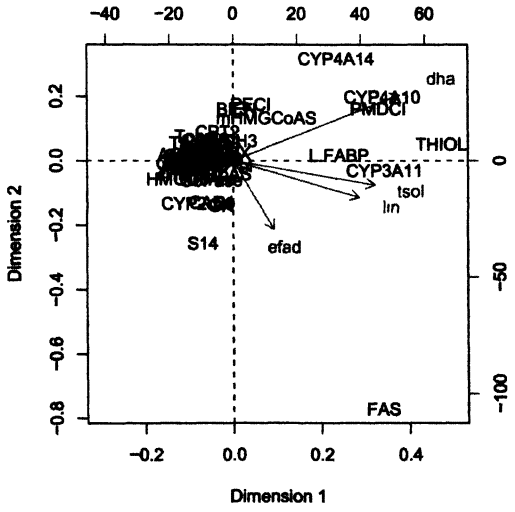

FIG 10. — Représentation sur le premier plan principal de l'ACP de l'opposé du logarithme des p-values des gènes différentiellement exprimés entre les deux génotypes à régime fixé

On observe que les gènes CYP3A11, CYP4A10, CYP4A14, L.FABP, PMDCI et THIOL différencient les deux génotypes pour les régimes dha, lin et tsol. Certains de ces gènes présentent des expressions constitutives différentielles entre les souris des deux génotypes. De plus, ces gènes sont régulés positivement par ces trois régimes riches en acides gras polyinsaturés d'une famille particulière (*Oméga 3* pour dha et lin et *Oméga 6* pour tsol) chez les souris WT, alors que la régulation de plusieurs de ces gènes est altérée chez les souris PPAR. Les gènes mHMGCoAS, PECI et BIEN apparaissent dans le contraste entre génotypes pour le régime dha, alors que les gènes S14 et FAS apparaissent pour le régime efad. Les souris des deux génotypes présentent là encore des régulations différentielles de ces gènes, soulignant ainsi le rôle du récepteur PPARα dans ces modulations d'expression provoquées par les régimes alimentaires.

Nous adoptons la même approche que précédemment, cette fois-ci sur les effets différentiels entre couples de régimes, à génotype fixé. L'ACP est donc ici réalisée sur le tableau de l'opposé du logarithme des *p-values* associées au différentiel d'expression des 57 gènes pour les dix différences entre les régimes pris deux à deux, à génotype fixé. La figure 11 présente le premier plan principal des gènes différentiellement exprimés entre régimes pour le génotype WT (à gauche) et pour le génotype PPAR (à droite). Les deux premiers axes, pour chacune des figures, représentent respectivement 79 % et 78 % de la variance totale. Les gènes Lpin et Lpin1 apparaissent dans des contrastes impliquant le régime efad pour le génotype WT et le régime tsol pour le génotype PPAR. Le gène CYP3A11 est impliqué dans le régime dha, quel que soit le génotype. Les gènes FAS et S14 apparaissent dans les contrastes impliquant le régime efad pour le génotype WT, alors que le même gène FAS apparaît dans les contrastes impliquant le régime ref pour le génotype PPAR. L'ensemble de ces résultats confirme les résultats obtenus avec l'ACP présentée en 3.4.

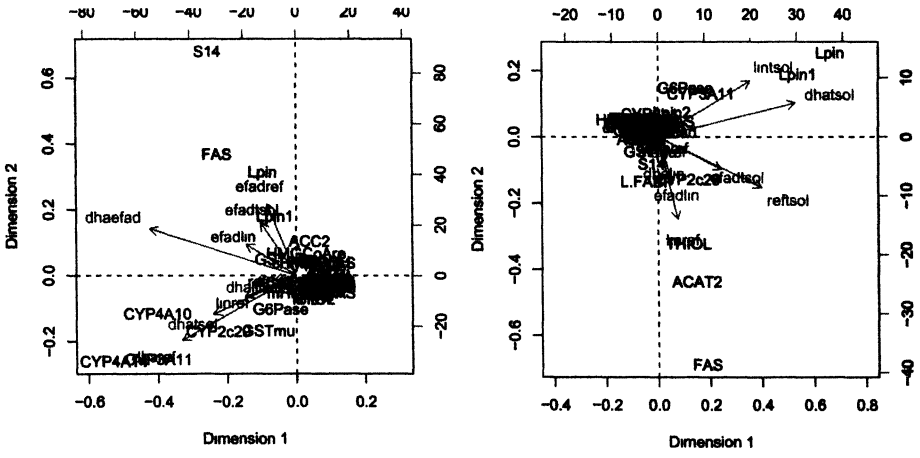

FIG 11. — Représentation sur le premier plan principal de l'ACP de l'opposé du logarithme des p-values des gènes différentiellement exprimés entre les régimes pour le génotype WT à gauche et PPAR à droite

4.1.3 Modèle mixte

Les souris étant issues d'une lignée consanguine, elles ont été considérées, dans un premier temps, comme des répétitions indépendantes et identiquement distribuées. Cependant, à l'aide d'un modèle linéaire mixte, chaque souris peut être considéree comme un tirage aléatoire dans une population plus large de souris. Le modèle linéaire mixte mis en œuvre s'écrit

$$y_{ijkl} = g_i + r_j + G_k + gr_{ij} + gG_{ik} + rG_{jk} + grG_{ijk} + souris_l + e'_{ijkl}, \quad (6)$$

où $souris_l$ représente l'effet aléatoire de la souris l, avec $souris_l \sim \mathcal{N}(0, \sigma_s^2)$, les différentes réalisations étant indépendantes, et les e'_{ijkl} représentent les résidus, avec $e'_{ijkl} \sim \mathcal{N}(0, \sigma_e^2)$, les résidus étant indépendants entre eux et indépendants de l'effet aléatoire souris.

Dans ce cas, les estimations des composantes de la variance sont pour la variance « souris » de 0.001 et pour la variance résiduelle de 0.007. La variabilité indivuelle est donc très faible. La variance des observations est identique à celle obtenue à l'aide d'une ANOVA (modèle à effets fixes) puisque nous sommes dans le cadre d'un plan équilibré et que la méthode d'estimation pour le modèle mixte est la méthode du maximum de vraisemblance restreinte (REML). Pour les mêmes raisons, les tests mis en œuvre dans ce modèle mixte pour tester la significativité des différents effets fixes sont identiques à ceux réalisés avec le modèle (2) ; les gènes mis en évidence sont donc les mêmes. Notons encore qu'il est possible d'étendre ce modèle aux cas de variances résiduelles hétérogènes, comme c'était le cas dans le modèle (5).

L'application du modèle linéaire mixte est beaucoup plus appropriée lorsque la variabilité due à la technique, à la diversité génétique ou aux gènes de la biopuce, présente un intérêt. C'est le cas dans l'étude transcriptomique décrite dans Bonnet *et al.* (2004) concernant la folliculogénèse chez la truie, dans laquelle le logarithme du signal est modélisé en fonction des facteurs membrane, animal, aiguille (ou bloc), jour d'hybridation, et des covariables logarithme de l'intensité du bruit de fond et de l'hybridation en sonde vecteur. Après une étape de choix de modèle (à l'aide du test de Fisher), le modèle linéaire mixte permet d'appréhender et de quantifier la part de variabilité due aux différentes sources de variation. Dans cet exemple relatif aux truies, la part de variabilité due à la diversité génétique représente 8 %, celle due à la technique 4 % et celle due aux gènes 75 %. Toute inférence basée sur ce modèle sera valide pour tout animal, toute membrane... car l'échantillonnage des animaux, des membranes... de cette étude, dans une population plus large d'animaux, de membranes... est pris en compte. Considérer les membranes (par exemple) comme effets fixes dans ce modèle aurait entraîné des conclusions valides uniquement sur les membranes de l'expérience. De plus, une structure de covariance non diagonale est prise en compte par ce modèle mixte puisque deux signaux d'une même membrane seront corrélés, la corrélation étant égale à $\sigma^2_{membrane}/\sigma^2_{totale}$.

4.2. Forêts aléatoires

Les techniques précédemment décrites et utilisées (ACP, MDS, modèles gaussiens) sont, par construction, linéaires et, pour les approches de type ANOVA, reposent principalement sur une significativité d'ajustement d'un modèle. Une autre approche (Speed, 2003) est basée sur la recherche d'un sous-ensemble de gènes susceptibles de construire le meilleur modèle de discrimination des génotypes, ou des régimes, à partir des expressions. Parmi les approches non linéaires, certaines semblent relativement prometteuses, à condition de satisfaire, avant tout, une règle incontournable de parcimonie pour éviter un sur-ajustement, ou sur-apprentissage de l'échantillon. Analyse discriminante linéaire ou quadratique, modèles polynômiaux, logistiques, réseaux neuronaux, arbres de classification, k plus proches voisins... s'en accomodent mal. Parmi les nombreuses approches non linéaires récentes qui se préoccupent de la complexité du modèle (*support vector machine* ou SVM, *boosting, bagging*...) nous avons privilégié les forêts aléatoires (*random forest*) proposées par Breiman (2001) qui sont construites par agrégation d'arbres de classification. La même démarche pourrait être mise en œuvre avec les SVM.

Le modèle de forêt aléatoire est un exemple de *bagging*, c'est-à-dire un algorithme d'agrégation de modèles, spécialement adapté aux arbres de décision. Son principe est simple ; il associe deux niveaux de randomisation :

1. chaque élément de la forêt est un arbre de décision de type CART (Breiman *et al.*, 1984) estimé sur un échantillon *bootstrap* de l'échantillon initial ;
2. au cours de la construction de l'arbre, la dichotomie d'un nœud, qui vise à réduire au mieux l'entropie au sein de ce nœud, est optimisée sur un sous-ensemble restreint des variables explicatives (toujours le même nombre) tiré au hasard.

Par une simple moyenne pondérée dans le cas de la régression et un vote majoritaire dans le cas de la discrimination, la prédiction d'une forêt est calculée à partir de l'ensemble des arbres estimés pour chaque échantillon *bootstrap*. À chaque itération, le calcul d'une erreur *out of bag*, c'est-à-dire estimée sur les observations exclues du tirage *bootstrap*, permet de contrôler l'évolution de l'apprentissage.

Comme pour beaucoup de méthodes « boîte noire » (k plus proches voisins, réseau de neurones, *bagging*, *boosting*, SVM), une forêt d'arbres n'est pas un modèle explicite dans lequel il est facile d'interpréter le rôle des variables. La stratégie proposée par Breiman (2001) consiste alors à rechercher un indicateur d'importance de chacune des variables explicatives. Le principe est simple et efficace, même s'il requiert des calculs lourds : plus la qualité de la prédiction se dégrade quand on remplace les valeurs d'une variable par un réarrangement aléatoire de ses valeurs, plus cette variable est considérée comme importante. Un tel indicateur possède, bien entendu, des faiblesses. Prenons le cas de deux variables importantes et fortement corrélées ; la perturbation d'une de ces variables peut ne conduire qu'à une faible dégradation de la qualité de la prédiction tant que la présence de l'autre variable la supplée.

En résumé, une forêt aléatoire est un modèle qui approche des surfaces de discrimination non linéaires dans l'espace des individus, sans pour autant sur-ajuster l'échantillon d'apprentissage, tandis que l'indice d'importance souligne les variables, ici les gènes, qui, en moyenne sur l'ensemble des tirages *bootstrap*, contribuent le mieux à discriminer les classes de la variable qualitative à prévoir.

Dans le cas élémentaire de la discrimination des génotypes des souris, les gènes qui apparaissent les plus significatifs sont, par ordre décroissant : `PMDCI`, `CAR1`, `THIOL`, `L.FABP`, `ALDH3`, `CYP3A11`, `PECI`, `GK`, `CYP4A10`, `ACBP`, `FAS`, `CPT2`, `BSEP`, `mHMGCoAS`, `ACOTH`. Avec ceux-ci, la prédiction des génotypes est presque sûre avec une estimation (*out of bag*) de l'erreur de prédiction de 2 %.

En revanche, la discrimination des régimes, beaucoup plus délicate, a été traitée conditionnellement au génotype. Le régime `ref` est dans les deux cas le plus difficile à reconnaître.

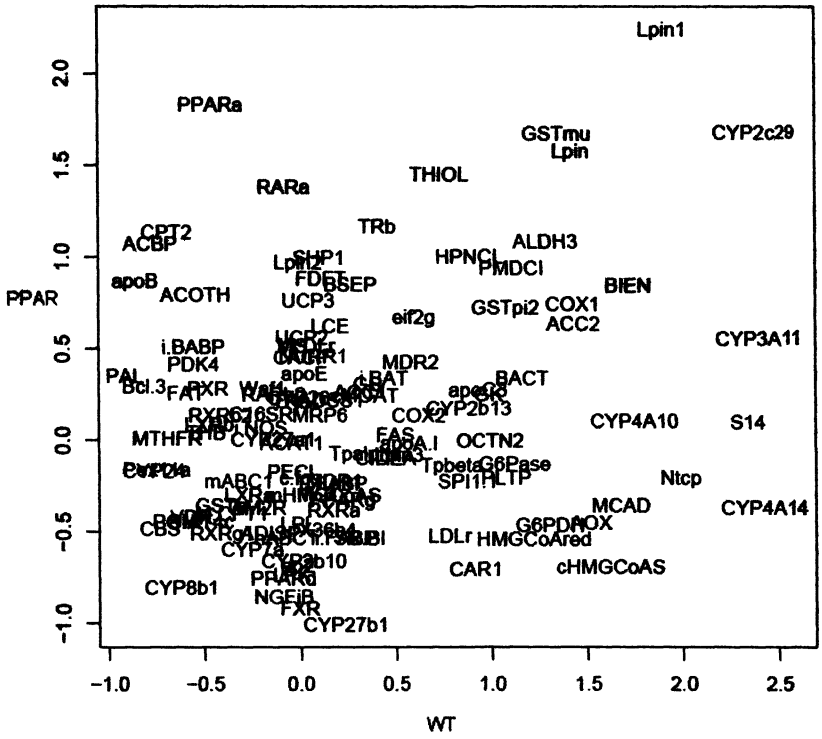

FIG 12. — Représentation des gènes en fonction de leur importance pour la discrimination des régimes à génotype fixé (`WT` sur l'axe horizontal et `PPAR` sur l'axe vertical)

La figure 12 représente les gènes en fonction de leur importance pour la discrimination des régimes pour chacun des génotypes. C'est pour les souris

PPAR que la discrimination des régimes est la plus difficile, ce qui confirme les observations déjà faites sur les résultats de la classification représentés par MDS (Fig. 9). Ce résultat s'interprète sur le plan biologique comme une implication du récepteur PPARα dans les régulations géniques provoquées par les régimes alimentaires.

5. Relations entre expressions des gènes et variables cliniques

Dans diverses applications biologiques, l'interprétation des données d'expression génique peut être avantageusement complétée par leur rapprochement avec des données mesurées par ailleurs sur les mêmes individus.

L'objectif général de l'analyse canonique (voir, par exemple, Mardia *et al.*, 1979) est d'explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d'individus. Nous l'avons mise en œuvre pour étudier les relations entre certains gènes et certains acides gras hépatiques de la souris.

Au préalable, une sélection des gènes et des acides gras a été nécessaire. Elle est décrite dans le point 5.1.

5.1. Sélections préliminaires

Les calculs effectués dans l'analyse canonique imposent de disposer d'un nombre de variables dans chaque groupe inférieur au nombre d'individus (40 dans notre exemple). Lors d'études préliminaires, nous avons de plus constaté que le nombre de variables devait être sensiblement inférieur au nombre d'individus. En effet, lorsque le nombre de variables est relativement important (disons supérieur à 25 dans notre cas), nous nous retrouvons avec plusieurs corrélations canoniques égales à 1 : la somme des dimensions des deux sous-espaces considérés dépassant la dimension de l'espace des variables, l'intersection de ces sous-espaces contient nécessairement des dimensions sans signification concrète qui correspondent aux plus grandes corrélations canoniques.

5.1.1 Sélection des gènes

Nous repassons en revue les différentes méthodes mises en œuvre dans les deux paragraphes précédents (3 et 4) afin d'extraire un groupe de gènes particulièrement intéressant vis-à-vis des facteurs expérimentaux. Cette synthèse est une façon indirecte mais, nous semble-t-il, efficace de répondre à la problématique de la détection de gènes différentiellement exprimés.

ACP.

Comme nous l'avons déjà évoqué, dans le cadre de l'ACP, il est possible de calculer la contribution des variables-gènes aux axes sélectionnés afin notamment d'alléger les représentations graphiques en ne représentant que les gènes qui contribuent le plus à ces axes. Dans une optique de sélection de

gènes, une représentation de la distribution de ces contributions permet dans notre exemple de détecter 12 gènes dont les contributions apparaissent comme *outlier* dans le diagramme en boîte (Fig. 13).

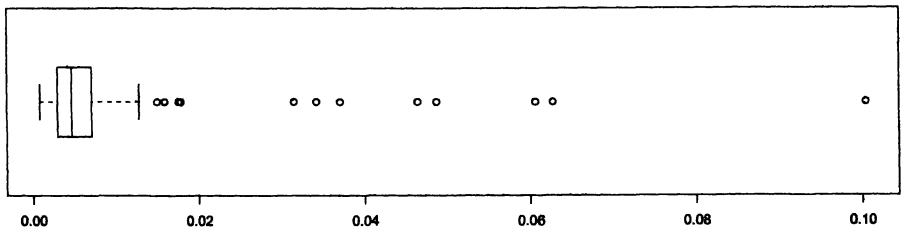

FIG 13. — Diagramme en boîte des contributions des gènes à la définition des trois premières composantes principales

Il s'agit des gènes suivants (dans l'ordre décroissant de leur contribution aux trois premiers axes) :

FAS, CYP4A14, CYP3A11, CYP4A10, THIOL, S14, PMDCI, L.FABP, mHMGCoAS, Lpin, Lpin1, GK.

Si on ne se limite pas à ces *outliers*, on peut étendre la liste aux 30 gènes contribuant le plus et ajouter ainsi :

PECI, CAR1, G6Pase, PAL, cHMGCoAS, BIEN, GSTpi2, ACC2, GSTmu, AOX, ALDH3, Lpin2, HPNCL, HMGCoAred, Tpbeta, TRa, SHP1, ACBP.

Classifications.

Pour sélectionner des gènes, nous nous basons essentiellement sur une interprétation visuelle de la carte induite par les classifications des lignes et des colonnes (Fig. 7). En effet, il nous semble pertinent de retenir les gènes pour lesquels l'aspect contrasté des colonnes est dominant, révélant ainsi une hétérogénéité des expressions géniques selon le génotype et le régime des individus-souris. Ces gènes sont également caractérisés par un regroupement « tardif » à l'ensemble des autres gènes au cours du déroulement de la classification hiérarchique ascendante, ce qui correspond à une expression sensiblement différente des autres. Ces 14 gènes sont :

PMDCI, THIOL, CYP3A11, CYP4A10, L.FABP, CYP4A14, FAS, Lpin, Lpin1, CYP2c29, GSTmu, GSTpi2, G6Pase, S14.

Contrairement aux autres méthodes, la liste des gènes sélectionnés est ici quasiment imposée par la lecture du graphique (Fig. 7). Il ne semblerait donc pas pertinent de vouloir étendre cette liste à des gènes supplémentaires.

Modélisation.

En procédant à une ANOVA selon le modèle (5) proposé dans le paragraphe 4, une façon de fournir une vision synthétique consiste à trier les gènes dans l'ordre croissant du minimum des 3 *p-values* associées aux doubles interactions (gène*génotype et gène*régime) et à l'interaction d'ordre 3 (gène*génotype*régime). Les 30 premiers gènes de ce classement sont :

PMDCI, THIOL, CYP3A11, CYP4A10, ALDH3, CAR1, HPNCL, L.FABP, mHMGCoAS, GK, PECI, CYP4A14, CPT2, BIEN, FAS, ACAT2, Lpin, Lpin1, Lpin2, CYP2c29, GSTmu, HMGCoAre, ACC2, ACBP, AOX, CACP, COX1, GSTpi2, PLTP, S14.

Forêts aléatoires.

Dans cette partie, nous n'utilisons que les résultats des forêts aléatoires à génotype fixé. En effet, presque tous les gènes mis en évidence par la discrimination des génotypes sont déjà signalés par les autres méthodes.

La sélection de gènes que nous proposons à partir de l'importance calculée par les forêts aléatoires consiste à extraire les gènes au-delà d'un arc de cercle centré en $(-1;-1)$ et de rayon approximatif 2.7 sur la figure 12. Cela correspond à sélectionner les gènes dont la racine carrée des sommes des carrés des importances est supérieure à un seuil. Par ce moyen, nous sélectionnons les 20 gènes suivants :

PMDCI, THIOL, CYP3A11, CYP4A10, ALDH3, HPNCL, CYP4A14, BIEN, Lpin, Lpin1, CYP2c29, GSTmu, COX1, ACC2, GSTpi2, Ntcp, S14, PPARa, RARa, TRb.

Synthèse.

La synthèse de la procédure de sélection de gènes selon les différentes méthodes est fournie dans le tableau 5. La première partie de ce tableau correspond aux 10 gènes sélectionnés par les quatre méthodes, la deuxième aux 7 gènes sélectionnés par trois méthodes sur quatre (à l'intérieur de chaque groupe, l'ordre est celui donné par la modélisation).

5.1.2 Sélection des acides gras

Sans reprendre l'intégralité du traitement appliqué aux gènes, nous nous sommes attachés à gommer les fortes corrélations entre acides gras pouvant pertuber l'analyse canonique. La démarche adoptée consiste à combiner les résultats d'une classification ascendante hiérarchique et d'une ACP.

Le dendrogramme de la figure 14 a été construit en utilisant la distance basée sur la corrélation (d_3) et le critère du saut moyen pour l'agglomération. Nous avons également procédé à l'agglomération selon le saut de Ward ; la classification ainsi produite entraîne les mêmes interprétations. Nous avons choisi ici le critère du saut moyen qui, associé à la distance basée sur la corrélation, permet d'interpréter simplement l'échelle verticale du dendrogramme (un moins la corrélation entre les différents lipides).

TABLEAU 5. – Synthèse de la sélection des gènes selon les différentes méthodes. Seuls les gènes détéctés par au moins trois méthodes sont présents dans le tableau. Le symbole * indique que le gène est détecté par la méthode de la colonne correspondante

Gène	ACP	Classification	Forêts aléatoires	ANOVA
PMDCI	*	*	*	*
THIOL	*	*	*	*
CYP3A11	*	*	*	*
CYP4A10	*	*	*	*
CYP4A14	*	*	*	*
Lpin	*	*	*	*
Lpin1	*	*	*	*
GSTmu	*	*	*	*
GSTpi2	*	*	*	*
S14	*	*	*	*
ALDH3	*		*	*
HPNCL	*		*	*
L.FABP	*	*		*
BIEN	*		*	*
FAS	*	*		*
CYP2c29		*	*	*
ACC2	*		*	*

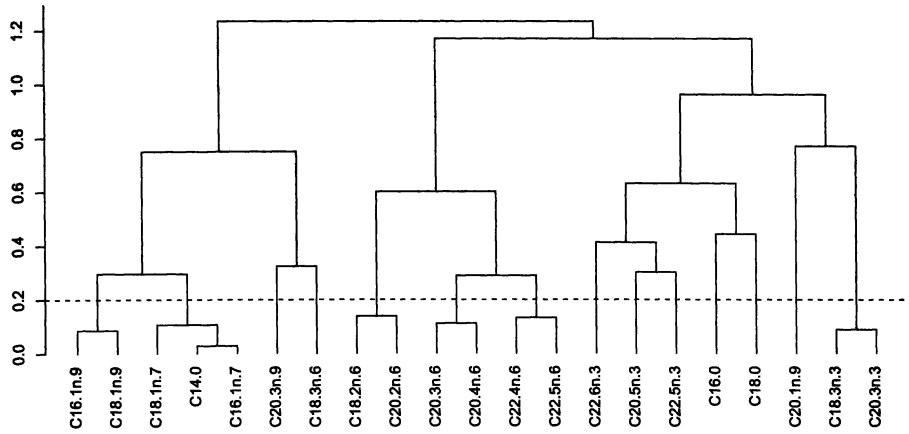

FIG 14. — Classification hiérarchique ascendante des acides gras selon d_3 (basée sur la corrélation) et le critère du saut moyen. Les gènes regroupés sous la ligne horizontale pointillée sont corrélés à plus de 0.8

De manière synthétique, on remarquera que cette réprésentation des acides gras tend à regrouper les principales familles : acides gras monoinsaturés à gauche, famille *Oméga 6* au centre et famille *Oméga 3* à droite. Cela est d'autant plus vrai si l'on fait abstraction de quelques acides gras qui atteignent la limite de détection de la méthode de dosage (chromatographie en phase gazeuse) dans plusieurs groupes de souris (notamment les `C20.3n-9` et `C18.3n-6`) et du `C20.1n-9` dont la proportion ne varie pas de manière sensible entre les groupes. Sur le plan biologique, cette représentation reflète bien les connaissances actuelles de la voie de biosynthèse des acides gras polyinsaturés (Nakamura & Nara, 2004).

Coupé au niveau 0.2 pour signifier que les gènes regroupés sous cette limite sont corrélés à plus de 0.8, l'arbre de la figure 14 fournit une classification en 14 groupes. Parmi ces groupes, nous éliminons les acides gras `C20.3n-9`, `C18.3n-6` et `C20.1n-9` pour les raisons indiquées ci-dessus. Le principe de notre sélection consiste à ne conserver qu'un seul acide gras pour chacun des 11 groupes restants.

Pour effectuer ce choix, nous nous basons sur les résultats de l'ACP des acides gras.

TABLEAU 6. – Parts de variance expliquée et cumuls pour les quatre premières dimensions de l'ACP des acides gras

	Axe 1	Axe 2	Axe 3	Axe 4
Parts de variance expliquée	0.41	0.29	0.17	0.09
Cumuls	0.41	0.70	0.86	0.96

Le tableau des valeurs propres (Tab. 6) nous conduit à conserver quatre dimensions à partir desquelles ont été calculées les contributions (voir équation (1)) des acides gras que nous avons classées ci-dessous par ordre décroissant :

$$\texttt{C18.2n}-6, \texttt{C18.1n}-9, \texttt{C18.3n}-3, \texttt{C22.6n}-3, \texttt{C20.4n}-6, \texttt{C16.0},$$
$$\texttt{C18.1n}-7, \texttt{C16.1n}-7, \texttt{C18.0}, \texttt{C20.5n}-3, \texttt{C22.5n}-3, \texttt{C14.0}, \texttt{C22.5n}-6,$$
$$\texttt{C20.3n}-9, \texttt{C20.3n}-6, \texttt{C18.3n}-6, \texttt{C16.1n}-9, \texttt{C22.4n}-6, \texttt{C20.2n}-6,$$
$$\texttt{C20.3n}-3, \texttt{C20.1n}-9.$$

Le représentant de chacun des 11 groupes est défini comme celui contribuant le plus aux quatre premiers axes de l'ACP à l'intérieur de son groupe.

Nous retenons finalement les 11 acides gras suivants :

$$\texttt{C18.1n}-9, \texttt{C18.1n}-7, \texttt{C18.2n}-6, \texttt{C20.4n}-6, \texttt{C22.5n}-6, \texttt{C22.6n}-3,$$
$$\texttt{C20.5n}-3, \texttt{C22.5n}-3, \texttt{C16.0}, \texttt{C18.0}, \texttt{C18.3n}-3.$$

Afin de contrôler l'effet des deux facteurs génotype et régime, globalement sur les dix gènes choisis, ainsi que sur les onze acides gras sélectionnés, nous avons réalisé deux analyses de variance multidimensionnelles (MANOVA), à dix et onze dimensions respectivement. Dans chaque cas, on a obtenu des effets extrêmement significatifs, que ce soit pour le génotype, pour le régime ou pour les interactions (*p-values* inférieures à 10^{-4} pour le test de Wilks).

5.2. Mise en œuvre de l'analyse canonique

Nous avons mis en œuvre deux analyses canoniques (AC) avec les onze acides gras sélectionnés précédemment : l'une avec les dix gènes retenus par les quatre méthodes de sélection, l'autre avec les dix-sept gènes retenus par au moins trois méthodes.

Dans la représentation graphique des variables, nous avons ajouté deux cercles : le cercle des corrélations, de rayon unitaire, justifié par la représentation des variables à partir des corrélations variables-facteurs; le cercle de rayon 0.5, pour faciliter la lecture en mettant en évidence les phénomènes les plus intéressants dans la couronne ainsi définie.

Dans la première analyse, les 10 corrélations canoniques sont les suivantes :

0.96 0.93 0.91 0.86 0.79 0.72 0.61 0.41 0.25 0.04

Les figures 15 et 16 représentent variables et individus dans les trois premières dimensions pour lesquelles les corrélations canoniques sont supérieures à 0.90.

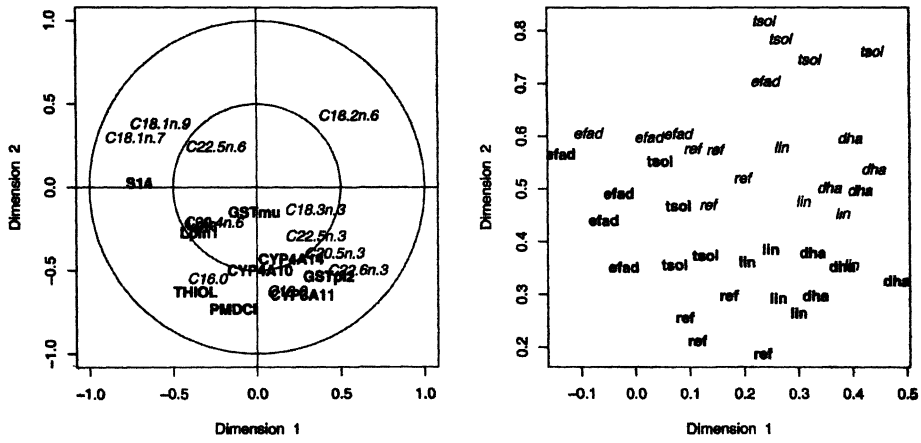

FIG 15. — Représentation des variables (10 gènes en gras, 11 acides gras en italique), à gauche et des individus (WT en gras et PPAR en italique), à droite, dans le premier plan canonique

Le premier plan de cette analyse canonique (Fig. 15) permet de résumer de manière originale les principaux résultats de cette étude. Dans le graphique des individus, on observe la séparation des deux génotypes, l'opposition entre le régime `efad` (faibles coordonnées sur l'axe 1) et les régimes riches en *Oméga 3*, `lin` et `dha` (fortes coordonnées sur l'axe 1), enfin la position particulière du régime `tsol` chez les souris PPAR. On retrouve, sur le graphique des variables, les regroupements de familles d'acides gras évoqués au paragraphe 5.1.2. La séparation des génotypes est principalement liée à l'accumulation préférentielle du `C18.2n-6` chez les souris PPAR au détriment du `C16.0`, du `C18.0` et des acides gras longs polyinsaturés `C20.5n-3` et `C22.6n-3` (DHA), ainsi

qu'à la plus forte expression des gènes THIOL, PMDCI, CYP3A11 et GSTpi2 chez les souris WT par rapport aux souris PPAR. On note les proximités entre le C16.0 et le gène THIOL, ainsi que les proximités entre CYP3A11 et GSTpi2 et les acides gras C18.0 et C22.6n-3. L'opposition entre le régime efad et les régimes lin et dha est liée aux particularités suivantes du régime efad : accumulation d'acides gras monoinsaturés (C18.1n-9 et C18.1n-7) chez les souris des deux génotypes (mais plus marquée chez les souris PPAR) et surexpression du gène S14 presque exclusivement chez les souris WT (d'où un décalage entre les acides gras monoinsaturés et S14 le long de l'axe 2). Sous régime riche en *Oméga 3* (lin et dha), on observe une accumulation préférentielle des acides gras C20.5n-3 (surtout pour le régime lin), C22.6n-3 (surtout pour le régime dha) et C18.0 accompagnée de régulations positives des gènes GSTpi2, CYP3A11 et des CYP4A qui, cependant, se révèlent moins marquées, voire absentes, chez les souris PPAR. Enfin, la position particulière du régime tsol chez les souris PPAR est liée à l'accumulation extrêmement marquée de C18.2n-6 dans le foie de ces souris sous le régime tsol (sous ce régime, la proportion de C18.2n-6 est presque deux fois plus importante chez les souris PPAR que chez les souris WT), soulignant ainsi le rôle primordial de PPARα dans la prise en charge de cet acide gras, que ce soit pour sa dégradation ou pour son utilisation dans la biosynthèse des acides gras longs polyinsaturés de la famille *Oméga 6*. Le troisième axe de cette analyse canonique (Fig. 16) n'apporte qu'une information vraiment nouvelle par rapport aux deux premiers axes, c'est la régulation positive du gène GSTmu chez les souris des deux génotypes (à nouveau, elle est moins marquée chez les souris PPAR) soumises au régime dha. Cette régulation n'est en revanche pas observée sous le régime lin.

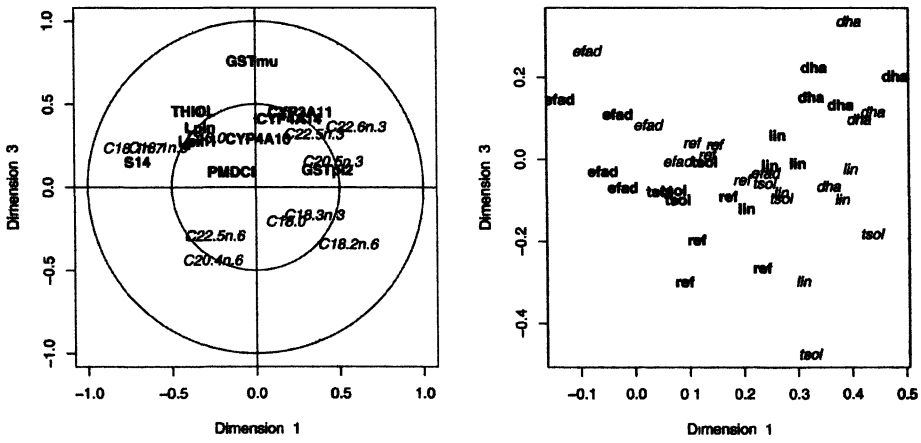

FIG 16. — Représentation de l'AC des 10 gènes et 11 acides gras dans le plan canonique 1-3 avec les mêmes conventions que pour la figure 15

Dans la seconde analyse canonique, les 11 corrélations sont :

0.98 0.98 0.94 0.93 0.87 0.83 0.76 0.72 0.68 0.53 0.44

L'information pertinente se concentre sur les quatre premières dimensions. En particulier, sur les dimensions cinq et six, toutes les variables sont projetées à l'intérieur du petit cercle.

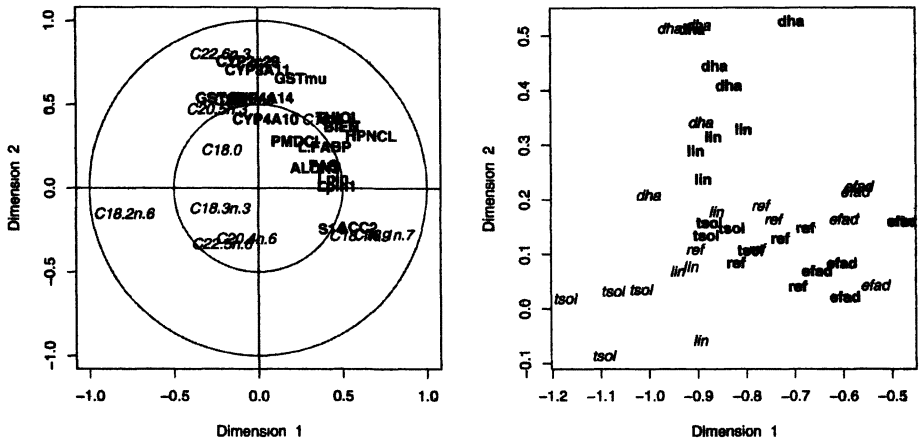

FIG 17. — Représentation de l'AC des 17 gènes et 11 acides gras sur le premier plan canonique avec les mêmes conventions que pour la figure 15

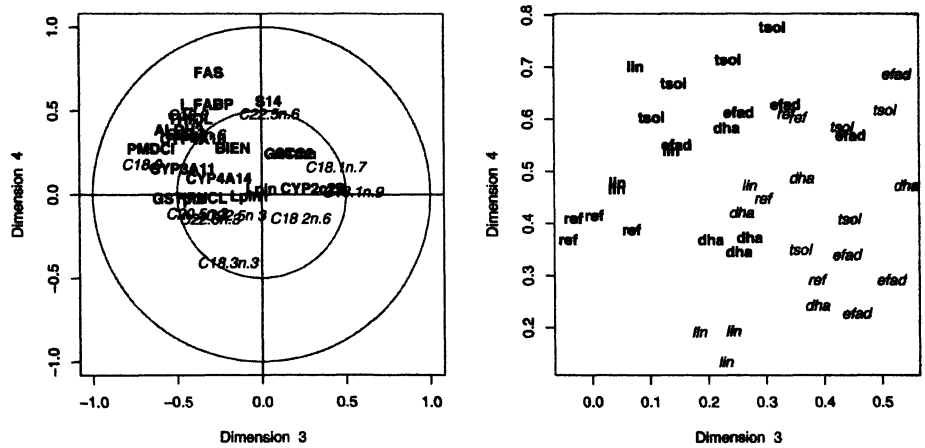

FIG 18. — Représentation de l'AC des 17 gènes et 11 lipides sur le plan 3-4 avec les mêmes conventions que pour la figure 15

Cette deuxième analyse canonique (Fig. 17 et 18) nous apporte des informations complémentaires grâce à l'ajout de plusieurs gènes qui se rapprochent de groupes de gènes et acides gras déjà évoqués lors de l'étude de la première analyse. Ainsi, dans le plan 1-2 (Fig. 17), on note que `CYP2c29` se comporte de manière assez similaire à `CYP3A11` et `GSTmu` et, dans une moindre mesure, à `GSTpi2` et `CYP4A14`. Ces gènes codant des cytochromes `P450` et des glutathion-S-transférases sont impliqués de manière globale dans les

phénomènes de détoxication. Leurs niveaux d'expression semblent reliés dans cette étude aux proportions d'acides gras long polyinsaturés de la famille *Oméga 3* comme le `C22.6n-3` (DHA), le `C22.5n-3` et le `C20.5n-3` (acide eicosapentanoïque, EPA). Comme nous l'avons évoqué, les régulations de ces gènes par les régimes alimentaires `dha` et `lin` sont généralement plus marquées chez les souris `WT` que chez les souris `PPAR`. Il est particulièrement rassurant de constater que plusieurs auteurs ont déjà observé les régulations de certains de ces gènes par ces mêmes acides gras (Berger *et al.*, 2002; Chen *et al.*, 2001) et que ces acides gras sont connus pour être des activateurs du récepteur PPARα (Forman *et al.*, 1997; Murakami *et al.*, 1999). De manière similaire, les gènes `BIEN` et `HPNCL` se rapprochent de `THIOL` et du `C16.0`. Ces trois gènes sont impliqués dans le catabolisme des acides gras par α- ou β-oxydation. En revanche, leur corrélation avec la proportion de `C16.0` constitue une donnée nouvelle qui nécessitera des expériences complémentaires pour être complètement comprise. Enfin, le gène `ACC2` est proche du gène `S14` et tous deux sont impliqués dans la lipogénèse. Leurs niveaux d'expression ainsi que l'abondance des acides gras monoinsaturés `C18.1n-7` et `C18.1n-9` sont maximaux avec le régime `efad`. Le plan 3-4 (Fig. 18) nous permet principalement de compléter la liste des gènes présentant des expressions plus fortes chez les souris `WT` par rapport aux souris `PPAR` en ajoutant les gènes `FAS`, `L.FABP` et `ALDH3`.

Ainsi, l'analyse canonique fournit au biologiste un outil intéressant pour mieux appréhender de manière conjointe deux jeux de variables sélectionnées. Elle offre des représentations graphiques facilement lisibles dans lesquelles il est possible de puiser des informations pertinentes. Il est important de souligner que l'analyse canonique intervient presque nécessairement après des étapes d'exploration des données puis de sélection de variables. Ainsi, lorsqu'elle est mise en œuvre, le biologiste dispose déjà d'une bonne connaissance de ses deux jeux de variables, facilitant d'autant plus l'interprétation des graphiques. Enfin, notons que l'analyse canonique met en évidence des corrélations qui ne peuvent donc être directement interprétées comme des phénomènes de cause à effet. Dans nos exemples, il nous est impossible de distinguer les effets des variations d'expression génique sur les proportions d'acides gras, les effets des proportions d'acides gras sur l'expression des gènes ou encore la simple concomittance de variations d'expression génique et de proportions d'acides gras. En revanche, nous avons souligné que des données publiées permettaient de conforter certaines proximité de gènes et d'acides gras et que d'autres soulevaient des questions nouvelles nécessitant la mise en œuvre d'expériences complémentaires.

6. Conclusion et perspectives

Cette étude, qui se veut essentiellement didactique, permet de mettre en avant le fait qu'il n'existe pas une méthode unique permettant de traiter des données d'expression. La question « Quelle méthode dois-je utiliser pour traiter mes données d'expression ? » n'a pas de sens de façon générale. En revanche, à une question précise du type « Puis-je effectuer une partition des gènes ? »,

une méthode statistique (ici la classification) peut apporter des éléments de réponses par des sorties numériques et/ou graphiques. Toutefois, la réponse précise à la question ne peut être apportée que par une collaboration étroite entre le statisticien, pour son expertise des méthodes utilisées, et le praticien, pour son expertise des phénomènes biologiques considérés. Soulignons encore une fois l'importance du choix de la méthode et de celui des options (qui fournissent différentes optiques). Ces choix ne sont pas neutres sur le plan biologique et doivent être pris en compte dans l'interprétation des graphiques et des modèles obtenus selon chaque optique.

Dans cet article, nous avons surtout essayé de répondre aux questions posées par l'expérience biologique mise en œuvre. De nombreux autres problèmes soulevés par l'analyse des données transcriptomiques n'ont pas été abordés. Signalons, pour mémoire, le *design* des oligos, la mise en œuvre de plans d'expériences pour le recueil des données, la normalisation, d'autres gestions des tests multiples, l'étude de cinétiques d'expression... Sur ce dernier point, on trouvera quelques pistes sur la classification de profils temporels d'expression ainsi que des ouvertures vers l'ANOVA fonctionnelle dans Baccini *et al.* (2005).

Dans le cas d'une puce comportant plusieurs milliers de gènes (en particulier une puce pangénomique), les méthodes présentées, notamment les méthodes exploratoires, peuvent s'appliquer de manière analogue, même si leur interprétation est alors plus délicate. En revanche, l'analyse canonique nous semble plus spécifique au cas d'un nombre restreint de gènes.

La démarche adoptée dans ce travail n'a pas été de rechercher la « meilleure » méthode pour détecter les gènes différentiellement exprimés, mais de voir comment chaque approche renseigne sur la problématique biologique. Une synthèse des différentes approches est plus susceptible de faire converger un faisceau de présomptions sur un ensemble de gènes, avec l'idée qu'elle permet d'éviter des artefacts (faux positifs) contrairement à une approche unique.

Remerciements : Les auteurs souhaitent remercier Thierry Pineau et Hervé Guillou pour leur participation active à l'acquisition des données et pour les discussions scientifiques concernant leur interprétation. D'autre part, ils remercient les relecteurs qui ont permis de bien améliorer la qualité de l'article.

Cette recherche a été partiellement financée par l'ACI IMPBio *Développement d'un environnement dédié à l'analyse statistique des données d'expression* (ENV-STAT-EXP).

Compléments

Diverses informations relatives aux activités des auteurs sur la thématique des biopuces sont accessibles en ligne à l'adresse

`www.lsp.ups – tlse.fr/Biopuces`

En particulier, on y trouvera les commandes R ayant permis de mettre en œuvre les analyses présentées dans cet article ainsi qu'un livret regroupant ces mêmes figures en couleur.

Références

AKAIKE H. (1974). A new look at the statistical model identification, *IEEE Transaction on Automatic control*, AC-19, 716-723.

AOYAMA T., PETERS J.M., IRITANI N., NAKAJIMA T., FURIHATA K., HASHIMOTO T., GONZALEZ F.J. (1998). Altered constitutive expression of fatty acid-metabolizing enzymes in mice lacking the peroxisome proliferator-activated receptor alpha (PPARalpha), *Journal of Biological Chemistry*, 273(10), 5678-84.

AUBERT J., BAR-HEN A., DAUDIN J.-J., ROBIN S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR, *BMC Bioinformatics*, 5, 125.

BACCINI A., BESSE Ph., DÉJEAN S., MARTIN, P., ROBERT-GRANIÉ C., SAN CRISTOBAL M. (2005). *Analyse statistique des données d'expression génomique*, support de formation disponible en ligne, www.lsp.ups-tlse.fr/Biopuces.

BACCINI A., BESSE Ph., DÉJEAN S., MARTIN, P., ROBERT-GRANIÉ C., SAN CRISTOBAL M. (2005). Étude de données cinétiques issues de biopuces, XXXVII^{èmes} *Journées de Statistique*, juin 2005, Pau.

BAR-HEN A., ROBIN S. (2003). An iterative procedure for differential analysis of gene expression, *Comptes rendus de l'Académie des sciences*, Série I, 337, 343-346.

BENJAMINI Y., HOCHBERG Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society - Series B*, 85, 289-300.

BERGER A., MUTCH D.M., GERMAN J.B., ROBERTS M.A. (2002). Dietary effects of arachidonate-rich fungal oil and fish oil on murine hepatic and hippocampal gene expression, *Lipids in Health and Disease*, 1 :2.

BLAND J., ALTMAN D. (1995). Multiple significance tests : the Bonferroni method, *British medical Journal*, 310, 170.

BONNET A., BENNE F., DANTEC C., GOBERT N., FRAPPART P.O., SAN CRISTOBAL M., HATEY F., TOSSER-KLOPP G. (2004). Identification of genes and gene networks involved in pig ovarian follicular development, by using c-DNA microarrays, *XIII International Workshop on the Development and Function of Reproductive organs*, 12-15 July 2004, Copenhagen, Denmark.

BREIMAN L. (2001). Random forest, *Machine Learning*, 45, 5-32.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. (1984). *Classification and regression trees*, Wadsworth & Brooks, Cole Advanced Books & Software.

BROËT P., LEWIN A., RICHARDSON S., DALMASSO C., MAGDELENAT H. (2005). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments, *Bioinformatics*, 20(16), 2562-2571.

CHEN H.W., YANG J.J., TSAI C.W., WU J.J., SHEEN L.Y., OU C.C., LII C.K. (2001). Dietary fat and garlic oil independently regulate hepatic cytochrome p(450) 2B1 and the placental form of glutathione S-transferase expression in rats, *The Journal of Nutrition*, 131(5), 1438-43.

DALMASSO C., BROËT P., MOREAU T. (2005). A simple procedure for estimating the false discovery rate, *Bioinformatics*, 21(5), 660-668.

DELMAR P., ROBIN S., TRONIK-LEROUX D., DAUDIN J.-J. (2005). Mixture model on the variance for the differential analysis of gene expression data, *Journal of the Royal Statistical Society - Series C, Applied Statistics*, 54, 1-20.

DRAGHICI S. (2003). *Data Analysis Tools for DNA Microarrays*, Mathematical Biology and Medicine Series, Chapman & Hall/CRC.

DUDOIT S., POPPER J., BOLDRICK J.C. (2003), Multiple Hypothesis Testing in Microarray Experiments, *Statistical Science*, 18(1), 71-103.

DUDOIT S., YANG Y., SPEED T., CALLOW M. (2002), Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, 12(1), 111-139.

EISEN M.B., SPELLMAN P.T., BROWN P.O., BOTSTEIN D. (1998), Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the USA*, 95(25), 14863-14868.

FORMAN B.M., CHEN J., EVANS R.M. (1997). Hypolipidemic drugs, polyunsaturated fatty acids, and eicosanoids are ligands for peroxisome proliferator-activated receptors alpha and delta, *Proceedings of the National Academy of Sciences of the USA*, 94(9), 4312-7.

FOULLEY J.-L., JAFFREZIC F., ROBERT-GRANIÉ C. (2000), EM-REML estimation of cavariances parameters in Gaussian mixed models for longitudinal data analysis, *Genetics Selection Evolution*, 32, 129-141.

KERR K., MARTIN M., CHURCHILL G. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, 7, 819-837.

LEBART L., MORINEAU A., PIRON M. (1995). *Statistique exploratoire multidimensionnelle*, Dunod.

LEE S.S., PINEAU T., DRAGO J., LEE E.J., OWENS J.W., KROETZ D.L., FERNANDEZ-SALGUERO P.M., WESTPHAL H., GONZALEZ F.J. (1995). Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators, *Molecular and Cellular Biology* 15(6), 3012-22.

MARDIA K., KENT J., BIBBY J. (1979). *Multivariate analysis*, Academic Press.

MARTIN P.G.P., LASSERRE F., CALLEJA C., VAN ES A., ROULET A., CONCORDET D., CANTIELLO M., BARNOUIN R., GAUTHIER B., PINEAU T. (2005a). Transcriptional modulations by RXR agonists are only partially subordinated to PPARalpha signaling and attest additional, organ-specific, molecular crosstalks, *Gene Expression*. Voir également le matériel supplémentaire associé à l'article à l'adresse :
http ://www.inra.fr/Internet/Centres/toulouse/pharmacologie/pharmaco-moleculaire/valorisations

MARTIN P.G.P. *et al.* (2005b). A nutrigenomic approach in mice reveals new aspects of PPARα-deficient phenotype with important implications in pharmacology, en préparation.

MCLACHLAN G.J., DO K.-A., AMBROISE C. (2004). *Analysing microarray gene expression data*, Wiley.

MURAKAMI K., IDE T., SUZUKI M., MOCHIZUKI T., KADOWAKI T. (1999). Evidence for direct binding of fatty acids and eicosanoids to human peroxisome proliferators-activated receptor alpha, *Biochemical and Biophysical Research Communications*, 260(3), 609-13.

NAKAMURA M.T., NARA T.Y. (2004). Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annual Review of Nutrition*, 24, 345-76.

NGUYEN D.V. (2004). On estimating the proportion of true null hypothesis for false discovery rate controlling procedures in exploratory DNA microarrays studies. *Computational Statistics and Data Analysis*, 47, 611-637.

PETERFY M, PHAN J, XU P, REUE K. (2001). Lipodystrophy in the fld mouse results from mutation of a new gene encoding a nuclear protein, lipin, *Nature Genetics*, 27(1), 121-4.

ROBERT-GRANIÉ C., BONAITI B., BOICHARD D., BARBAT A. (1999). Accounting for variance heterogeneity in French dairy cattle genetic evaluation, *Livestock Production Science*, 60, 343-357.

SAN CRISTOBAL M., ROBERT-GRANIÉ C., FOULLEY J-L. (2002). Hétéroscédasticité et modèles linéaires mixtes : théorie et applications en génétique quantitative, *Journal de la Société Française de Statistique*, 143, 1-2.

SAPORTA G. (1990). *Probabilités analyse des données et statistique*, Paris : Éditions Technip.

SCHWARZ G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.

SPEED T. (2003). *Statistical Analysis of Gene Expression Microarray Data*, Interdisciplinary Statistics, Chapman & Hall/CRC.