Université Paris Sud 11

Approche méthodologique pour l'intégration de données de neuroimagerie et de génétique

Christophe Lalanne

Master 2 Bioinformatique et Biostatistiques

2008–2009

Remerciements

Ce travail de recherche a été effectué sous la direction de M. Édouard Duchesnay.

J'en profite pour le remercier vivement de ses précieux conseils, ainsi que M. Jean-Baptiste Poline qui a relu une partie du manuscript et suggéré certaines analyses complémentaires. Enfin, j'adresse tous mes remerciements à l'équipe de Neurospin (I^2BM, CEA) avec qui j'ai eu l'occasion d'échanger, notamment Édith Lefloch, Vincent Frouin et Bertrand Thirion, ainsi qu'Arthur Tenenhaus (Supelec).

Table des matières

Table des figures

Liste des tableaux

Résumé

La possibilité d'acquérir à la fois des données de génotypage et des mesures de l'activité cérébrales grâce à l'imagerie fonctionnelle offre de nouvelles perspectives pour une meilleure compréhension des réponses comportementales observables chez le sujet sain et des facteurs de risque analysables chez le sujet porteur d'une maladie spécifique.

Nous nous intéressons dans ce travail à la possibilité de recourir aux techniques d'analyses multivariées, telle que la régression PLS et l'analyse canonique des corrélations, pour mettre en évidence des associations spécifiques entre les données recueillies en neuroimagerie chez des sujets sains participant à de tâches de psycholinguistique et leur génome (polymorphismes de séquence, SNP). Nos résultats suggèrent que ces techniques, éprouvées au travers d'une validation croisée, permettent effectivement d'isoler des SNP constituant de bons prédicteurs du degré d'asymétrie du signal BOLD, bien qu'elles se révèlent moins efficaces que dans le cadre des données de trancriptomiques étudiées jusqu'à présent. Le recours à des techniques telles que la régression pénalisée de type LASSO ou les forêts aléatoires, en considérant une variable réponse composite, permet également d'arriver à des conclusions favorables à l'existence d'un lien entre neuroimagerie et génétique.

Chapitre 1

Introduction

Le thème de ce mémoire repose sur l'idée que les variations génétiques puissent expliquer les modulations des activations cérébrales, au niveau individuel. Bien que cette idée ne soit pas nouvelle, les études dans ce sens demeurent assez récentes et la méthodologie nécessaire à ces études n'est pas encore établie.

Les études précédemment réalisées dans le domaine transcriptomique (Lê Cao et al., 2008) ou génomique (Parkhomenko et al., 2007) suggèrent que les analyses multivariées de type régression PLS ou analyse canonique des corrélations sont de bons candidats pour ce genre d'approche.

Ce travail constitue ainsi une première approche de l'étude des liens entre neuroimagerie et génétique et adresse plus généralement deux problèmes fondamentaux en biostatistiques :

– la sélection de variables : comment sélectionner le meilleur sous-ensemble d'attributs parmi un ensemble de prédicteurs potentiels ?
– la sélection de modèles : comment sélectionner le meilleur modèle, tant sur le plan de la précision de ses estimations que de ses capacités de prédiction et de généralisabilité ?

Ces problèmes ne sont pas nouveaux mais, dans le cas qui nous intéresse, ils prennent un sens tout particulier dans la mesure où l'on travaille avec des données de grande dimension et où plusieurs types de modèle peuvent être mis en compétition.

1.1 Génotype, phénotype et environnement

L'idée que les variations génétiques puissent expliquer les modulations des activations cérébrales, au niveau individuel, n'est pas nouvelle mais les études dans ce sens demeurent assez récentes (e.g., Glahn et al., 2007; Myers et al., 2007). Sans entrer dans les discussions relatives à la relation de causalité elle-même (les mutations génétiques sont-elles la source directe des modulations observées au niveau de l'activité cérébrale, ou bien existe-t-il un phénomène de médiation ou d'interaction, ou encore les activités cérébrales sont-elles à mêmes de réguler l'expression de certains gènes ?), on notera tout de même que dans la plupart de ces approches intégratives, on néglige souvent les phénomènes d'aggrégation familiale ou d'hétérogénité de population pour ne pas alourdir la complexité des modèles.

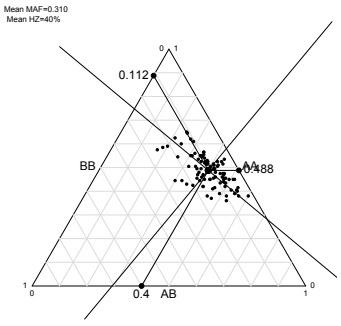
Le cadre de l'étude rejoint celui des études d'association entre un trait, ici quantitatif (mesures en neuroimagerie), et des données de génotypage. En particulier, on s'intéresse aux polymorphismes de séquence (Single nucleotide polymorphism, SNP) qui sont des variations génétiques au niveau d'un locus de la séquence. Ces polymorphismes se retrouvent tous les 600–1000 bp sur le génome humain et peuvent être considérés comme des biomarqueurs bialléliques (Guerra and Yu, 2006). Leur fréquence dans la population est au minimum de 1 % et ils font l'objet du vaste programme d'études HapMap (The International HapMap, Consortium, 2003, `www.hapmap.org`). Ils constituent une signature potentielle du trait d'intérêt (e.g. maladie ou phénotype particulier). On fait généralement l'hypothèse que certains de ces polymorphismes sont fonctionnels, c'est-à-dire qu'ils influencent directement ou indirectement le trait étudié. L'encadré ci-dessous illustre l'approche généralement retenue pour l'étude de l'effet des mutations géniques sur un trait quantitatif.

Approche explicative (régression) pour les SNP

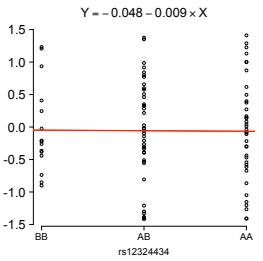
On considère un trait quantitatif Y et un SNP candidat dont les allèles sont notés A et B. Supposons par ailleurs que l'équilibre de Hardy-Weinberg est vérifié (p. ex., tableau ci-dessous) et notons la proportion des allèles par p_A et $p_B = 1 - p_A$, respectivement. Il existe trois génotypes possibles : AA, AB et BB, l'allèle le moins fréquent étant nommé « allèle variant ».

rs17749840 \ rs11660741	BB	AB	AA
BB	1	3	0
AB	7	23	5
AA	8	18	29

Fréquences de l'allèle mineur (MAF) :
0.228 (rs17749840), 0.404 (rs11660741)
Test de l'équilibre de Hardy-Weinberg :
$\chi^2 = 0.288$, $p = 0.576$ (rs17749840),
$\chi^2 = 0.075$, $p = 0.832$ (rs11660741)



Génotypage de 200 SNP



Codage quantitatif (MAF)

Les valeurs génotypiques moyennes, μ_1, μ_2 et μ_3, représentent la valeur moyenne de Y estimée sur chaque de l'un des génotypes (figure de droite ci-dessus). On supposera que la variance résiduelle est homogène entre ces trois populations. On définit les effet génétique additif α et dominant δ comme

$$\alpha = \frac{1}{2}(\mu_1 - \mu_3), \quad \text{et} \quad \delta = \frac{1}{2}(2\mu_2 - \mu_1 - \mu_3).$$

L'interprétation de δ est relativement simple : cette valeur reflète la déviation moyenne des individus hétérozygotes (AB) par rapport à la moyenne des homozygotes (AA et BB). Si $\delta = 0$, il y a additivité complète au niveau du trait mesuré (Liu, 1997, p. 377–379).

C'est dans ce contexte que s'inscrivent les approches de type gène candidat à large spectre (Genome-wide association studies, GWAS), à la différence près que l'on s'intéresse alors à un ensemble de SNP. Lorsqu'une région chromosomale contenant un gène caractéristique d'une maladie a été isolée, les études d'association permettent d'aller plus loin dans la caractérisation structurelle et fonctionnelle des réseaux biologiques impliqués dans la genèse et l'évolution de la maladie. Spécifiquement, trois grands domaines d'étude peuvent être distingués (Thomas, 2004) :

– localiser de manière plus précise la région dans laquelle le gène impliqué (mais non identifié) se situe en isolant d'autres marqueurs en déséquilibre de liaison avec celui-ci (*LD mapping*) ;

– tester l'hypothèse qu'un gène spécifique déjà identifié dans une région donnée est le bien le gène responsable de la maladie que l'on recherche en montrant qu'il comporte des polymorphismes associés spécifiquement à la maladie (*association studies*) ;
– estimer le facteur de risque associé à un gène connu et fonctionnel en fonction des divers facteurs susceptibles d'affecter sa pénétrance (*gene characterization*).

La collecte de ce type de données très volumineuses appelle des méthodes de réduction de données et des procédures de tests d'association spécifiques que nous introduisons dans les paragraphes suivants.

1.2 Approche méthodologique pour les données de grande dimension

1.2.1 Le cas $n \ll p$

La volonté de traiter et visualiser de gros volumes de données, particulièrement en génétique où l'on travaille le plus souvent avec un nombre de sujets largement inférieur au nombre de variables (cas $n \ll p$), nécessite soit d'adapter les méthodes statistiques usuelles, soit de réduire la dimensionnalité des données. Une façon d'aborder ce problème de dimensionnalité est de « réduire » le nombre de variables, soit en en sélectionnant un sous-ensemble représentatif, soit en dérivant de nouvelles variables qui résument au mieux l'information portée par l'ensemble des variables de départ.

Par exemple, la visualisation des gènes exprimés sur tout ou partie du génome humain est rendue possible grâce aux techniques de classification (analyse en clusters, dans la littérature anglo-saxonne) qui permettent de regrouper les variables, ici les gènes, selon une mesure de (dis)similarité particulière. Le choix de la mesure de similarité retenue par Eisen et al. (1998) est par exemple un score de similarité inter-gènes dans différentes conditions expérimentales qui reflète la corrélation entre leurs profils d'expression. L'utilisation d'une telle métrique dans l'espace des variables et d'une méthode d'aggrégation des classes entre elles permet alors de visualiser les sous-ensembles de gènes partageant le même profil d'expression sous forme d'un arbre de classification (dendrogramme). Des techniques d'élagage (*gene shaving*) permettent de raffiner cette classication et de ne conserver que de petits groupes de gènes fortement corrélés entre eux et maximisant la variance inter-individuelle (Hastie et al., 2000).

Outre ces approches de classification, les techniques factorielles, incluant l'analyse en composantes principales, permettent également de résumer l'ensemble des variables en une ou plusieurs combinaisons linéaires, orthogonales entre elles et qui maximisent l'information (pour un ouvrage récent sur le sujet, Izenman, 2008).

Toutefois, ces deux types d'approches ne permettent de traiter qu'un seul bloc de données à la fois. Dans le cas où le phénotype d'intérêt est constitué d'un ensemble de mesures, comme c'est le cas en neuroimagerie (que l'on travaille sur des régions d'intérêt ou sur l'ensemble de l'image 4D), il devient nécessaire de repenser les approches multivariées classiques pour prendre en compte le cas $n \ll p + q$, où p et q désignent la dimension de deux jeux de données.

1.2.2 Un problème de puissance

On retrouve également le problème classique des comparaisons multiples puisque l'on effectue des millions de tests dans le même espace probabilisé, c'est-à-dire sur le même critère d'intérêt, d'où la non-indépendance des résultats obtenus et l'inflation du risque de conclure à tort en faveur de l'existence d'un lien significatif. Cette multiplicité des tests est naturellement aggravée si l'on prend en considération des co-facteurs additionnels, comme le sexe, l'âge, la plateforme de génotypage, des mesures histologiques, ou d'autres caractéristiques portant soit sur les individus testés soit sur les mesures recueillies.

L'encadré page suivante résume les principaux moyens de contrôler les risques d'erreur, mais le lecteur intéressé trouvera dans Foulkes (2009, chap. 4) une bonne introduction aux procédures de tests multiples, et dans l'ouvrage de Dudoit and van der Laan (2008) un traitement plus approfondi de leurs applications dans le domaine génétique. On retiendra que les approches univariées permettent souvent d'opérer un filtrage des prédicteurs potentiellement intéressants. En revanche, comme le rappellent Ioannidis et al. (2009), les p-valeurs à considérer dans les études d'association doivent être corrigées pour la multiplicité des tests effectués sur tout le génome et des méta-analyses sont nécessaires pour confirmer les effets retrouvés dans des études de taille modérée (< 3000 sujets). Sans que des valeurs de référence soient clairement établies, des p-valeurs de 10^{-6} à 10^{-8} semblent répondre aux exigences attendues.

Les p-valeurs ne sont d'ailleurs pas les seuls critères utilisés pour évaluer le degré de significativité d'une association entre un trait mesuré et des données génomiques. Par exemple, la q-valeur, reposant sur le FDR et proposée dans le cadre des études d'association (Storey, 2003), est définie comme la borne supérieure du FDR positif, pFDR, où $\text{pFDR} = \mathbb{E}\left(\frac{V}{R} \mid R > 0\right)$ (on conditionne ici sur la présence d'au moins un résultat positif). Ce type de degré de significativité est très utilisé dans les analyses de puces à ADN (e.g., Tusher et al., 2001).

On remarquera en passant que dans le domaine de l'épidémiologie génétique, le risque de type I n'est sans doute pas le plus délicat, et que c'est bien le risque de seconde espèce (faux négatifs), et donc la puissance (le complémentaire de ce risque), qui apparaît critique. L'intérêt des procédures de validation croisée est de pouvoir se prononcer sur la capacité de prédiction et donc sur la puissance des tests utilisés.

Procédures de comparaisons multiples

Le tableau ci-dessous résume les différents cas de figure rencontrés dans le cas d'un simple test d'hypothèse (à gauche) et de tests inférentiels multiples (à droite) :

		Réalité	
		H_0	H_1
Test	H_0	$1-\alpha$ (TN)	β (FN)
	H_1	α (FP)	$1-\beta$ (TP)

		Réalité		
		H_0	H_1	
Test	H_0	U	T	$m-R$
	H_1	V	S	R
		m_0	$m-m_0$	m

Dans le cas d'un simple test, si l'on note H_0 l'hypothèse nulle reflétant l'absence d'effet, et H_1 l'hypothèse alternative (stochastique ou non), on retrouve 4 scénarios possibles : on rejette H_0 à tort alors qu'en réalité il n'y a pas d'effet, avec un risque α (souvent fixé à 5 %, *erreur de type I*), et on conclue correctement dans $100(1-\alpha)$ % des cas ; on ne rejette pas H_0 alors qu'en réalité il existe un effet, avec probabilité β (*erreur de type II*), et la puissance $(1-\beta)$ se définit comme la probabilité de conclure positivement lorsque l'effet existe réellement. Selon le schéma d'inférence proposé par Neyman et Pearson, on rejetera H_0 lorsque $P(s \geq s_c^\alpha \mid H_0) < \alpha$, c'est-à-dire lorsque la probabilité d'observer une statistique de test, s, au moins aussi extrême (ici dans un cas unilatéral) que la statistique de référence s_c sous H_0 est inférieure au risque consenti, α.
Par extension, dans le cas de m hypothèses, on se retrouve avec la situation de droite où l'ensemble des tests de $H_0^1, \ldots, H_0^m$ amène à V erreurs de type I et T erreurs de type II.
Globalement, il existe deux types de stratégies pour contrôler les risques d'erreur de type I (*faux positifs*) ou II (*faux négatifs*) :

– contrôler le risque d'erreur global (FWER), qui représente la probabilité de commettre au moins une erreur de type I (p. ex., Bonferroni, Scheffe ou Sidák),
– contrôler le taux de faux positifs (FDR), qui représente la proportion attendue d'hypothèses nulles correctement rejetées parmi l'ensemble des hypothèses nulles rejetées, R (p. ex., Benjamini & Hochberg ou Benjamini & Yekutieli).

De manière plus formelle, on peut écrire

$$\text{FWER} = P(V \geq 1) \quad \text{et} \quad \text{FDR} = \mathbb{E}\left(\frac{V}{R}\right)$$

Les procédures de test peuvent contrôler le FWER de manière stricte ou faible [a] mais nous ne rentrerons pas dans les détails pour ne pas alourdir l'exposé, de même que le FDR peut être exprimé de manière conditionnelle ou sous la forme d'un FDR positif. On retiendra que FDR $\leq$ FWER, puisque si les hypothèses nulles ne sont pas toutes vraies, de sorte que $V < R$, on a alors $V/R < 1$ et

$$\begin{aligned}\mathbb{E}\left(\frac{V}{R}\right) &= \big(V/R\big)P(V \geq 1) + \big(0/R\big)P(V = 0) \\ &= \big(V/R\big)P(V \geq 1) < P(V \geq 1)\end{aligned}$$

Par conséquent toute procédure contrôlant le FDR contrôle également le FWER.

a. Le FWER peut en fait être conditionné sur $H_0^c = [H_0^1, \ldots, H_0^m]$ (toutes les hypothèses nulles sont vraies) ou $H_0^{P_1} = [H_0^1, \ldots, H_0^k]$ (un sous-ensemble de k hypothèses nulles sont vraies) auquel cas on parlera, respectivement, de FWER sous nullité complète (FWEC) et de FWER sous nullité partielle (FWEP).

1.2.3 Un problème de sélection de variables

Les problèmes soulevés par le traitement de blocs de données de grande dimension sont au cœur des théories modernes de l'apprentissage supervisé. Ils prennent tout leur sens lorsque l'on travaille directement avec la séquence au lieu de considérer une centaine de gènes candidats (pour

une revue de questions, Liang and Kelemen, 2008).

Les méthodes de sélection de variables ou d'attributs (*feature selection*) peuvent se regrouper en trois grandes catégories (pour une revue, Guyon et al., 2006, chapitres 3 à 5) :

– les *méthodes de filtrage*, généralement univariées, permettent de sélectionner des variables indépendemment de l'algorithme de classification appliqué sur le résultat de cette sélection ; en ce sens, ces méthodes n'incorporent pas directement de procédure d'apprentissage mais peuvent être associées à des connaissances *a priori*, c'est-à-dire opérer dans un cadre bayésien (e.g., Bo and Jonassen, 2002; Long et al., 2001) ;
– les *méthodes d'ensemble* (*wrapper methods*), permettent d'effectuer la sélection à l'aide d'une procédure d'apprentissage, avec validation croisée, avant de soumettre itérativement les résultats de cette sélection à l'algorithme de classification lui-même ; la possibilité d'évaluer la qualité de la sélection (taux de classification correcte ou erreur de prédiction minimale) permet également d'ordonner les variables par pertinence (Inza et al., 2002; Guyon et al., 2002), sans toutefois qu'il soit possible d'incorporer des données additionnelles concernant la structure des fonctions de classification ou de régression (Lal et al., 2006) ;
– les *méthodes intégrées* ou *enchâssées* (*embedded methods*), dans lesquelles le processus de sélection de variables est intégré à l'algorithme d'apprentissage, sont aussi puissantes que les méthodes d'ensemble mais moins exigeantes en termes de calcul (Lal et al., 2006).

Les procédures de sélection de variables sont donc très diversifiées et répondent souvent à des finalités différentes, comme on peut voir dans la figure 1.1 : cherche-t-on à isoler des variables expliquant au mieux une association spécifique, sur l'échantillon observé, ou bien cherche-t-on à généraliser des associations plus complexes à une population de référence ? En tout état de cause, cette étape constitue un préalable indispensable dans l'analyse des données génétiques lorsque le phénotype d'intérêt n'est pas un simple trait, comme dans la plupart des études d'association de type GWAS, mais un ensemble de mesures.

1.2.4 Méthodes de régularisation et traitement de k-blocs

Le traitement des données de grande dimension pose donc deux problèmes essentiels : d'une part, le nombre de paramètres à estimer est largement supérieur au nombre d'observations disponibles (problème de dimensionnalité) entraînant un risque de sur-ajustement, et, d'autre part, la significativité des prédicteurs doit être modérée par le nombre de tests effectués. Ces deux problèmes ne sont pas équivalents. Il est donc nécessaire de :

1. réduire le nombre de prédicteurs en utilisant une méthode de régularisation, ou de pénalisation des prédicteurs peu informatifs ;
2. utiliser une procédure de validation croisée pour évaluer la qualité du modèle résultant et surtout sa généralisabilité (taux de classification ou erreur de prédiction).

En d'autres termes, on recherche la « meilleure » projection dans un espace de dimension réduit, le terme « meilleur » dépendant comme on va le voir du critère optimisé par l'algorithme. Chaque solution est construite sur un sous-ensemble des observations, et sa pertinence est évaluée sur les observations non utilisées pour construire le modèle. Le degré de significativité de chaque prédicteur peut être évalué en tenant compte de cette procédure de validation croisée (tests de

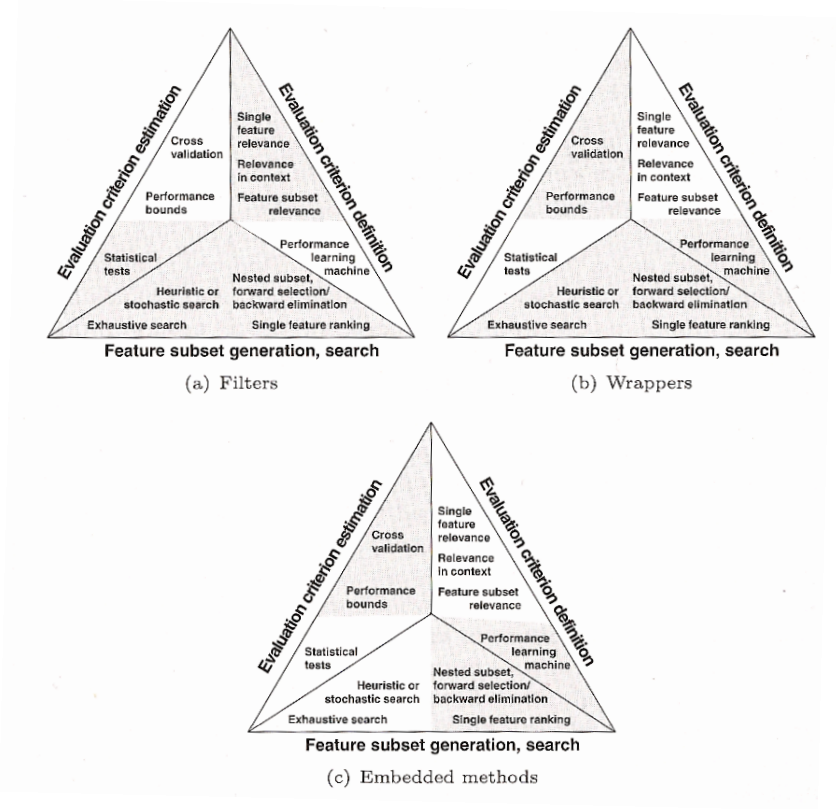


FIGURE 1.1 – Différentes approches pour la sélection de variables. (Tiré de Guyon et al. (2006, chap. 1, page 6))

significativité sur les coefficients ou mesure de l'importance de chaque variable dans le modèle). Les procédures de validation croisée sont assez nombreuses (e.g., Hastie et al., 2001, chap. 7), mais les plus utilisés reposent sur des techniques de rééchantillonnage telles que le bootstrap, le rééchantillonnage par blocs aléatoires (k-fold) ou par sortie individuelle (leave-one-out). Il existe un compromis entre la précision des estimations et la généralisabilité selon la méthode employée. Lorsque le nombre d'individu est petit (e.g. $n < 100$), il est préférable d'utiliser une procédure de type « leave-one-out » dans laquelle le modèle est évalué sur $n-1$ individus et testé sur l'individu exclu ; cette procédure est répétée n fois en excluant à chaque itération un nouvel individu.

Une revue de littérature des résultats publiés en 2004 sur les analyses de données d'expression en cancérologie effectuée par Dupuy and Simon (2007) révèle l'absence de prise en compte de la multiplicité des tests dans 39 % des cas, l'utilisation de critères de classification erronés dans 46 % des cas et enfin, des procédures de test biaisées par absence de validation croisée dans 43 % des cas.

Nous proposons dans les paragraphes suivants une brève description des techniques que nous avons retenues sur la base de différents critères : (1) la possibilité d'appliquer des méthodes de validation croisée, (2) la capacité de traiter des variables réponses numériques ou catégorielles, (3) la capacité de traiter un bloc de variables réponses, et (4) la possibilité de pénaliser ou régulariser les variables prédictrices et/ou réponses.

Régression pénalisée. Partant d'un modèle linéaire classique à p prédicteurs $x_1, \ldots, x_p$, la réponse prédite par un tel modèle prend la forme

$$g\big(\mathbb{E}(y)|X\big) = \hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j \tag{1.1}$$

où g désigne la fonction de lien identité ; les coefficients de régression $\hat{\beta}_j$, et l'ordonnée à l'origine $\hat{\beta}_0$, sont estimés par le critère usuel des moindres carrés

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \parallel Y - X\beta \parallel^2, \tag{1.2}$$

c'est-à-dire en recherchant les β qui minimisent l'erreur quadratique entre les valeurs prédites et les valeurs observées. La précision de ces estimateurs étant directement fonction du nombre d'observations et de la matrice de dessin X, il devient clair que lorsque le nombre de prédicteurs (p) est grand, le modèle fournit des estimations de moins en moins précises et que l'on a un problème de sur-ajustement.

Une solution à ce problème consiste à pénaliser les termes du modèle, en considérant soit une pénalisation de norme L_2 (*ridge regression*, Hoerl and Kennard, 1988), soit une pénalisation de norme L_1 (*lasso regression*, Tibshirani, 1996) :

$$\hat{\beta}_{\text{RIDGE}} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda\|\beta\|^2, \quad \text{avec } \|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 \tag{1.3}$$

$$\hat{\beta}_{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda\|\beta\|_1, \quad \text{avec } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \tag{1.4}$$

La régression régularisée de type *ridge* permet de meilleures prédictions que le modèle OLS classique, grâce à un meilleur compromis entre biais et variance ; malheureusement, elle a le défaut de conserver tous les prédicteurs dans le modèle, donc le modèle final n'est pas vraiment parcimonieux en termes de nombre de paramètres. La régression de type *sparse* (LASSO) permet au contraire la sélection automatique des variables en imposant la nullité de certains coefficients, selon les valeurs du paramètre de pénalisation λ. Toutefois, elle fournit de moins bons résultats que la régression *ridge* lorsque les prédicteurs sont fortement corrélés, c'est-à-dire en présence de multi-colinéarité (Efron et al., 2004). En termes de sélection de variables, cela se révèle peu avantageux pour les données de génomique dans lesquels on a souvent affaire à des groupes de gènes.

Une solution proposée plus récemment consiste à utiliser un compromis des deux critères. Le critère elasticnet proposé par (Zou and Hastie, 2005) est de la forme :

$$L(\lambda_1, \lambda_2, \beta) = \|Y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{1.5}$$

où $\|\beta\|^2$ et $\|\beta\|_1$ sont définis comme en (1.3), et l'estimateur $\hat{\beta}$ est défini comme :

$$\hat{\beta}_{\text{elasticnet}} = \underset{\beta}{\operatorname{argmin}}\, L(\lambda_1, \lambda_2, \beta) \tag{1.6}$$

Les deux paramètres de régualarisation permettent d'une part la sélection de variables et d'autre part d'autoriser dans le cas $n \ll p$ la sélection de $m > p$ variables.

Arbres de décision et forêts aléatoires. Les arbres de classification et de régression ou CART (pour *Classification And Regression Trees*, Breiman et al., 1984) reposent sur un algorithme récursif dans lequel on cherche à partitionner les individus en classe tout en minimisant la variance ou hétérogénité intra-classe [1]. Le critère d'arrêt peut porter sur l'effectif minimal dans chaque nœud terminal, ou le nombre total de nœuds. Selon le type de variable réponse considérée, continue ou catégorielle, on parlera d'arbres de régression ou de classification, respectivement.

Si l'on considère m variables, le partitionnement est effectué à partir d'une mesure quantitative du degré d'hétérogénité intra-classe, encore appelée impureté nodale, par exemple l'*indice de Gini* (ou erreur de plus proche voisin) dans le cas d'une variable réponse catégorielle :

$$i(\Omega) = \sum_{r \neq s} p(r \mid \Omega) p(s \mid \Omega) \tag{1.7}$$

où $p(r|\Omega)$ est la probabilité d'être dans la classe r ($r = 1, \ldots, m$) parmi les individus se retrouvant au nœud Ω. D'autres indices peuvent être utilisés, par exemple l'indice d'information ou de déviance, $i(\Omega) = -p_\Omega \log(p_\Omega) - (1 - p_\Omega) \log(1 - p_\Omega)$, avec p_Ω désignant toujours une probabilité conditionnelle. Dans le cas d'une variable réponse numérique, on remplace l'indice précédent par l'erreur quadratique moyenne :

$$i(\Omega) = \frac{1}{n_\Omega} \sum_{i \in \Omega} (y_i - \bar{y})^2 \tag{1.8}$$

où $\bar{y}$ désigne le trait moyen dans l'échantillon.

Ce type d'approche très flexible ne suppose pas de modèle sous-jacent et permet de s'accomoder d'un nombre important de variables. Toutefois, la stabilité des solutions proposées n'est pas assurée, et il est nécessaire de recourir à une procédure d'élagage de l'arbre et de validation croisée pour minimiser le risque de sur-ajustement.

Pour cette raison, entre autres, Breiman (2001) a proposé une extension intéressante de cette approche, dénommée forêts aléaoires (*random forests*, RF), dans laquelle au lieu d'évaluer un seul arbre, on construit un ensemble d'arbres de décision grâce à un rééchantillonnage par bootstrap. Le procédé est donc analogue aux méthodes d'aggrégation de modèles, tel que le *bagging*. Pour les aspects de prédiction, une procédure de votes (ou de moyennage) permet de sélectionner le meilleur arbre parmi les sous-ensembles constitués. En d'autres termes, en combinant les prédictions fournies par différents arbres de décision construits à partir de sous-ensembles obtenus par rééchantillonnage, on arrive à compenser l'instabilité inhérente aux arbres de décision et à obtenir des prédictions relativement précises (Breiman, 1996; Bauer and Kohavi, 1999).

1. L'approche CART donne des résultats généralement comparables aux algorithmes C4.5 et C5.0 de Quinlan (1993).

Le principe général est résumé ci-dessous :

1. on spécifie le nombre de variables n qui servira d'ensemble de prédicteurs parmi les N variables de départ (généralement, $n \sim \sqrt{N}$) ;
2. chaque arbre (de profondeur maximale) est construit à partir d'un échantillon bootstrap des individus de l'ensemble d'apprentissage ;
3. à chaque nœud, n variables sont sélectionnées aléatoirement parmi les N variables ;
4. la division de l'arbre se fait selon un critère de maximisation du gain d'information sur ces n variables.

Pour chaque arbre, environ un tiers des individus sert d'individus de test (équivalent de l'algorithme .632+ pour le bootstrap) : on parle d'individus « out-of-bag » (OOB).

On voit donc qu'à la différence des arbres de décision, à chaque étape de construction d'un nouvel arbre, on n'utilise pas de procédure d'élagage ; de même, on n'utilise pas l'ensemble des prédicteurs pour construire un arbre, mais seulement un sous-ensemble aléatoire (correspondant généralement aux deux tiers de l'ensemble initial). Ceci permet *a priori* de mieux gérer le problème de colinéarité entre variables puisque du fait de la procédure d'échantillonnage des variables, les prédicteurs corrélés entre eux ne seront pas nécessairement sélectionnés ensemble. Dans le cas qui nous intéresse, cela offre la possibilité d'évaluer l'importance relative des SNP, même lorsque ceux-ci sont en déséquilibre de liaison.

La mise en œuvre des RF dans le domaine génétique, en particulier les données d'expression de gènes (e.g., Lunetta et al., 2004; Díaz-Uriarte and Alvarez de Andrés, 2006; Strobl et al., 2008), suggère que ce type d'approche est à même d'identifier de bons prédicteurs parmi un large ensemble de variables, tout en tenant compte des potentielles interactions entre ces variables. Dans une étude cas-témoins, Bureau et al. (2005) ont utilisé les RF pour explorer l'importance relative de SNP isolés ou bien de paires de SNP localisés sur le gène ADAM33 (impliqué dans les troubles asthmatiques) et leurs associations spécifiques avec le statut des patients. Leurs résultats suggèrent que l'importance relative, et par là le pouvoir prédicteur, des SNP converge avec les mesures d'association, à quelques fluctuations près.

Régression logique. La régression logique (Ruczinski et al., 2004; Kooperberg et al., 2005) revient à estimer un modèle consistant en une somme d'expressions booléennes et prenant la forme classique d'un modèle linéaire

$$g\big(\mathbb{E}(y)|X\big) = \beta_0 + \sum_{j=1}^{p} \beta_j L_j \tag{1.9}$$

où L_j désigne une combinaison booléenne de prédicteurs binaires, par exemple $(x_1 \wedge x_2) \vee (x_3 \wedge \bar{x}_4)$ avec $\wedge$, $\vee$ et $\bar{x}$ désignant l'intersection, l'union et le complémentaire, respectivement. Dans le cas des données de type SNP, un codage binaire peut être obtenu en considérant la présence/absence d'au moins 1 allèle variant.

Dans la mesure où il est impossible d'énumérer et tester toutes les combinaisons linéaires lorsque p est grand, on utilise des algorithmes de recherche et de recuit simulé pour réduire l'espace des

prédicteurs. Comme dans le cas des CART, on évalue l'apport d'une nouvelle variable dans le modèle à l'aide d'une fonction score basée sur la variable réponse. Dans le cas d'une réponse quantitative, il peut s'agir du critère des moindres carrés. Enfin, le meilleur modèle prédictif est recherché au travers d'une procédure de validation croisée, le plus souvent en 10-fold, et la distribution de référence du modèle final peut être obtenue par une procédure de test par randomisation (permutations).

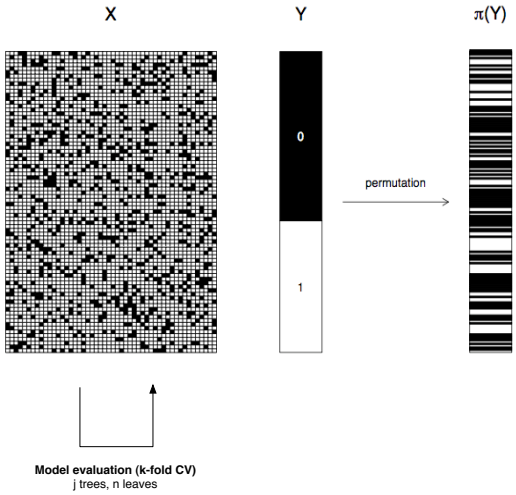


FIGURE 1.2 – Illustration de la procédure d'estimation des paramètres d'un modèle de régression logique.

Régression PLS. On peut approcher le problème de sélection de variable par des techniques d'ordination en espace réduit, telle que l'analyse en composantes principales (PCA) dans sa version pénalisée, ou *sparse* (Zou et al., 2006). Toutefois, l'application de cette technique reste limitée au cas où l'on n'a qu'une seule variable réponse. Dans le cas multivarié conjoint, c'est-à-dire en présence de deux tableaux de données, dont l'un peut être considéré comme un ensemble de variables réponse, la régression PLS offre une alternative intéressante puisqu'elle permet de modéliser de manière asymétrique le lien entre un bloc de variables prédictrices et un bloc de variables réponses (Tenenhaus, 1998). Suivant le type de variable réponse considérée, scalaire ou vecteur, on parlera de PLS1 et de PLS2, respectivement. Nous ne nous intéresserons qu'à ce second type de PLS.

Le principe général de la PLS2 est le suivant : on cherche à construire une succession de combinaisons linéaires (orthogonales entre elles) des variables de chaque bloc telles que leur covariance soit maximale. En d'autres termes, on cherche à construire de nouvelles variables (appelées également composantes ou facteurs) contenant le maximum d'information sur le bloc X $(n \times p)$ et qui permettent de prédire Y $(n \times q)$, tout en réduisant la dimension des données, sur X et/ou Y.

Plus formellement, si l'on part de la représentation proposée dans la figure 1.3, cela revient à chercher les vecteurs u_h et v_h, vecteurs des poids de chaque variable dans les composantes

principales de X et Y (resp.), de normes unité, sous la contrainte

$$\max_{\|u_h\|=1,\|v_h\|=1} \operatorname{cov}(X_{h-1}u_h, Yv_h) \quad \big(\equiv \max \operatorname{cov}(\xi_h, \omega_h)\big) \tag{1.10}$$

où X_{h-1} représente la matrice des résidus de X après régression de Y sur X. Ici, h représente le rang de la composante et ce processus peut être itéré pour $h = 1, \ldots, H$ composantes.

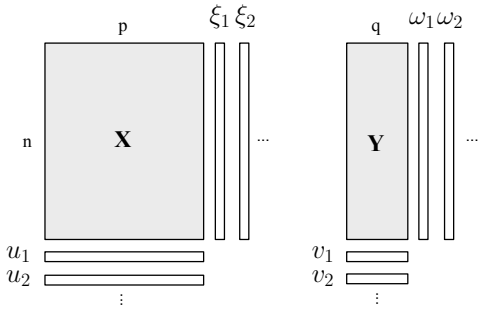


FIGURE 1.3 – Principe de la régression PLS.

Bien qu'il existe plusieurs modes de déflation de la matrice X (PLS mode A *vs.* PLS2), les premières composantes sont toujours identiques quelle que soit la méthode employée.

PLS et décomposition en valeurs singulières

N'importe quelle matrice Z $(p \times q)$ de rang r peut se décomposer (principe de la décomposition en valeurs singulières) sous la forme

$$Z = U\Delta V^t$$

où U $(p \times r)$ et V $(q \times r)$ sont orthogonales et Δ est une matrice diagonale dont les éléments δ_{kk} $(k = 1, \ldots, r)$ sont les valeurs singulières. Celles-ci correspondent à la racine des valeurs propres de la matrice Z^tZ (mais également ZZ^t). Si on prend $Z = X^tY$, alors les vecteurs singuliers droits et gauches, $(u_1, \ldots, u_r)$ et $(v_1, \ldots, v_r)$, qui sont les vecteurs colonnes de U et V, correspondent strictement aux vecteurs de poids (« charges ») de la PLS de X et Y.

Ainsi, les premiers vecteurs canoniques de la PLS (u_1 et v_1) constituent la meilleure approximation de la matrice de covariance de X et Y dans les deux directions.

L'implémentation algorithmique à l'aide d'une décomposition en valeurs singulières (encadré ci-dessus) optimise nettement les temps de calcul. Par ailleurs, il est possible d'intégrer une étape de sélection de variables directement dans cet algorithme, grâce à une fonction de seuillage des poids dans n'importe laquelle des deux composantes à l'itération h. Par exemple, l'application d'une fonction de type [2]

$$g_\lambda(y) = \operatorname{sign}(y)\big(|y| - \lambda\big)_+,$$

2. où l'on définit les deux fonctions

$$(x)_+ = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad \text{et} \quad \operatorname{sign}(x) = \begin{cases} -1, & x < 0 \\ 1, & x > 0 \\ 0, & x = 0 \end{cases}$$

telle que proposée dans la PCA pénalisée de Zou et al. (2006), permet de ne conserver que les variables ayant un poids supérieur à λ (en valeur absolue). On reconnaît là la pénalisation LASSO décrite en § 1.2.4 qui est celle proposée par Lê Cao et al. (2008), mais Chun and Keleş (2007) ont également proposé une pénalisation reposant sur elasticnet. Malgré la pénalisation des vecteurs u et v, les composantes de chaque bloc restent généralement proches de l'orthogonalité.

Analyse canonique des corrélations. L'analyse canonique des corrélations, CCA, permet également d'analyser le lien entre deux blocs de données, mais cette fois-ci de manière symétrique (aucun bloc ne possède le statut de variables réponse) et en maximisant la corrélation, plutôt que la covariance, entre les composantes principales successives de chacun des blocs. En reprenant les notations précédentes et en s'appuyant sur le schéma de la figure 1.3, ceci revient à trouver les composantes principales de chaque bloc (i.e., les vecteurs principaux u_h et v_h) sous la contrainte

$$\max_{\|u_h\|=1,\|v_h\|=1} \frac{u_h^t \Sigma_{XY} v_h}{\left(u_h^t \Sigma_{XX} u_h\right)^{1/2} \left(v_h^t \Sigma_{YY} v_h\right)^{1/2}} \tag{1.11}$$

où Σ_{XX}, Σ_{YY} et Σ_{XY} désignent les matrices de corrélations et de covariances de X et Y. Comme précédement, ceci est strictement équivalent à la recherche de l'argument qui rend maximum la corrélation entre les vecteurs scores associés, soit $\operatorname{argmax} \operatorname{cor}(\xi_h, \omega_h)$, sous les mêmes contraintes de norme unité des vecteurs u_h et v_h.

Si l'on considère le bloc X, par analogie avec l'analyse en composantes principales, les vecteurs scores individuels, ξ_h et ξ_{h+1}, donnent les coordonnées des individus dans le plan factoriel $(h, h+1)$ tandis que les vecteurs principaux (encore appelés vecteurs canoniques) donnent les coordonnées des variables dans le cercle des corrélations défini à partir des mêmes composantes factorielles. La différence avec la PCA est que la CCA permet d'effectuer cette réduction factorielle sur les deux blocs de variables, tout en maximisant l'information mutuelle entre les blocs : on maximise à la fois l'information intra-bloc (les composantes sont celles qui expliquent le mieux la variance de leur propre bloc) et l'information inter-bloc (les composantes de même rang sont maximalement corrélées).

Parkhomenko et al. (2009) ont proposé une implémentation d'une version pénalisée de la CCA, reposant comme dans le cas de la PLS sur l'utilisation d'une décomposition en valeurs singulières et d'une pénalisation de type LASSO sur les vecteurs singuliers droit et gauche (cf. algorithme en Annexe, page 51). Une version pénalisée de la CCA à 2 ou k blocs de données a également été développée par Witten et coll. (Witten and Tibshirani, 2009; Witten et al., 2009).

1.2.5 Quel algorithme retenir ?

Le choix d'un algorithme relève de considérations à la fois informatiques (temps de calcul, occupation mémoire) et statistiques (codage des variables prédictrices, régression ou classification). Les SNP étant par nature des variables catégorielles nominales, tout porte à privilégier les méthodes pour données catégorielles. Or, comme nous venons de le voir, celles-ci ne permettent pas de considérer des blocs de variables-réponse. Il semble toutefois justifié de considérer les données

de génotypage comme un codage discret de la fréquence de l'allèle mineur, voire de considérer le statut de ces variables comme purement numérique. Dans ce cas, les deux méthodes multivariées décrites aux paragraphes précédents peuvent être appliquées, nonobstant l'hypothèse de normalité multivariée pour les vecteurs-réponse. En comparaison de méthodes telles que les forêts aléatoires ou la régression logique, elles offrent l'avantage de permettre de prendre en considération à la fois les interrelations entre les SNP (résultant de leur corrélation spatiale ou d'interactions d'ordre supérieur) et celles entre les variables réponse.

1.3 Organisation de l'étude

L'objectif de ce travail est donc de comparer les performances de différentes méthodes de sélection de variables lorsque l'on considère deux blocs de variables, dont l'un peut être assimilé à un ensemble de variable réponse ou non. Nous porterons un accent tout particulier à la capacité de généralisation de chacune de ces méthodes, tout en prenant en considération les temps de calculs inhérents à leur mise en œuvre.

Ce travail est composé de deux parties (non indépendantes) : dans un premier temps, nous nous intéresserons à la robustesse de deux méthodes multivariées dans leur version pénalisée, la régression PLS et l'analyse canonique des corrélations, qui permettent toutes les deux l'étude des liens entre deux tableaux de données (Chap. 3). Avant d'appliquer ces méthodes sur des jeux de données de « dimension raisonnables » (entre une centaine et un millier de variables), nous effectuerons un filtrage univarié sur l'ensemble des données de génotypage (Chap. 2). Dans tous les cas, les SNP seront considérés selon un codage quantitatif discret, assimilé à une variable aléatoire continue. Dans un second temps, nous mettrons en pratique ces méthodes sur un cas concret où une liste de SNP candidats nous est fournie (Chap. 4). Nous comparerons les résultats de la CCA et de la PLS à une régression pénalisée de type LASSO, et à une méthode autorisant le traitement des prédicteurs d'intérêt dans leur forme qualitative (*random forests*). L'intérêt et les limites de ces approches sont ensuite discutés dans le dernier chapitre, de même que les perspectives soulevées par ce champ d'investigation volontairement restreint (Chap. 5).

Les analyses statistiques sont réalisées avec le logiciel R (version 2.9.1, 2009–06–26, i386-apple-darwin8.11.1, `cran.r-project.org`) sur un Pentium Intel Core Duo 2.8 GHz (4 Go RAM). Pour la représentation et la manipulation des données génétiques, nous utilisons le package snpMatrix (Clayton and Cheung, 2007).

Chapitre 2

Approche univariée

Cette étape d'analyse univariée a pour principal objectif de fournir une liste de « SNP candidats » que l'on utilisera comme variables prédictrices dans les approches multivariées.

2.1 Description du jeu de données

2.1.1 Échantillon de sujets

Les individus, tous locuteurs français natifs, ont été recrutés sur la base du volontariat. Ils sont composés pour près des deux tiers d'hommes (60 %), et leur niveau d'études, bien que variable, est généralement celui des études supérieures (48 %). Différentes mesures de performance dans des tâches de psycholinguistique ont été recueillies sur cet échantillon, mais celles-ci ne seront pas analysées dans le cadre de cette étude.

2.1.2 Données de neuroimagerie

Les acquisitions ont été réalisées dans 2 scanners de même résolution (SHFJ, Orsay) avec 33 % des individus dans l'un des deux scanners.

On dispose de scores standardisés en imagerie fonctionnelle calculés à partir de deux tâches – lecture et parole – sous la forme d'un index de latéralisation représentant le différentiel de signal BOLD entre les deux hémisphères $((\text{BOLD}_D - \text{BOLD}_G)/\sqrt{\text{BOLD}_D^2 + \text{BOLD}_G^2})$. Cet indice est recueilli sur 34 régions d'intérêt, et sa distribution entre-sujets sur l'ensemble des régions d'intérêt est représentée dans la figure 2.1 (c). Les distributions sont légèrement asymétriques (vers les valeurs négatives) mais demeurent confinées dans le même intervalle $[-1.5; 1.5]$ sans que l'on relève la présence de valeurs extrêmes particulièrement flagrantes.

Si l'on réalise une analyse en composantes principales pour vérifier la structure de corrélation de ces régions, on constate que ces mesures sont très corrélées entre elles et peuvent se résumer sous la forme d'une, voire deux combinaisons linéaires. L'examen du diagramme des valeurs propres (Figure 2.1, a) sur lequel on a superposé les résultats observés avec des matrices simulées de mêmes dimensions (principe de l'analyse parallèle, REF) montre que l'on peut dégager deux axes

principaux expliquant un peu moins de 20 % de l'inertie ; au-delà, la constitution de nouveaux axes est sujette aux fluctuations d'échantillonnage. Le cercle des corrélations est représenté dans la partie droite de la figure 2.1 (c) et on ne voit pas apparaître de projections spécifiques des régions selon le type de tâche (surligné en couleur pour faciliter la lisibilité). On note toutefois que certaines variables demeurent relativement mal représentées dans ce plan factoriel $(1, 2)$, avec des contributions relatives aux deux axes < 2 % pour 11 d'entre elles (dont 8 sont des conditions de lecture). Enfin, l'examen du nuage des individus ne fait pas apparaître de différence notable selon le scanner d'acquisition (Figure 2.1, b).

2.1.3 Préparation des données génétiques

Les données génétiques dont nous disposons concernent $n = 138$ individus génotypés sur 1054068 SNP avec un taux de valeurs manquantes variable entre les individus et entre les biomarqueurs. Comme nous disposons de mesures d'imagerie pour 94 sujets seulement, nous ne considérerons par la suite que les données de génotypage de ces 94 individus, mais le filtrage des données génétiques est réalisé sur l'ensemble des individus disponibles.

Le filtrage comprend trois étapes réalisées en parallèle :

1. suppression des SNP dont la fréquence de l'allèle mineur est inférieure à 10 %,
2. suppression des SNP ne vérifiant pas l'équilibre de Hardy-Weinberg à $p < 0.005$,
3. suppression des SNP dont le taux de génotypage est inférieur à 95 %.

L'application de ces trois critères doit permettre de constituer une base de données assainie et conforme aux critères habituellement retenus dans les études d'association portant sur le génome entier.

Cette procédure de filtrage, réalisée sous R amène à retenir 622534 SNP, l'application des différents critères étant résumée dans le tableau 2.1.

	MAF	HWE	GENO	Total conservé
R	355225	164232	66235	*622534*
PLINK	359244	30137	64821	*628127*

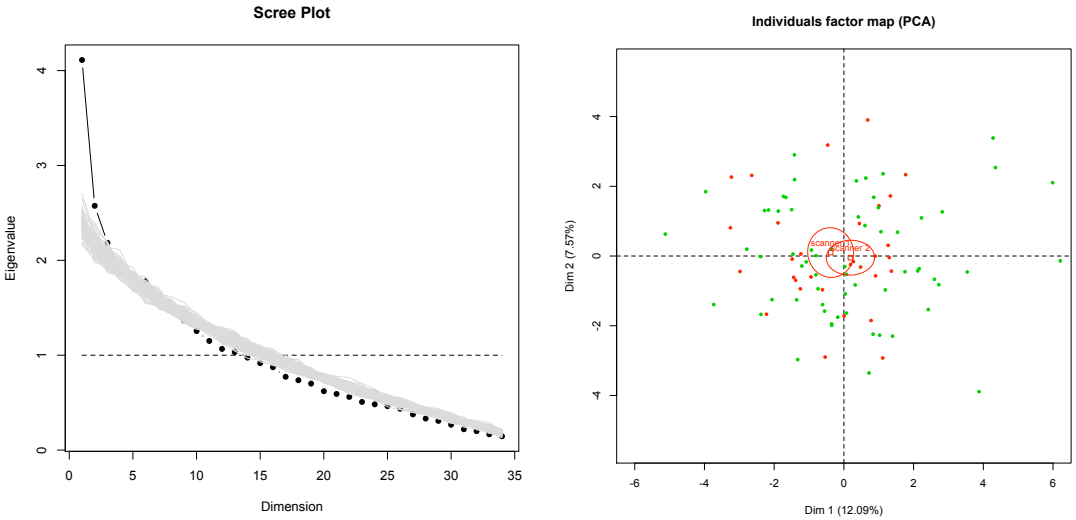
TABLE 2.1 – Résultats du filtrage des données génétiques sous R et plink.

Le même type de filtrage réalisé avec PLINK (Purcell et al., 2007), avec la commande suivante

```
plink --noweb --bfile localizer_V1 --out localizer_V1_filt --make-bed
      --maf 0.1 --hwe 0.005 --geno 0.05
```

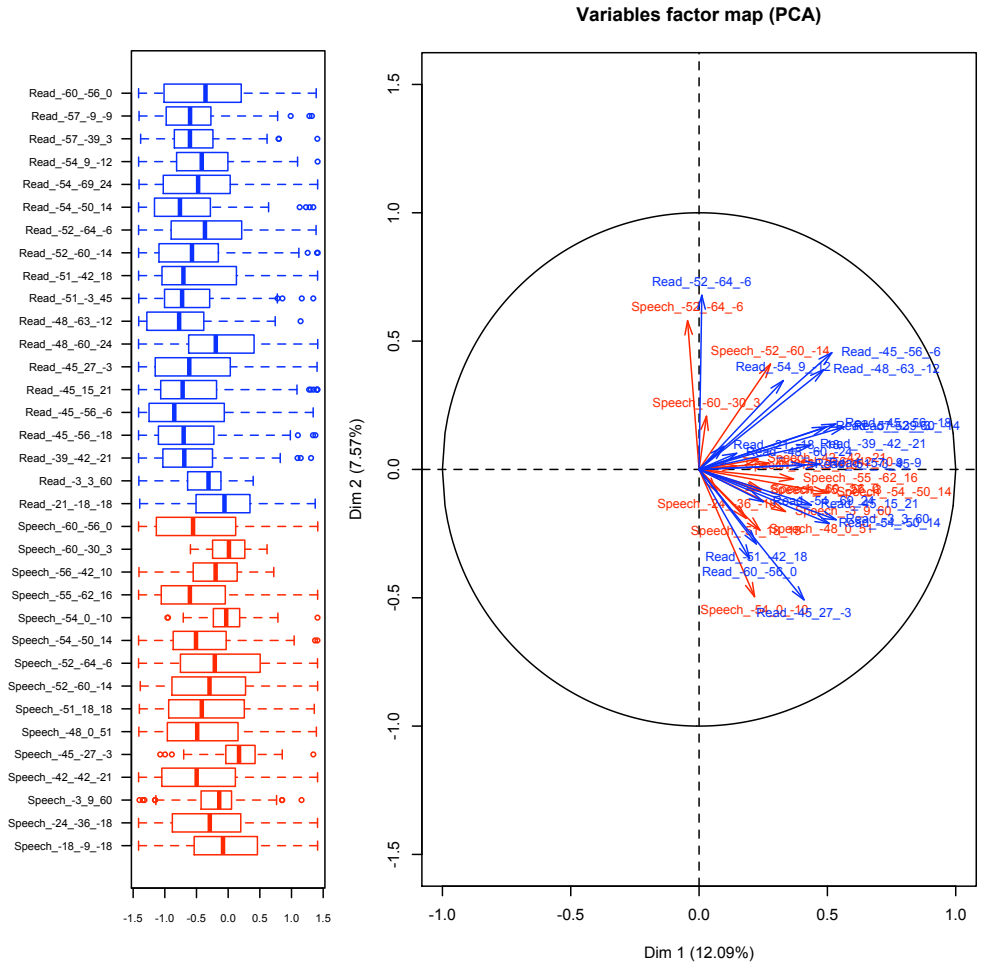
nous amène à conserver 628127 (Tableau 2.1), la différence s'expliquant essentiellement par l'ordre dans lequel sont effectués les tests sous PLINK.

Une fois le filtrage effectué, nous ne conservons que les données de génotypage concernant les 94 individus disposant de mesures en imagerie sur les 34 régions d'intérêt. On notera cependant que cette procédure de filtrage ne garantit pas de conserver un MAF ≥ 10 % lorsqu'on restreint



(a)

(b)



(c)

FIGURE 2.1 – (a) Diagramme des valeurs propres (en gris, 100 simulations sur des matrices de mêmes dimensions). (b) Nuage des individus dans le plan factoriel $(1, 2)$ en fonction du scanner. (c) Distribution des scores d'imagerie (tâche de lecture en bleu, tâche de parole en rouge) et représentation des variables dans le cercle des corrélations.

les données de génotypage aux 94 individus. Toutefois, le MAF minimum est de 0.085 avec la moitié des SNP ayant un MAF compris entre 0.181 et 0.399 (MAF médian, 0.277).

Au final, nous avons donc deux blocs de données de dimensions 94×34 (imagerie, bloc Y) et 94×622534 (génétique).

2.2 Tests d'association imagerie *vs.* génétique

Pour réaliser les tests d'association univariés, nous calculons la statistique F (Fisher-Snedecor) associée à la régression linéaire de chacun des 34 scores d'imagerie sur l'ensemble du génome, en considérant les données de génotypage comme numériques. Ceci revient donc à réaliser un peu plus de 2.10^7 (34×622534) modèles linéaires. Afin d'optimiser la procédure de test, nous n'utilisons pas la décomposition QR pour la méthode OLS mais calculons la corrélation linéaire entre chaque couple (X_i^j, Y_i^k) ($j = 1 \ldots 622534$, $k = 1 \ldots 34$) et transformons la statistique r en F, à l'aide de la relation suivante :

$$F = \nu_a \frac{r^2}{1 - r^2}, \tag{2.1}$$

où $\nu_a = (n-2) - \#\{(x_i = \text{NA}, y_i = \text{NA})\}$ représentent les degrés de liberté ajustés du fait des valeurs manquantes éventuellement présentes conjointement dans les observations considérées. Dans le cas où il n'y a pas de paires de données manquantes, on retrouve la formule usuelle d'un coefficient de détermination $(n-2) \times r^2/(1-r^2)$, qui traduit le rapport entre la part de variance expliquée dans la variable réponse et la résiduelle.

Les figures A-1 à A-4, reproduites p. 42–45, résument la distribution des p-valeurs (exprimées en $-\log_{10}$) sur l'ensemble du génome pour chacune des 34 régions d'intérêt. Seules les p-valeurs (non corrigées) significative à 10^{-3} sont représentées. Les p-valeurs significatives au seuil 10^{-6} sont quant à elles représentées en rouge. Il apparaît que peu de SNP remplissent cette dernière condition.

Sachant qu'un ajustement très conservateur telle que la correction de Bonferroni consiste à multiplier les p-valeurs estimées par le nombre de tests effectués (ici, $34 \times 622534 = 21166156$), cela revient à considérer des p-valeurs $\leq 2,4.10^{-9}$. En faisant abstraction des tests sur les différentes régions d'intérêt, les p-valeurs corrigées devraient être inférieures à $8,0.10^{-8}$. Dans ce dernier cas, nous arrivons à dégager 3 SNP significatifs au seuil $\alpha = 0.05$ (corrigé).

Chapitre 3

Approches multivariées

3.1 Objectif

L'analyse univariée nous a permis de dégager une liste de SNP associés significativement aux 34 variables réponse considérées, ci-après regroupées sous l'appellation bloc Y. Nous considérerons dans la suite les 1000 SNP ayant la valeur de probabilité observée la plus basse sur l'ensemble des 34 régions, en ne considérant qu'une seule instance des SNP lorsque ceux-ci sont significativement associés à plusieurs régions d'intérêt [1].

Ceci appelle d'emblée deux remarques concernant (1) le choix du nombre de SNP et (2) la manière de sélectionner ces 1000 SNP :

1. le choix du nombre de SNP peut être considéré comme arbitraire, mais si l'on se replace dans l'optique d'une approche prospective de comparaison des performances de différents algorithmes, ce nombre permet déjà de se placer dans des dimensions « respectables », et comparable à certaines études déjà publiées (e.g., Parkhomenko et al., 2007) ;
2. la sélection des SNP est ici effectuée de manière très simple, par simple classement et élimination de la redondance, alors que d'autres méthodes de sélection de variables pourraient être employées, comme on l'a discuté en introduction.

Toutefois, cette approche se justifie par notre volonté de nous situer, d'une part, dans une gamme de prédicteurs acceptable du point de vue de leur nombre, et d'autre part de ne retenir que des variables potentiellement intéressantes afin d'éviter d'appliquer des algorithmes de sélection de variables sur du « bruit ».

3.2 Méthodologie

3.2.1 Sélection des SNP

À partir de la matrice de p-valeurs estimées durant l'approche univariée, nous procédons à une sélection des 1000 p-valeurs les plus basses sur chaque région d'intérêt. Or, comme nous utilisons

1. Le taux de redondance des SNP sur les différentes régions est < 5 %.

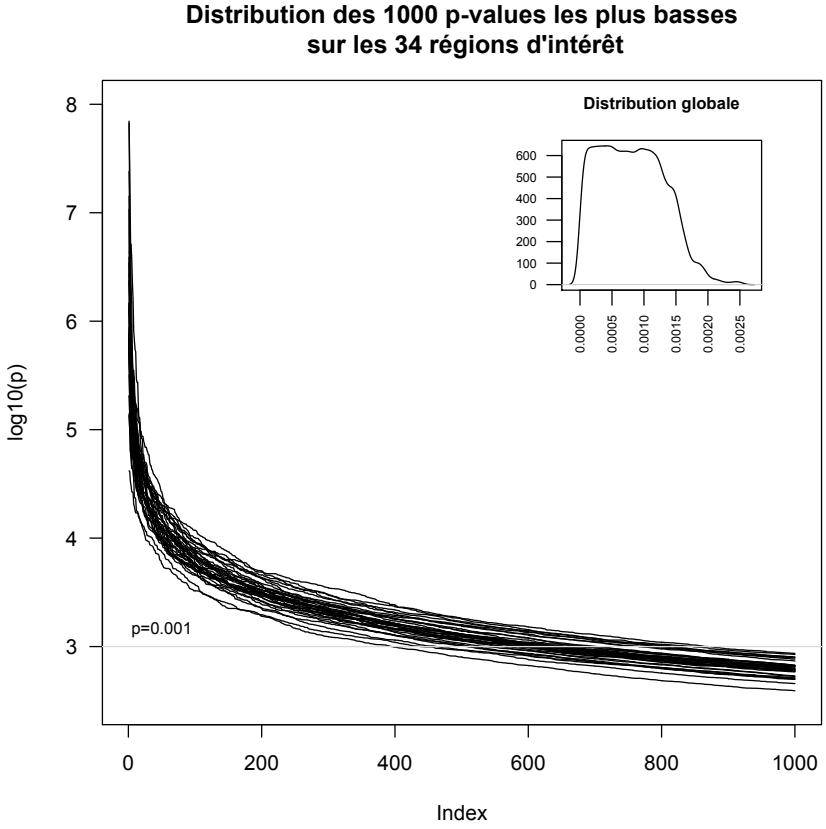


FIGURE 3.1 – Distribution des p-values les plus basses sur les différentes régions de neuroimagerie.

une procédure de validation croisée « leave-one-out », cette procédure est répétée 94 fois, en considérant à chaque fois des blocs de 93 individus [2].

La distribution des 1000 « meilleures » p-valeurs (non corrigées) par régions d'intérêt est illustrée dans la figure 3.1. On voit que, parmi les SNP candidats, peu de SNP ont une p-valeur $< 10^{-6}$.

3.2.2 Comparaison des méthodes multivariées

Étant donnés les critères exposés plus haut, la comparaison entre les différentes méthodes portera sur :

– l'erreur de prédiction pour la régression PLS ;
– les scores prédits sur la première composante de chaque bloc, pour la CCA.

Par ailleurs, on relèvera dans chaque condition le nombre total de SNP sélectionnés, ainsi que la corrélation canonique entre les deux premières composantes dans le cas de la CCA.

L'erreur de prédiction sera calculée comme la distance euclidienne entre la valeur prédite, $\hat{y}$, et la valeur observée, y (vecteur de réponse) Le choix d'une distance de norme L_2 peut être discuté dans la mesure où l'on fait une hypothèse de normalité pour la distribution des réponses, mais

2. À raison de 15 min de calcul sur un Pentium Dual Core cadencé à 2,8 GHz, et en ne stockant que les 1000 p-values les plus basses, cela représente environ 24h de calcul et 35 Mo de données. Au final, nous disposons de 94 matrices de p-values sur l'ensemble des régions d'intérêt et des SNP.

elle offre l'avantage d'être facile et rapide à calculer. Notre statistique de test sera donc de la forme :

$$\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right)^{1/2}, \tag{3.1}$$

où l'indice i réfère à l'individu de test issu de la procédure LOO.

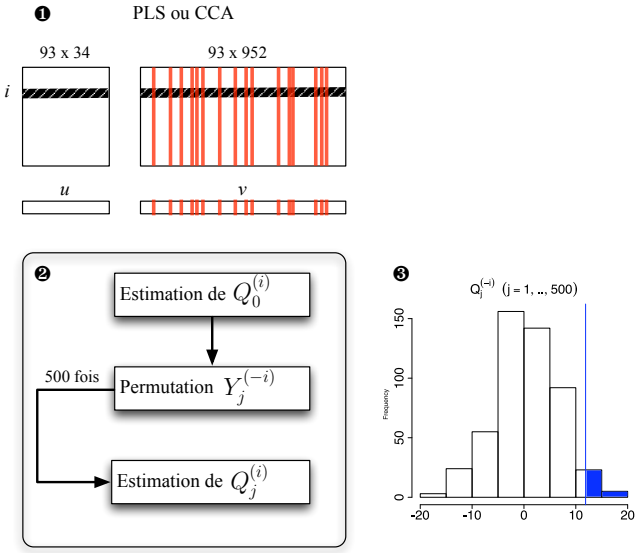
La statistique de test pour la CCA prend la forme :

$$(x_i^t u)(y_i^t v), \tag{3.2}$$

où $(x_i u, y_i v)$ désignent les vecteurs canoniques chez l'individu de test i. Il s'agit simplement des vecteurs réponse X et Y de l'individu i pondérés par les vecteurs de poids des premières composantes. Leur produit scalaire produit est un scalaire.

Pour estimer la distribution sous H_0 de ces deux statistiques de test, notées Q, on effectue pour chaque combinaison des paramètres d'étude 500 permutations des lignes du bloc d'apprentissage Y et on réestime notre statistique, ce qui nous permet *in fine* d'obtenir une p-valeur empirique de notre statistique.

L'ensemble de ces étapes est résumé dans le schéma ci-dessous :



L'étape ❶ consiste à estimer les paramètres du modèle, en l'occurence les combinaisons linéaires des variables de chaque bloc qui maximise la covariance (PLS) ou la corrélation (CCA), avec pénalisation (schématisée ici par les bandes de couleur rouge). Cette étape est réalisée en omettant l'individu i.

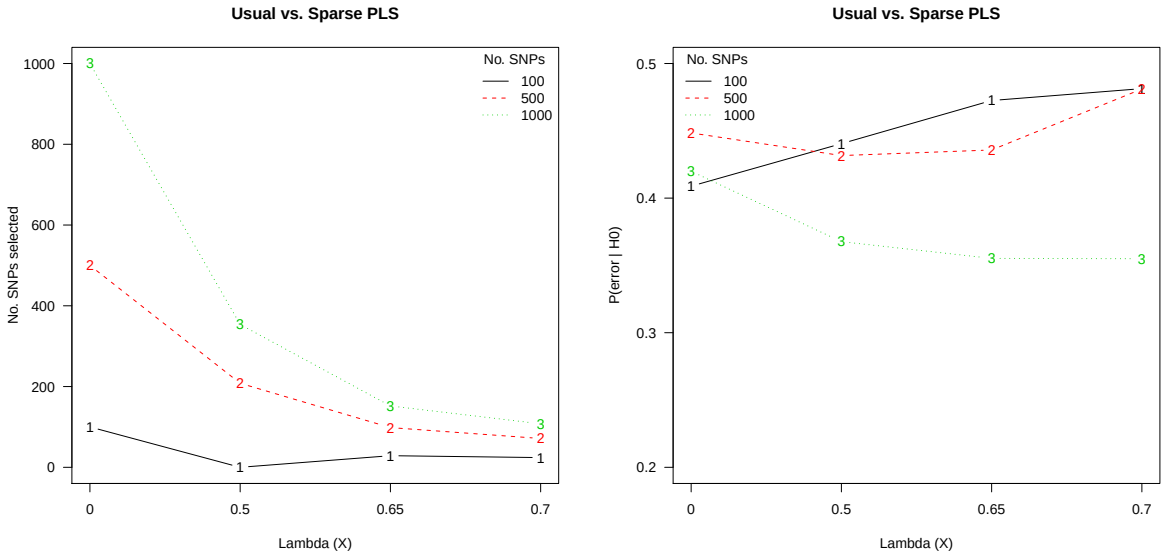
À l'étape ❷, on estime la statistique de test pour l'individu i, $Q_0^{(i)}$, sur ces mêmes données, puis on réestime les paramètres du modèle et la statistique de test $Q_j^{(i)}$ associée, après la $j$ème permutation aléatoire du bloc Y. Ceci permet d'estimer la distribution de Q sous H_0 et la p-valeur correspondante à l'étape ❸.

Ces étapes sont réalisées pour les 94 individus (procédure *leave-one-out*).

3.3 Résultats

3.3.1 Régression PLS

La distribution du nombre de SNP sélectionnés et celle de la statistique de test considérée est illustrée dans la figure 3.2 (a et b, respectivement). Le cas $\lambda = 0$ correspond à une régression PLS classique, c'est-à-dire sans pénalisation du bloc X.



(a) (b)

FIGURE 3.2 – (a) Nombre de SNP sélectionnés selon les paramètres appliqués au modèle PLS (degré de pénalisation et nombre de variables initiales). (b) Distribution sous H_0 de la statistique de test considérée (500 permutations).

Le nombre de SNP sélectionnés diminue à mesure que le degré de pénalisation augmente, comme on pouvait s'y attendre. Le taux de sélection varie de 15 % (1000 SNP) à 29 % (100 SNP) pour un paramètre de pénalisation intermédiaire ($\lambda = 0.65$). En ce qui concerne la distribution de la statistique de test sous randomisation, on constate en revanche que les degrés de significativité (p-valeurs) sont toujours supérieurs à 10 % suggérant que l'erreur estimée ne s'éloigne pas vraiment de ce que l'on pourrait observer sous H_0 (prédiction « au hasard »).

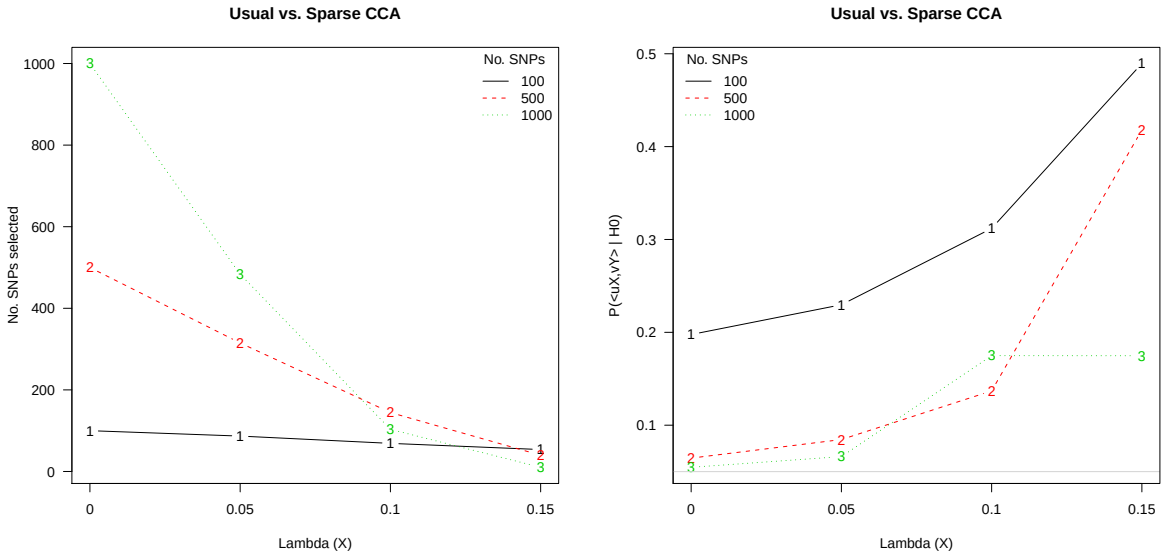
Comme on peut le vérifier dans le tableau suivant, l'erreur de prédiction ne semble diminuer qu'en présence de pénalisation et pour un nombre de variables initiales > 500. Dans le cas où X comporte 100 ou 500 variables, on constate une augmentation de l'erreur en présence de pénalisation, quelle que soit la valeur de λ.

		Degré de pénalisation (λ)			
		0	0.05	0.1	0.15
Variables	100	3.933	3.948	3.964	3.975
	500	3.950	3.947	3.958	3.987
	1000	3.922	3.899	3.901	3.897

3.3.2 Analyse canonique des corrélations (CCA)

Dans la figure 3.3 sont résumés les résultats portant sur le nombre de SNP sélectionnés (a) et l'extrêmalité de notre statistique de test après permutation (b). Le cas $\lambda = 0$ correspond à une CCA classique (pas de pénalisation du bloc X).

De manière comparable à la PLS, le nombre de SNP sélectionnés diminue naturellement lorsque le paramètre de pénalisation augmente, et cet effet est beaucoup plus prononcé lorsque le nombre



(a) (b)

FIGURE 3.3 – (a) Nombre de SNP sélectionnés selon les paramètres appliqués au modèle CCA (degré de pénalisation et nombre de variables initiales). (b) Distribution sous H_0 de la statistique de test considérée (500 permutations).

de SNP présents initialement dans la matrice X est élevé. Avec $\lambda = 0.1$, on atteint un taux de sélection variant de 10 (1000 SNP) à 69 % (100 SNP). En ce qui concerne le degré de significativité de la statistique de test, on constate que celui évolue sensiblement lorsque λ augmente, et ce quel que soit le nombre de SNP fournis en entrée. Lorsqu'aucune pénalisation n'est imposée (CCA classique), la p-valeur dans le cas où X comporte 1000 SNP vaut 0.055. Elle est donc à la limite des 5 % considérés comme notre seuil critique. Dans tous les autre cas, elle se révèle > 0.06.

Quant à la corrélation canonique dont la distribution selon les paramètres choisis, est illustrée dans le tableau suivant, on voit clairement qu'elle augmente lorsque le nombre de variables du bloc X augmente, sauf lorsque la pénalisation est trop importante.

		Degré de pénalisation (λ)			
		0	0.05	0.1	0.15
Variables	100	0.827	0.821	0.804	0.772
	500	0.927	0.918	0.873	0.752
	1000	0.949	0.931	0.834	0.615

Enfin, pour avoir une estimation de la distribution de notre statistique dans le cas général où les prédicteurs ne seraient pas nécessairement effectifs, nous avons également répliqué la même analyse en sélectionnant aléatoirement les SNP sur le génome. Les résultats, reproduits en Annexe (page 46), montrent que les p-valeurs ainsi obtenues sont toujours > 0.2 à l'exception d'une condition (500 SNP, $\lambda = 0.1$). Ceci permet de conforter l'hypothèse que ce type de statistique demeure sensible à la quantité d'information portée par les prédicteurs, même si dans le cas présent, la sélection des meilleures p-valeurs univariées ne permet pas de clairement mettre en évidence la supériorité de la CCA pénalisée en comparaison de la CCA classique.

3.4 Discussion

Les résultats obtenus suggèrent que la pénalisation n'améliore pas nécessairement la qualité du lien entre les blocs X et Y. Or, en pénalisant le bloc de prédicteurs X, on aurait pu s'attendre à améliorer le modèle descriptif (CCA) ou explicatif (PLS) dans la mesure où ceci permet de réduire les facteurs de nuisance susceptibles de nuire à la modélisation du lien inter-blocs. D'autre part, il semble que la régression PLS soit moins performante, selon le critère d'évaluation que nous avons retenu (distance euclidienne assimilée à une erreur de prédiction moyenne).

Deux arguments peuvent être avancés pour expliquer ces écarts à nos prédictions. D'une part, nous avons délibéremment choisi de sélectionner des prédicteurs en fonction de leur degré de significativité dans des tests univariés. Dans ce cas, nous négligeons la corrélation spatiale sur la séquence, observable au travers du déséquilibre de liaison, et ce faisant nous considérons un ensemble de SNP plus ou moins indépendants car situés sur différents chromosomes ou à plus de 50 Mb l'un de l'autre en général. Or, les techniques de modélisation incluant une phase de sélection de variables, comme la PLS ou la CCA, sont plutôt destinées à isoler de bons prédicteurs parmi un ensemble de variables liées à des degrés divers. Ici, il n'existe *a priori* plus de lien modéré ou faible entre les SNP puisque nous n'avons retenu que les 1000 premiers SNP (au sens de leur p-valeur). D'autre part, dans le cas de la PLS, le choix de la distance euclidienne comme mesure de la qualité de la prédiction n'est sans doute pas le plus approprié. Il serait intéressant de comparer les résultats observés avec des mesures plus classiques telles que le RMSE (erreur moyenne utilisée dans un modèle linéaire), qui représente la somme des racines des erreurs quadratiques calculées variable par variable, ou le Q^2_h qui est le rapport entre la variance résiduelle prédite et la la variance résiduelle observée, également sommé sur chaque variable. Ce critère, proposé par Tenenhaus (1998), permet d'évaluer la contribution de chaque composante au modèle PLS (pour plus de détails, consulter Lê Cao et al., 2008, page 11).

Enfin, comme nous avons conservé le même paramètre de pénalisation dans les trois cas de figure (100, 500 ou 1000 variables), nous ne contrôlons pas exactement le taux de sélection des SNP. Par exemple, lorsque nous travaillons en CCA avec 100 SNP, quelle que soit la valeur de λ on ne peut descendre en deça de 50 % (54 SNP en moyenne), alors qu'avec 1000 SNP on peut sélectionner 1 % des SNP seulement. Il serait donc intéressant de travailler en fixant le nombre de SNP sélectionner plutôt que le paramètre de régularisation. C'est d'ailleurs de cette manière que procèdent Lê Cao et al. (2008).

Chapitre 4

Application : approche SNP candidats

Les analyses présentées dans ce chapitre ont en partie motivé l'approche méthodologique développée aux chapitres précédents. Ici, nous nous focalisons sur la recherche d'association spécifique entre les scores d'imagerie recueillis sur 94 sujets et leurs données de génotypage restreintes à environ un millier de SNP sélectionnés sur la base de recherche dans la littérature. Ces SNP sont connus pour être impliqués de près ou de loin dans les processus liés à la lecture ou à la dyslexie.

4.1 Données et hypothèses

Les données de neuroimagerie ont déjà été décrites à la section 2.1.2. Quant aux données génétiques, il s'agit d'une liste de 1438 SNP candidats, répartis dans 5 chromosomes. La sélection de ces SNP a été effectuée sur la base de recherches bibliographiques visant à identifier des SNP ou des gènes potentiellement impliqués dans les troubles de lecture ou la dyslexie. L'objet de ce travail n'est pas d'étudier la qualité de cette sélection mais bien de vérifier si les techniques exposées précédemment nous permettent de mettre en évidence des liens particuliers entre les scores d'imagerie et le profil génotypique des individus sur ces SNP.

Après application des filtres mentionnés à la section 2.1.3, nous disposons d'un ensemble de 952 SNP, soit près des deux tiers, dont la distribution sur le génome est indiquée dans le tableau 4.1.

On notera que les SNP sur les chromosomes 2 et 18 se regroupent en 3 à 4 clusters bien distincts (cf. Tableau 4.1). Quant aux SNP situés sur les chromosomes 3 et 6, ils sont pour la plupart adjacents sur le génome et en forts déséquilibres de liaison (cf. Figure A-7 en Annexe pour une illustration sur le chromosome 6).

4.2 Résultats

4.2.1 Sélection de variables avec variable réponse univariée

LASSO. La régression régularisée par la méthode LASSO est effectuée en considérant les données de génétique et les scores des individus sur la première composante principale issue de l'ACP

chromosome	15	18	2	3	6
n	46	251	402	180	73
f	0.05	0.26	0.42	0.19	0.08

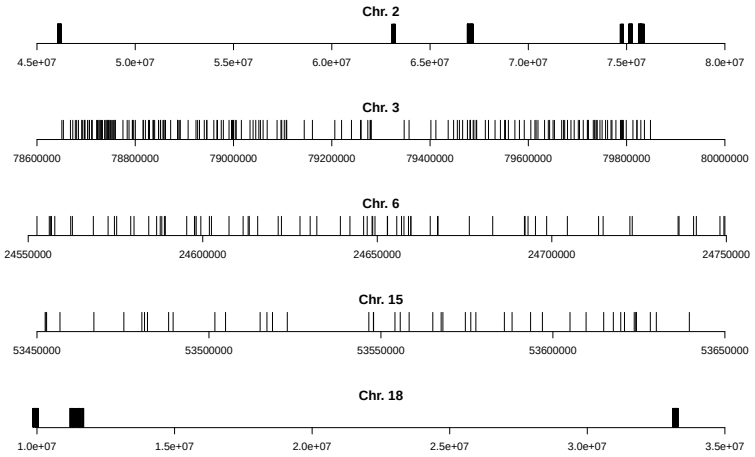


TABLE 4.1 – Répartition des SNP candidats entre les chromosomes (nombre et fréquence). Leur distribution spatiale est schématisée pour chaque chromosome, dans une fenêtre limitée aux positions les plus extrêmes.

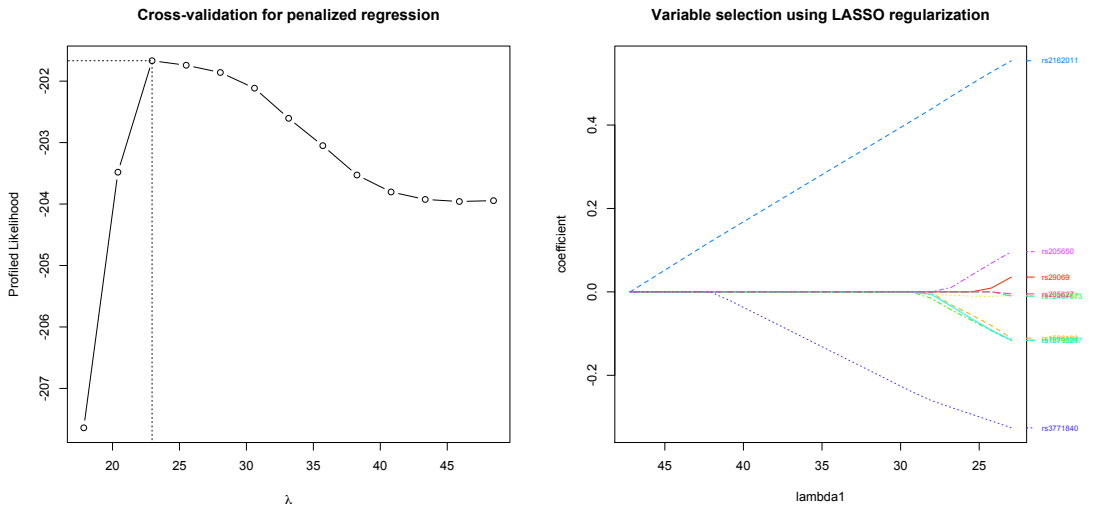
SNP	β_{LASSO}	univ. p-valeurs	ROI min p	chr.	position	gène	locus
rs29069	0.035	[0.010–0.999]	Speech (-42,-42,-21)	18	9945960	VAPA	9218
rs1566198	-0.111	[0.003–0.940]	Read (-57,-39,3)	18	11249831	–	–
rs1420358	-0.009	[0.042–0.940]	Speech (-52,64,-6)	2	67116062	–	–
rs7579521	-0.117	[0.013–0.988]	Read (-51,-3,45)	2	67120211	–	–
rs12467673	-0.010	[0.082–0.993]	Speech (-48,0,51)	2	46162233	PRKCE	5581
rs12712967	-0.116	[0.004–0.975]	Read (-51,-3,45)	2	46176109	PRKCE	5581
rs2162011	0.554	[0.001–0.938]	Speech (-55,-62,16)	2	46176109	OTX1	5013
rs3771840	-0.326	[0.014–0.943]	Speech (-52,-60,-14)	2	75239139	TACR1	6869
rs205650	0.096	[0.001–0.953]	Read (-45,-56,-18)	2	74738119	SEMA4F	10505
rs205627	-0.005	[0.003–0.994]	Read (-48,-63,-12)	2	74817176	–	–

TABLE 4.2 – Résumé des SNP sélectionnés par la régression LASSO.

(§ 2.1.2, page 15). Ceci nous permet de nous positionner dans un cadre à peu près comparable à la régression PLS et à la CCA dans la mesure où l'on travaille directement avec une variable latente.

La figure 4.1 (a) montre l'évolution de la vraisemblance en fonction des valeurs du paramètre de régularisation à l'issue d'une procédure de validation croisée 10-fold.

On voit que la régression régularisée nous amène à retenir 10 SNP localisés sur les chromosomes 2 et 18 (Figure 4.1, b) dont le résumé est fourni dans le tableau 4.2. Les p-valeurs indiquées correspondent aux tests univariés effectués sur le génome entier (§ 2.2) ; en l'occurence, nous avons considéré l'étendue des p-valeurs estimées sur les 34 régions. Nous avons également indiqué les gènes associés identifiés à partir des information contenues dans la base de données dbSNP (`www.ncbi.nlm.nih.gov/projects/SNP`).



(a) (b)

FIGURE 4.1 – (a) Optimisation du critère basé sur la vraisemblance « profilée » en fonction du paramètre de régularisation. (b) Profil de la procédure de sélection de variables et coefficients de régression associés en fonction du paramètre de régularisation.

Forêts aléatoires. L'algorithme utilisé pour construire les modèles par forêts aléatoires est celui de Breiman and Cutler (voir Breiman, 2001), proposé dans le package R **randomForest**. Le nombre de variables sélectionnées aléatoirement comme variables candidates à chaque coupure de l'arbre est fixé à $p/3$ (valeur `mtry` par défaut dans un cadre de régression) et le nombre d'arbres à construire est de 500. La taille des sous-échantillons générés par bootstrap vaut $n = 94$ et les effectifs minimums par nœuds terminaux sont fixés à 5.

La variance résiduelle (MSE) est estimée à 4.562 et le nombre de variables testées à chaque coupure est de 317 (pour chacun des 500 arbres de régression). La figure 4.2 montre la contribution relative des 30 meilleurs prédicteurs dans le modèle final [1]. Les SNP rs2162011 et rs11125947 apparaissent les meilleurs indicateurs si l'on se réfère à la part de variance expliquée ou, de manière équivalente, l'augmentation relative de la résiduelle lorsque ceux-ci ne sont pas présents dans l'arbre de régression (graphique de gauche).

Si l'on augmente le nombre d'arbres à 1000 et que l'on spécifie 400 variables pour la sélection par arbre, la résiduelle vaut alors 4.622 ne suggérant aucune amélioration sensible.

4.2.2 Sélection de variables avec variable réponse multivariée

PLS. Le choix des paramètres optimaux pour la régression PLS régularisée (version *sparse*) résulte également d'une procédure de validation croisée 10-fold (Chun and Keleş, 2007). On cherche alors à minimiser l'erreur moyenne de prédiction, tout en faisant varier les deux paramètres de

1. L'importance d'un prédicteur j est estimée à partir de $\hat{\theta}_j = \frac{1}{B}\sum_{b=1}^{B} \delta_{bj}$ pour chaque échantillon bootstrap $b = 2, \ldots, B$, avec $\delta_{bj} = \pi_{bj} - \pi_b$ où π_b désigne l'impureté de l'arbre et π_{bj} l'impureté de l'arbre après permutation des x_j ($j = 1, \ldots, p$). Dans le contexte des RF, $\hat{\theta}_j$ s'assimile à un score d'importance dans la mesure où il reflète l'effet réel du prédicteur en comparaison de son effet sous randomisation. Une mesure naturelle de l'importance du prédicteur j est alors $\hat{\theta}_j / SE(\hat{\theta}_j)$ (%IncMSE dans le graphique).

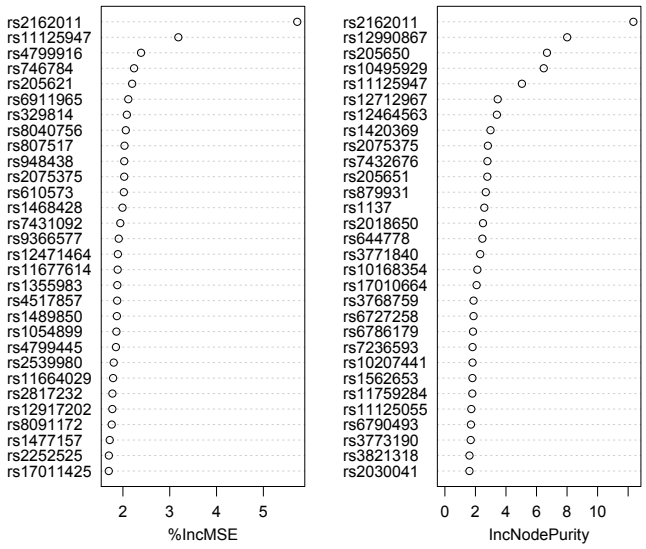


FIGURE 4.2 – Mesures d'importance des variables.

régularisation du modèle (*eta*, pour le seuillage sur X et K pour le nombre de composantes à retenir sur le bloc Y). À la lecture de la carte représentée dans la figure 4.3, on constate que les paramètres optimaux sont $\eta = 0.8$ avec 2 composantes (paramètre K).

Le modèle PLS sélectionne 6 SNP parmi les 952 prédicteurs : rs4799445, rs8094149, rs1539847, rs3771840, rs10207441, rs3821318. On peut constater que rs3771840 figure parmi la liste des SNP sélectionnés par la régression LASSO et qui est localisé dans le gène TACR1 (tachykinin receptor 1) qui est chargé d'encoder le récepteur de la neurokinin 1 (substance tachykinin P). L'effet de chacun de ces prédicteurs sur les scores d'imagerie est représenté dans la figure 4.4, associés à des intervalles de confiance à 95 % estimés par une procédure bootstrap (1000 rééchantillonnages). On constate que leur effet est assez localisé sur certaines régions d'intérêt puisqu'une grande majorité des IC recouvrent la valeur 0 (absence d'effet). Plus spécifiquement, les effets les plus marqués apparaissent sur la région 34 pour les SNP rs4799445 et rs1539847.

Si maintenant l'on ne considère que les deux premières composantes et que l'on fixe le nombre de variables à retenir dans le modèle à 20 (sur chaque composante), au lieu de fixer directement le critère de régularisation, comme proposé par Lê Cao et al. (2008), on retrouve des résultats quelque peu différents. Toutefois, trois des six SNP précédement identifiés – rs3771840, rs10207441 et rs3821318 – sont toujours présents parmi les meilleurs prédicteurs en utilisant une pénalisation de type LASSO plutôt que elasticnet. Les SNP les plus influents sur les variations des scores de neuroimagerie sont indiqués dans la figure 4.5. La mesure d'influence retenue est ici l'importance relative de chaque prédicteur. Les prédicteurs sélectionnés sur les deux composantes sont symbolisés en rouge.

Les 20 variables ayant des charges non nulles sur les deux axes factoriels sont représentées sur un cercle des corrélations dans la figure 4.8 (a). Il apparaît clairement que le second axe factoriel est principalement déterminé par une opposition entre deux clusters de SNP.

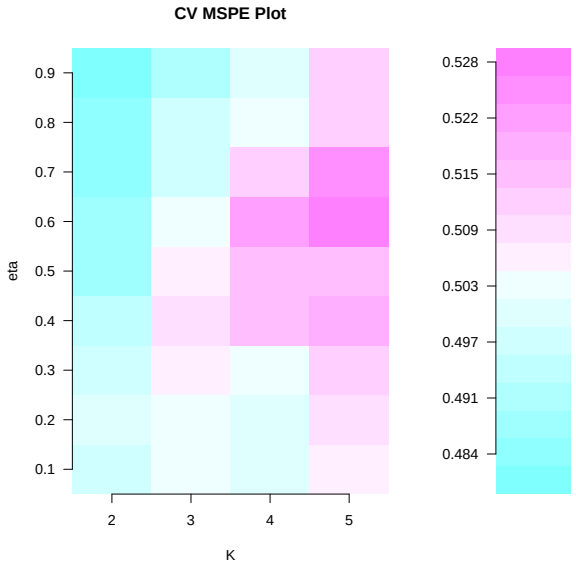


FIGURE 4.3 – Valeur moyenne de l'erreur de prédiction selon les paramètres de seuillage (η) et le nombre de composantes (K), avec une procédure de validation croisée 10-fold.

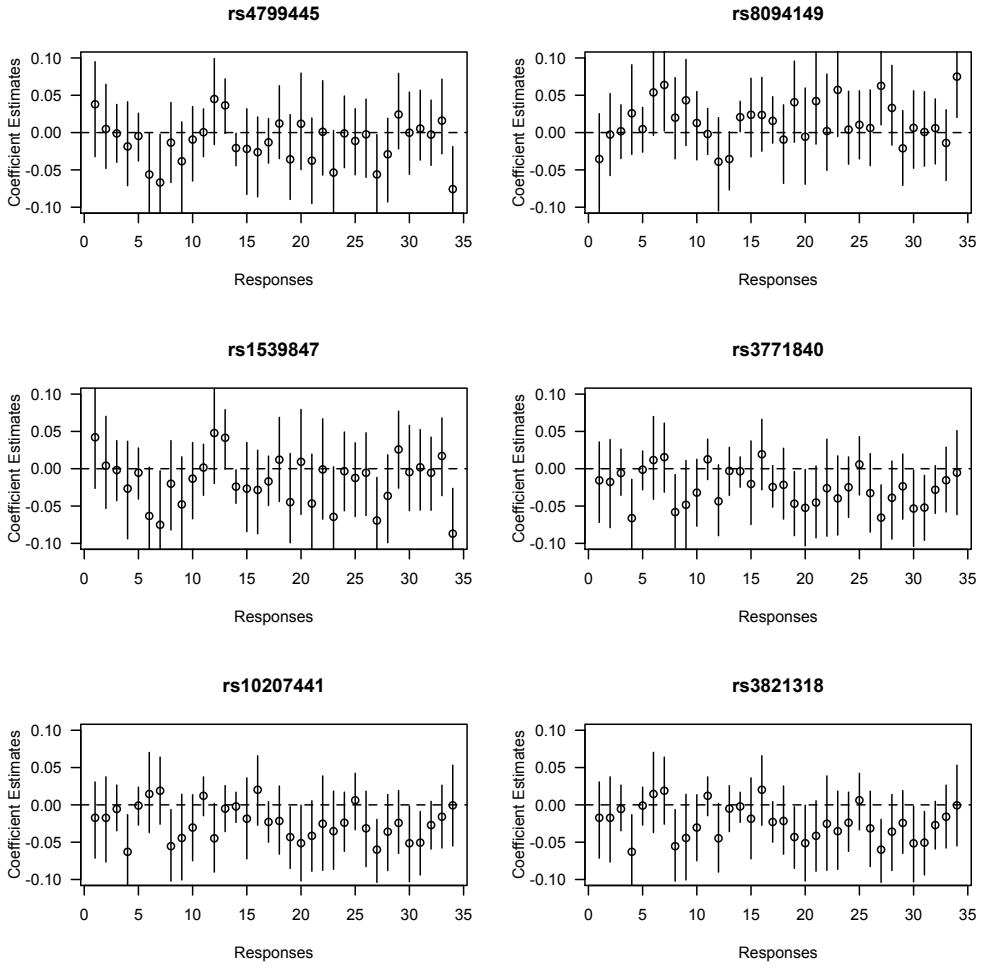


FIGURE 4.4 – Effet de chaque prédicteur retenu par le modèle PLS sur l'ensemble des variables réponses (axe horizontal), avec des intervalles de confiance à 95 %.

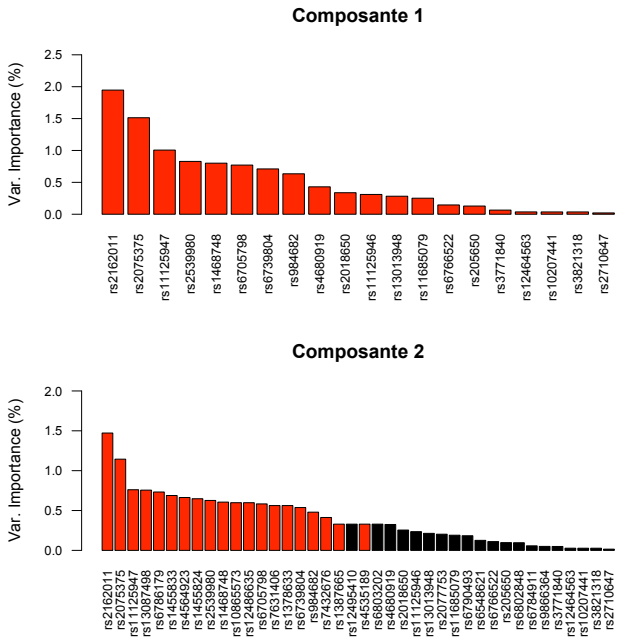


FIGURE 4.5 – Importance relative (en %) des SNP sélectionnés sur les deux premières composantes dans un modèle PLS avec régularisation de type LASSO.

CCA. Dans le cadre de la CCA, nous utilisons directement l'algorithme proposé par Parkhomenko et al. (2009). Celui-ci est basé sur une décomposition en valeurs singulières et consiste à appliquer une pénalisation de type LASSO durant l'étape d'estimation des vecteurs canoniques. Cette pénalisation peut être effective sur les deux blocs de données (ce qui revient à régulariser les vecteurs singuliers droit et gauche de la SVD), mais dans notre cas nous avons délibérement opté pour la pénalisation du seul bloc de données de génétique (X).

La figure 4.7, qui représente les matrices de corrélation des deux blocs, ainsi que la matrice de covariance, permet d'apprécier le déséquilibre de liaison dans le bloc des prédicteurs X qui se manifeste par des blocs de forte corrélation entre SNP adjacents. En revanche, la matrice de covariance ne laisse pas apparaître de lien fort entre les deux blocs de données.

Avec un paramètre de régularisation sur X fixé à 0.19 [2], la corrélation canonique vaut 0.614, et le modèle sélectionne 18 SNP dont le poids relatif dans la composante est illustré dans la figure 4.6. Parmi ceux-ci, les SNP rs1539847 et rs4799445 ont déjà été sélectionnés dans le modèle PLS (cf. page 28). Ce ne sont pourtant pas les SNP ayant le plus fort poids dans le vecteur canonique. Ce résultat est donc intéressant en soi puisque cela signifie que la CCA permet d'isoler certains SNP en commun avec la PLS sans que ceux-ci contribuent le plus à l'explication de leur propre bloc.

Si l'on n'impose pas de pénalisation sur les prédicteurs [3], et que l'on s'intéresse uniquement aux prédicteurs les mieux corrélés avec leurs propres composantes, en d'autres termes les variables

2. La valeur choisie pour le paramètre de pénalisation ne résulte pas d'une estimation par validation croisée, telle que celle appliquée par Parkhomenko et al. (2007).

3. Dans ce cas de figure, il y a toujours une régularisation des données, pour éviter les problèmes numériques lors de l'inversion de la matrice X^tX. Cette régularisation consiste à considérer $X^tX + \lambda I$, où I est une matrice identité $p \times p$, à la place de X^tX lorsque l'on procède à l'inversion de la matrice.

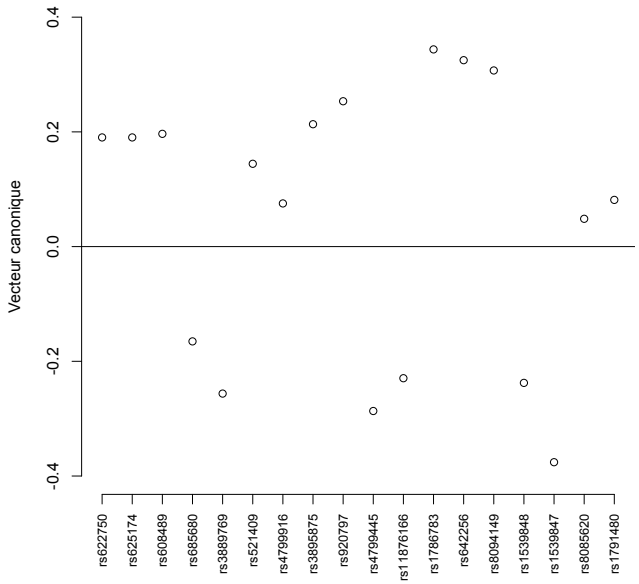


FIGURE 4.6 – Poids relatif des SNP (vecteur canonique) dans le modèle de CCA *sparse* avec un paramètre de régularisation fixé à $\lambda = 0.19$.

qui expliquent le mieux leur propre bloc, on a les résultats illustrés dans la figure 4.8 (b). Dans ce cas de figure, on ne retrouve pas les mêmes SNP, mais cette représentation présente l'avantage de permettre la visualisation des associations entre régions de neuroimagerie et SNP. En l'occurence, on voit apparaître des clusters de SNP corrélant fortement avec des clusters d'imagerie, avec une opposition entre les tâches de lecture et de parole. Ceci est moins visible dans le cas de la PLS (Figure 4.8, a).

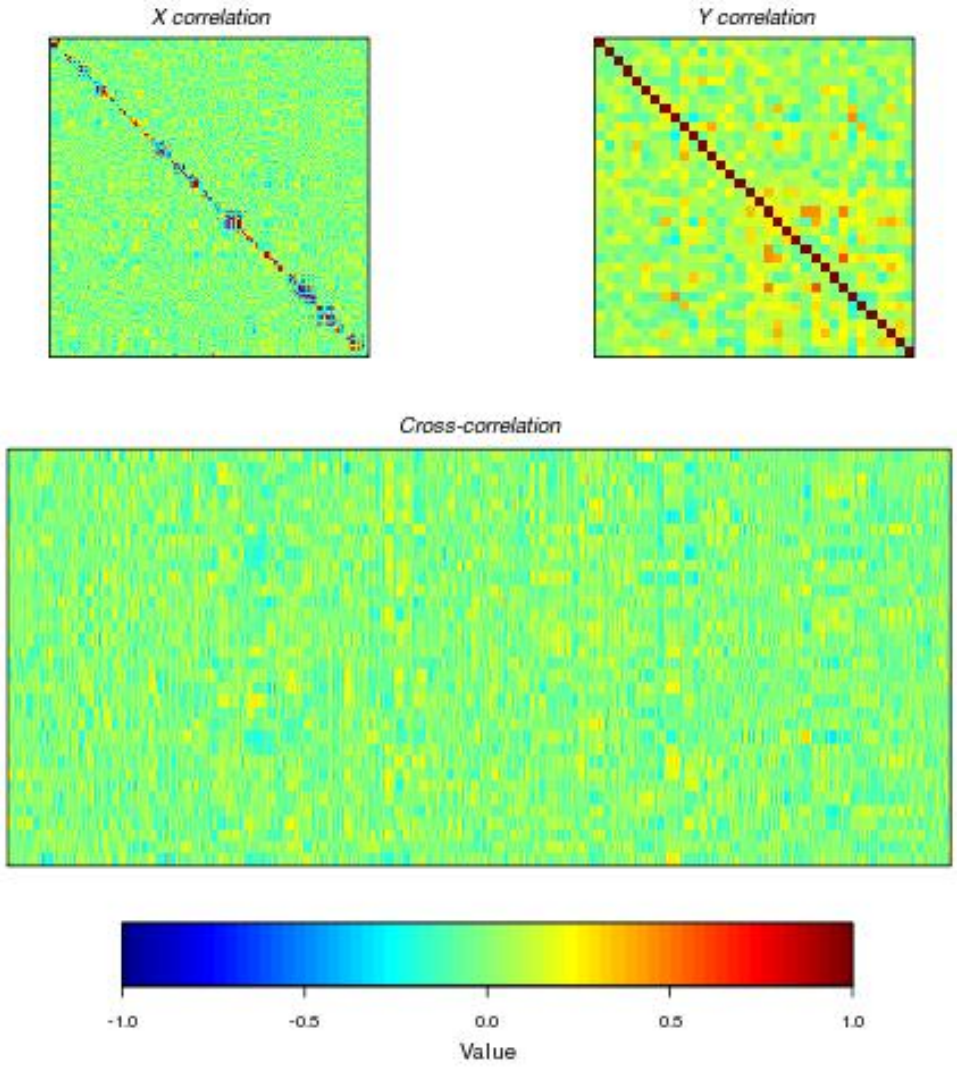
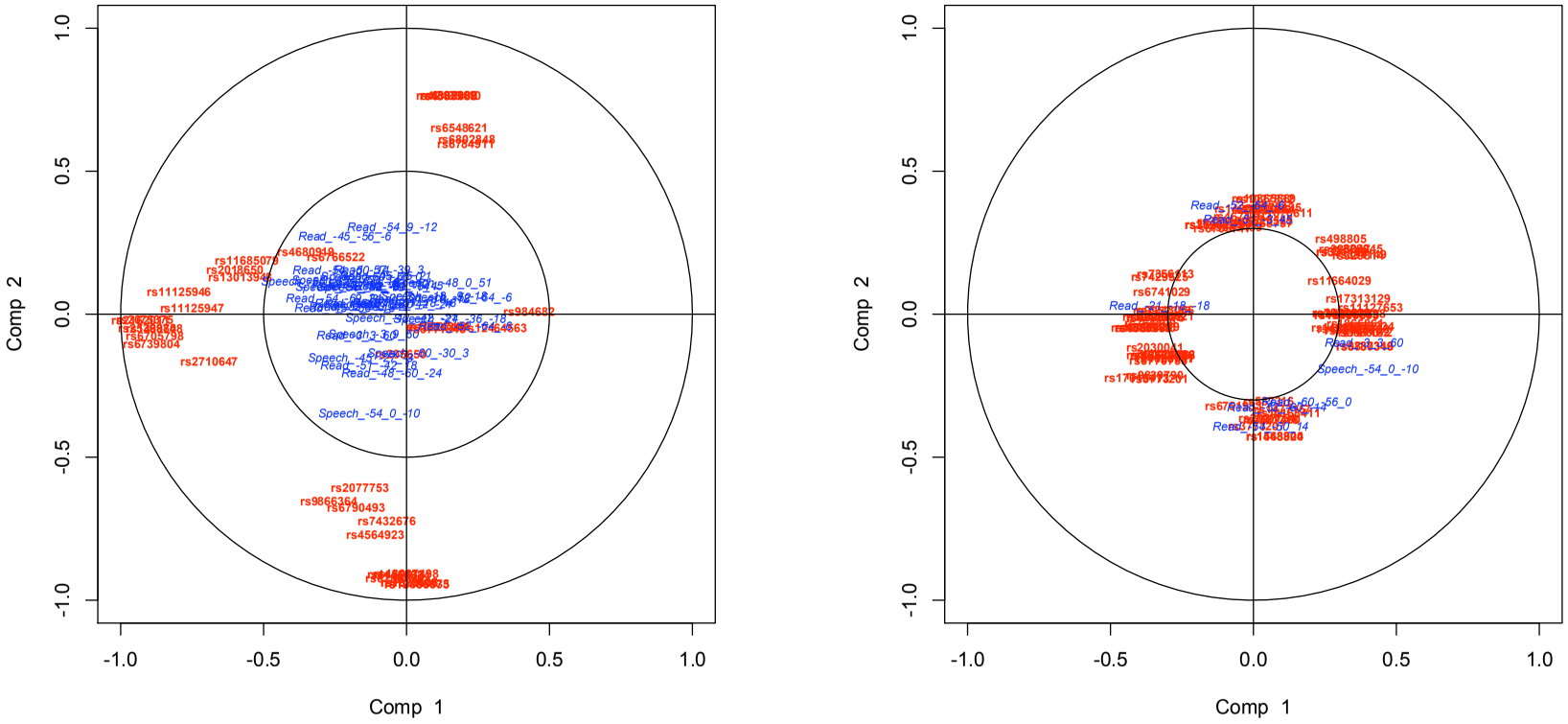


FIGURE 4.7 – Matrices de corrélation X^tX (952×952), Y^tY (34×34), et Y^tX (34×952).



(a) (b)

FIGURE 4.8 – Représentation des variables des blocs X et Y dans le cercle de corrélation pour (a) la PLS et (b) la CCA. Afin de faciliter la lecture, seules les variables sélectionnées ($k = 20$) sont indiquées dans le cas de la PLS. Dans le cas de la CCA, seules les variables (SNP et imagerie) pour lesquelles la corrélation des variables est > 0.35 sont affichées.

4.3 Synthèse des résultats

Nous avons appliqué quatre techniques de modélisation différentes : régression LASSO et forêts aléatoires en considérant comme variable réponse les scores des individus dérivés par une ACP ; régression PLS et CCA pénalisées en considérant l'ensemble des scores de neuroimagerie. Les 10 meilleurs SNP, en termes d'importance relative ou de coefficient de régression (en valeur absolue), sont résumés dans le tableau 4.3.

LASSO	RF	PLS	CCA
rs29069	rs2162011	rs4799445	rs1539847
rs1566198	rs11125947	rs8094149	rs1786783
rs1420358	rs4799916	rs1539847	rs642256
rs7579521	rs746784	rs3771840	rs8094149
rs12467673	rs205621	rs10207441	rs4799445
rs12712967	rs6911965	rs3821318	rs3889769
rs2162011	rs329814		rs920797
rs3771840	rs8040756		rs1539848
rs205650	rs807517		rs11876166
rs205627	rs948438		rs3895875

TABLE 4.3 – Liste des SNP prédicteurs identifiés par différentes méthodes.

Bien que les résultats ne convergent pas tous vers la mise en évidence d'un sous-ensemble unitaire de SNP en association avec les modulations des activités cérébrales dans nos régions d'intérêt, il apparaît que certains SNP sont tout de même « détectés » par au moins deux techniques différentes. ceci est intéressant car comme le rappellent Liang and Kelemen (2008, p. 48), la reproductibilité des résultats dans les études d'association est souvent affectée par la diversité des techniques utilisées.

Ceci est d'autant plus intéressant car dans le cas de la régression LASSO et de l'approche par forêts aléatoires, nous considérons une variable composite qui, comme on l'a vu au chapitre 2 (page 15), représente moins de 20 % de la variance de l'ensemble des scores de neuroimagerie. Or, dans la régression PLS et la CCA, nous avons délibérement choisi de ne pas pénaliser le bloc Y. Ceci suggère que le score composite retenu permet déjà de réduire la dimension du bloc Y et que les techniques utilisant une seule variable y restent applicables (e.g., régression logique). En l'occurence, dans le cadre de la PLS, les scores des individus sur la première composante du bloc Y sont relativement bien corrélés (en valeur absolue) avec ceux d'une ACP classique ($r = 0.823$, avec un IC à 95 % de $[0.843; 0.928]$). Pour la CCA, la corrélation est beaucoup plus modérée bien que toujours significativement différente de zéro ($r = 0.333$, $t(92) = 3.388$, $p = 0.001$, IC 95 % $[0.140; 0.502]$). ceci pourrait éventuellement expliquer le fait que le SNP rs3771840 n'est pas détecté par la CCA, contrairement à la PLS.

En conclusion, chacune des méthodes utilisées présente ses avantages et ses inconvénients, tant du point de vue du type de prédicteur qu'il est possible de considérer (catégoriel *vs.* numérique) que de la possibilité offerte de visualiser les résultats et évaluer la qualité du modèle appliqué. Même s'il semble difficile d'isoler le meilleur sous-ensemble de SNP qui permettrait de résumer au mieux les liens entre les variations génétiques et les endophénotypes recueillis chez ces 94 sujets,

nos résultats nous confortent dans l'idée que ce type d'approche multivariée permet d'aborder de manière satisfaisante les relations entre génétique et neuroimagerie. D'un autre côté, on peut souligner un clair manque de puissance dans ces résultats du fait que l'on utilise le même échantillon pour sélectionner les variables d'intérêt et évaluer la qualité du modèle résultant. Ceci devrait faire l'objet d'une étude de validation croisée telle que celle décrite au chapitre précédent.

Chapitre 5

Conclusions et perspectives

Bien que notre étude de la PLS et de la CCA pénalisées mériterait d'être étendue dans différentes directions, elle suggère tout de même qu'un simple filtrage univarié de variables candidates ne permet pas vraiment d'isoler par la suite un bon sous-ensemble de SNP prédicteurs lorsque l'on considère des données recueillies dans une expérience de neuroimagerie et des données de génotypage. Ceci n'est visiblement pas le cas lorsque l'on considère des données transcriptomiques (Lê Cao et al., 2008) ou des interactions gènes/SNP (Parkhomenko et al., 2007).

Le choix du filtrage univarié et du nombre de variables considérées dans l'approche multivariée pourrait être remis en question. S'il est plus ou moins comparable à ce que l'on retrouve dans d'autres études (e.g., van't Veer et al., 2002; Furlanello et al., 2003), cela pose le problème de l'élimination de SNP sans prise en compte de leurs potentielles interactions avec d'autres SNP, interactions qui se manifestent notamment au travers du déséquilibre de liaison comme nous l'avons déjà mentionné. Díaz-Uriarte and Alvarez de Andrés (2006) ont formulé la même remarque au sujet des études de classification de gènes.

Par conséquent, il semble nécessaire d'explorer plus en avant ces deux méthodes du point de vue de leur intégration avec une procédure de sélection de variables. Comme nous l'avons évoqué en introduction, nous avons définitivement opté pour une approche par filtrage univarié alors que des méthodes d'ensemble seraient sans doute plus appropriées dans ce contexte. Par ailleurs, il faut souligner que notre procédure d'estimation de la distribution sous H_0 de nos statistiques de test est biaisée dans la mesure où les permutations devraient être effectuées au niveau de la boucle externe de sélection des SNP, et non pas dans la boucle interne à la validation croisée par « leave-one-out ». En effet, dans notre cas, nous ne nous prémunissons pas d'un éventuel risque de sur-ajustement.

D'autre part, nous avons délibéremment choisi de fixer des paramètres variables pour la pénalisation des composantes de X. *A priori*, les résultats semblent plutôt indiquer que la pénalisation n'entraîne pas de bénéfice clair par rapport à des CCA ou des PLS classiques. Comme le soulignent Lê Cao et al. (2008), en travaillant avec de petits échantillons, l'estimation des bons paramètres de pénalisation peut être délicate. Il est donc possible que nous n'ayons pas choisi les paramètres optimaux et que, du simple fait d'avoir sélectionné les 100 à 1000 SNP les plus significatifs, la pénalisation ne soit pas nécessaire. Pour vérifier la sensibilité de ces techniques

pénalisées au nombre de paramètres, il est envisageable de travailler avec un plus grand nombre de prédicteurs dès le départ. Ceci nous rapprocherait d'ailleurs des études d'association « restreinte » où plusieurs milliers de SNP sont étudiés en même temps.

Quant à notre application de différentes techniques multivariées sur une liste de SNP candidats, nos résultats suggèrent que, malgré les divergences, elles sont à mêmes de recouvrer l'effet de certains SNP alors même qu'une approche massivement univariée n'indique aucune association significative. Bien entendu, la même remarque que celle précédemment formulée au sujet du filtrage de variables d'intérêt peut être formulée : la sélection univariée ne tient pas compte des interactions spatiales locales entre SNP. Une solution dans ce cas consiterait à adapter la technique de « fenêtre glissante » proposée par Neale and Sham (2004) dans le cadre des études cas-témoins : les p-valeurs de plusieurs tests univariés sont combinées ensemble en considérant comme distribution de référence $\chi^2 = -2\sum_{i=1}^{m}\log(p_i) \sim \chi^2_{2m}$, où m désigne le nombre de SNP testés dans la même fenêtre spatiale et p_i la p-valeur du test d'association. Cette approche permet d'isoler des régions chromosomiques d'intérêt plutôt que des SNP isolés (Dudbridge and Koeleman, 2004). Toutefois, cette approche ne tient pas compte de la distance entre les SNP sur la séquence, ce qui en rend délicate l'application dans une approche SNP candidats lorsque les gènes associés sont répartis sur plusieurs chromosomes. Liang and Kelemen (2008, p. 47–48) passent en revue d'autres possibilités, comme la statistique scan qui tient compte de la distance inter-SNP et de leur distribution sur les chromosomes (Hoh and Ott, 2000).

En termes de perspectives, il nous reste encore à :
- redéfinir le filtrage des SNP candidats, en tenant compte par exemple des interactions entre SNP et scores de neuroimagerie grâce à une procédure de type GLASSO (Friedman et al., 2008) ;
- évaluer les méthodes employées dans la partie applicative – régression LASSO et forêts aléatoires – à l'aide d'une procédure de validation croisée pour mieux comparer leurs résultats avec ceux de la CCA et de la PLS pénalisées ;
- définir un meilleur critère d'erreur pour la PLS pénalisée ;
- étudier, grâce à un modèle génératif, dans quelle mesure ces méthodes peuvent s'appliquer à un nombre de variables X et Y plus grand ;
- optimiser les procédures de validation croisée et les interfacer avec les packages R existants.

Bibliographie

Bauer, E. and Kohavi, A. (1999). An empirical comparison of voting classification algorithms : Bagging, boosting, and variants. *Machine Learning*, 36(1-2) :105–139.

Bo, T. and Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4) :0017.1–0017.11.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2) :123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32.

Breiman, L., Friedman, R. A., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Inc.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K., Hayward, B., Keith, T., and Van Eerdewegh, P. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, 28 :171–182.

Chun, H. and Keleş, S. (2007). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. `http://www.stat.wisc.edu/~keles/Papers/spls_jrssb.pdf`.

Clayton, D. and Cheung, H.-T. (2007). An r package for analysis of whole-genome association studies. *Human Heredity*, 64 :45–51.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3).

Dudbridge, F. and Koeleman, B. P. C. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics*, 75(3) :424–435.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*. Springer.

Dupuy, A. and Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of The National Cancer Institute*, 99 :147–157.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2) :407–499.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25) :14863–14868.

Foulkes, A. S. (2009). *Applied Statistical Genetics with R*. Springer.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441.

Furlanello, C., Serafini, M., Merler, S., and Jurman, G. (2003). An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks*, 16 :641–648.

Glahn, D. C., Thompson, P. M., and Blangero, J. (2007). Neuroimaging endophenotypes : Strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, 28 :488–501.

Guerra, R. and Yu, Z. (2006). Single nucleotide polymorphisms and their applications. In Zhang, W. and Shmulevich, I., editors, *Computational and Statistical Approaches to Genomics*, chapter 16, pages 311–349. Springer.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A., editors (2006). *Feature Extraction : Foundations And Applications*. Springer-Verlag.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. N. (2002). Gene selection for cancer classification using Support Vector Machines. *Machine Learning*, 46(1-3) :389–422.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer.

Hastie, T., Tibshirani, T., Eisen, M. B., Alzadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. O. (2000). 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2) :1–21.

Hoerl, A. and Kennard, R. (1988). Ridge regression. In *Encyclopedia of Statistical Sciences*, volume 8, pages 129–136. Wiley.

Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences*, 97 :9615–9617.

Inza, I., Sierra, B., Blanco, R., and Larranaga, P. (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent and Fuzzy Systems*, 12(1) :25–34.

Ioannidis, J. P. A., Thomas, G., and Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, 10 :318–329.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer.

Kooperberg, C., Ruczinski, I., LeBlanc, M., and Hsu, L. (2005). Identifying interacting snps using Monte Carlo logic regression. *Genetic Epidemiology*, 28(2) :157–170.

Lal, T. N., Chapelle, O., Weston, J., and Elisseeff, A. (2006). Embedded methods. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A., editors, *Feature Extraction : Foundations And Applications*, pages 137–162. Springer-Verlag.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Liang, Y. and Kelemen, A. (2008). Statistical advances and challenges for analyzing correlated high dimensional snp data in genomic study for complex diseases. *Statistics Surveys*, 2 :43–60.

Liu, B. H. (1997). *Statistical Genomics : Linkage, Mapping, and QTL Analysis*. CRC Press.

Long, A., Mangalam, H., Chan, B., Tolleri, L., Hatfield, G., and Baldi, P. (2001). Improved statistical inference from dna microarray data using analysis of variance and a Bayesian statistical framework. *Journal of Biological Chemistry*, 276 :19937–19944.

Lunetta, K. L., Hayward, L. B., Segal, J., and Eerdewegh, P. V. (2004). Screening large-scale association study data : Exploiting interactions using random forests. *BMC Genetics*, 5(32).

Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., Zismann, V. L., Joshipura, K., Huentelman, M. J., Hu-Lince, D., Coon, K. D., Craig, D. W., Pearson, J. V., Holmans, P., Heward, C. B., Reiman, E. M., Stephan, D., and Hardy, J. (2007). A survey of genetic human cortical gene expression. *Nature Genetics*, 39(12) :1494–1499.

Neale, B. and Sham, P. (2004). The future of association studies : Gene-based analysis and replication. *American Journal of Human Genetics*, 75 :353–362.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings*, 1(1) :S119.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink : a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81(3) :559–575.

Quinlan, R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufman.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12 :475–511.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004). Exploring interactions in high dimensional genomic data : An overview of logic regression, with applications. *Journal of Multivariate Analysis*, 90 :178–195.

Storey, J. (2003). The positive false discovery rate : A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6) :2013–2035.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeilis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307).

Tenenhaus, M. (1998). *La régression PLS : théorie et pratique.* Editions Technip.

The International HapMap, Consortium (2003). The international hapmap project. *Nature*, 426(6968) :789–796.

Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology.* Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58 :267–288.

Tusher, V., Tinshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9) :5116–5121.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415 :530–536.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

Witten, D. M., Tibshirani, R. J., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3) :515–534.

Zou, H. and Hastie, T. (2005). Regression and variable selection via the elastic net. *Journal of the Royal Statistical Society, B*, 67 :301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15 :265–286.

Annexes

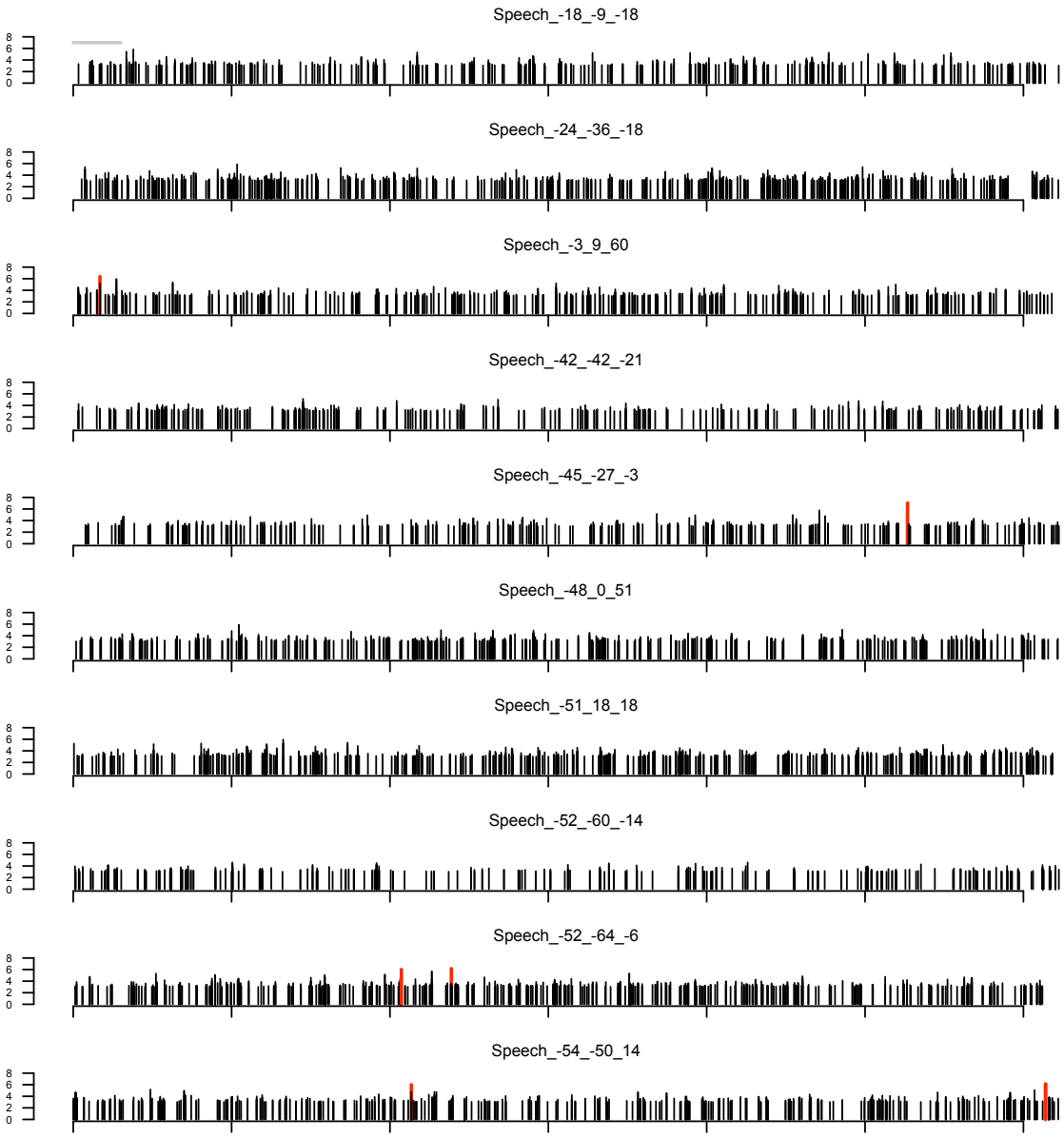


FIGURE A-1 – Carte de p-valeurs ($-\log_{10}$) calculées sur l'ensemble des SNP pour chaque région d'intérêt. Le trait horizontal gris dans le panneau supérieur représente 30000 SNP. Seules les p-valeurs $< 10^{-3}$ sont représentées et les p-valeurs $< 10^{-6}$ sont surlignées en rouge. (1)

FIGURE A-2 – cf. A-1. (2)

FIGURE A-3 – cf. A-1. (3)

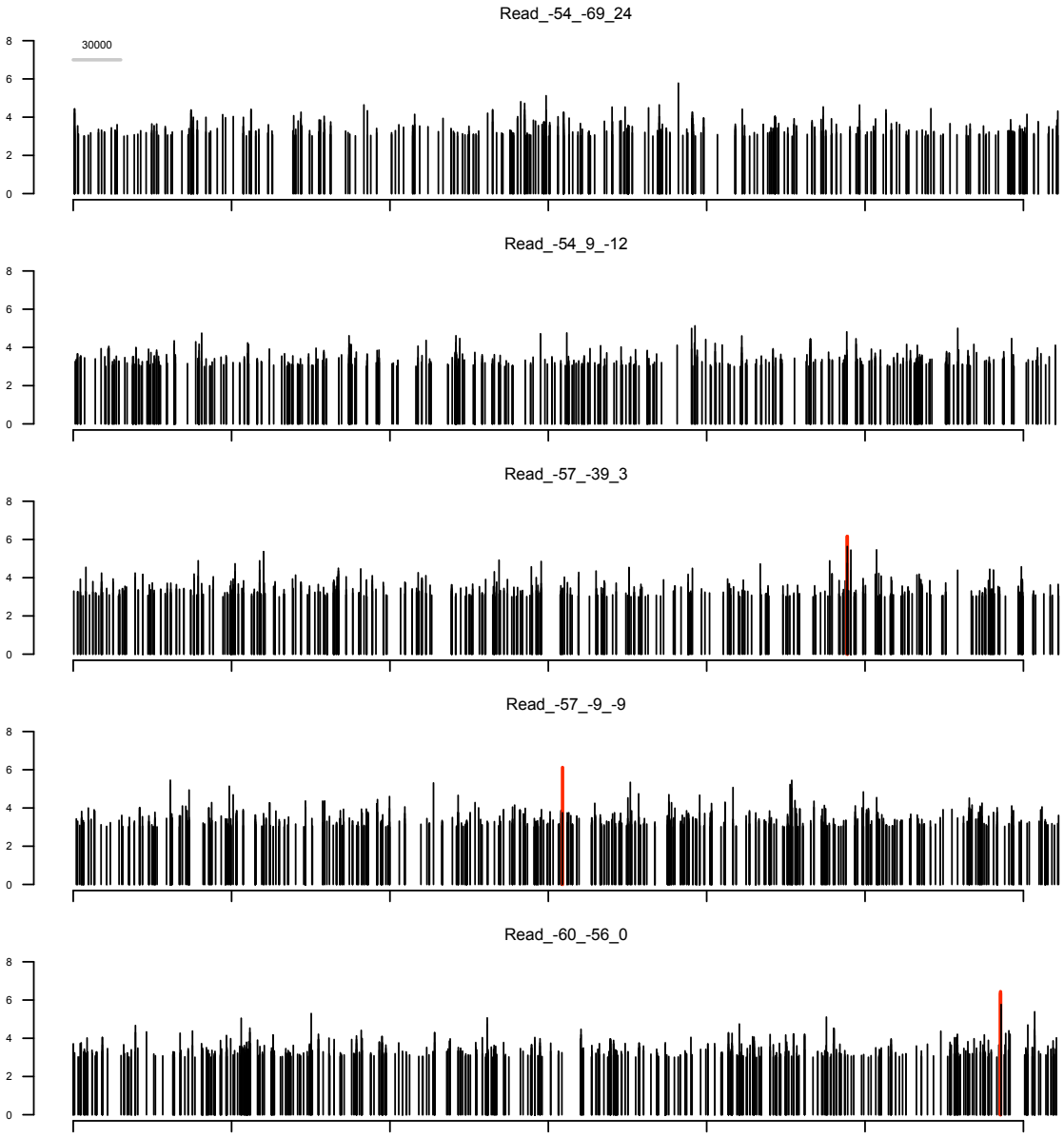
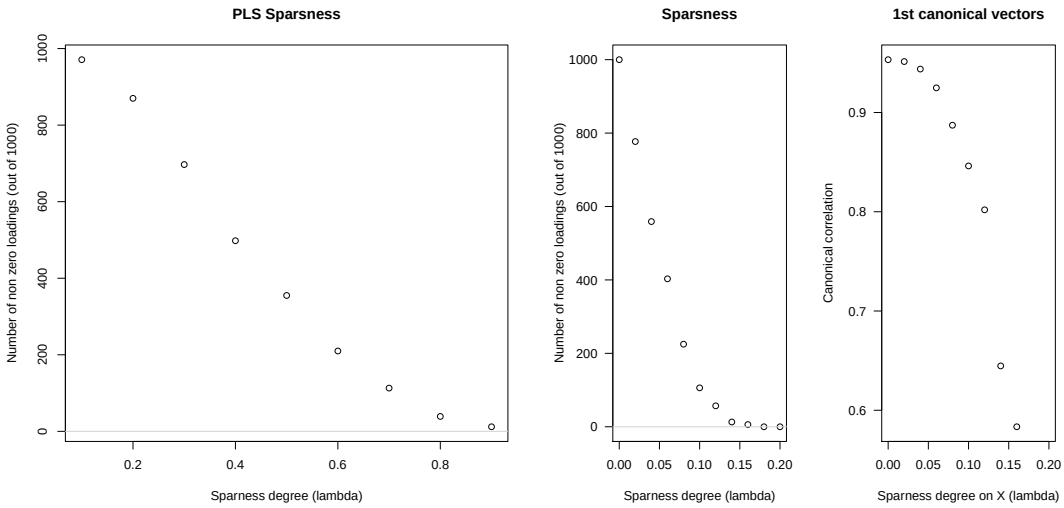
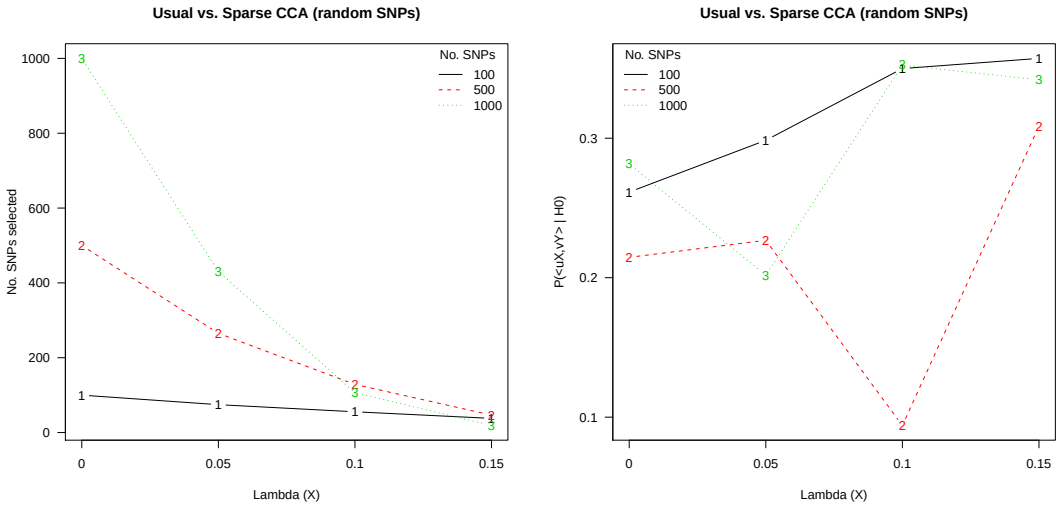


FIGURE A-4 – cf. A-1. (4)



(a) (b)

FIGURE A-5 – (a) Nombre de SNP sélectionnés par la méthode PLS en fonction du critère de pénalisation (λ). (b) Idem pour la CCA.



(a) (b)

FIGURE A-6 – (a) Nombre de SNP sélectionnés selon les paramètres appliqués au modèle CCA (degré de pénalisation et nombre de variables initiales), en considérant des SNP répartis aléatoirement sur la séquence. (b) Distribution sous H_0 de la statistique de test considérée (500 permutations), à comparer aux résultats de la figure 3.3, page 23.

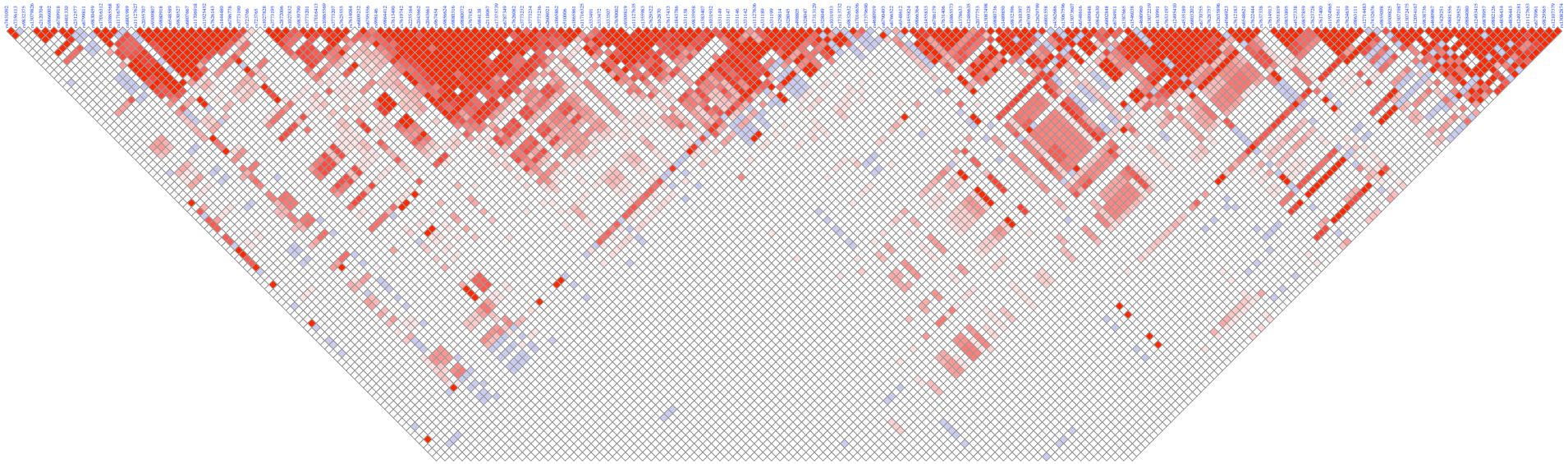


FIGURE A-7 – Carte de déséquilibre de liaison pour 180 SNP du chromosome 6 (position 24552513 à 24749584), avec une profondeur de 100.

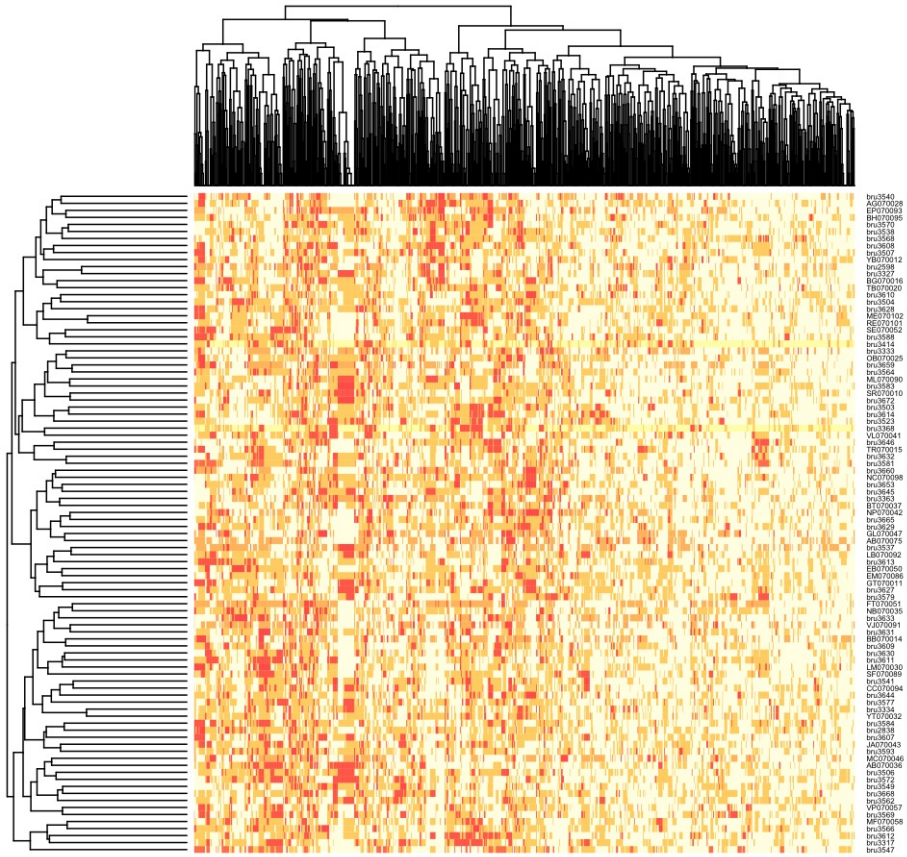


FIGURE A-8 – Double classification des individus ($n = 94$) et des SNP candidats ($q = 952$).

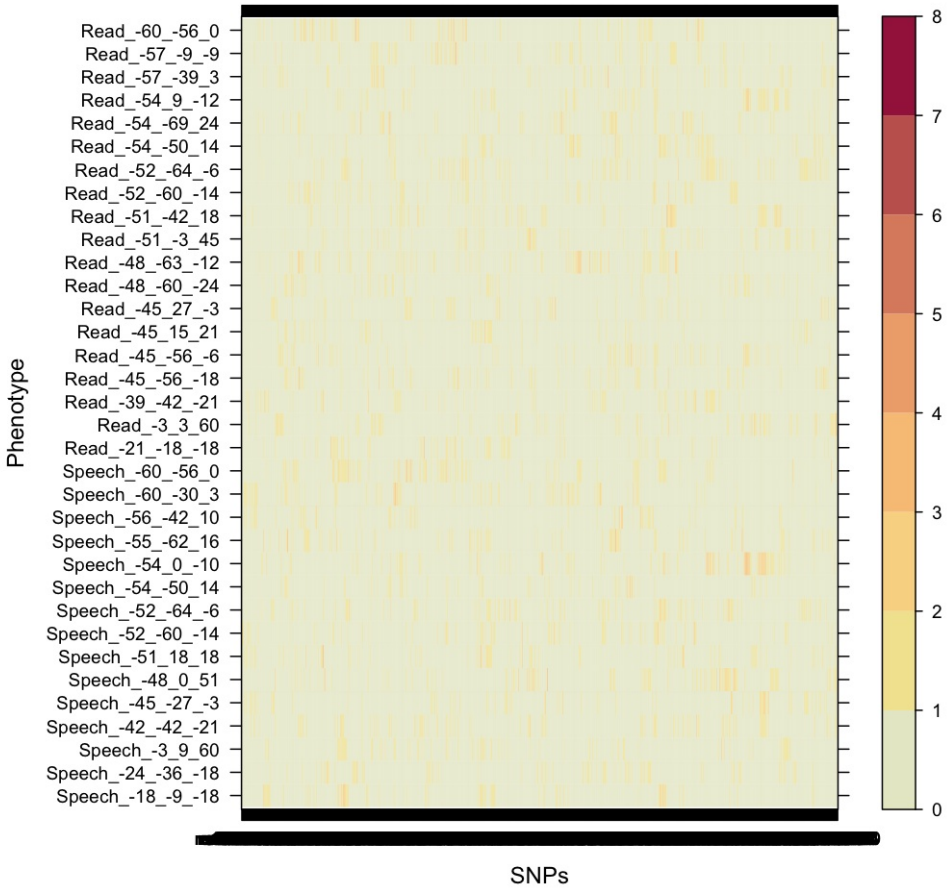


FIGURE A-9 – Carte de p-valeurs ($-\log_{10}$) calculées sur l'ensemble des SNP « candidats » pour chaque région d'intérêt.

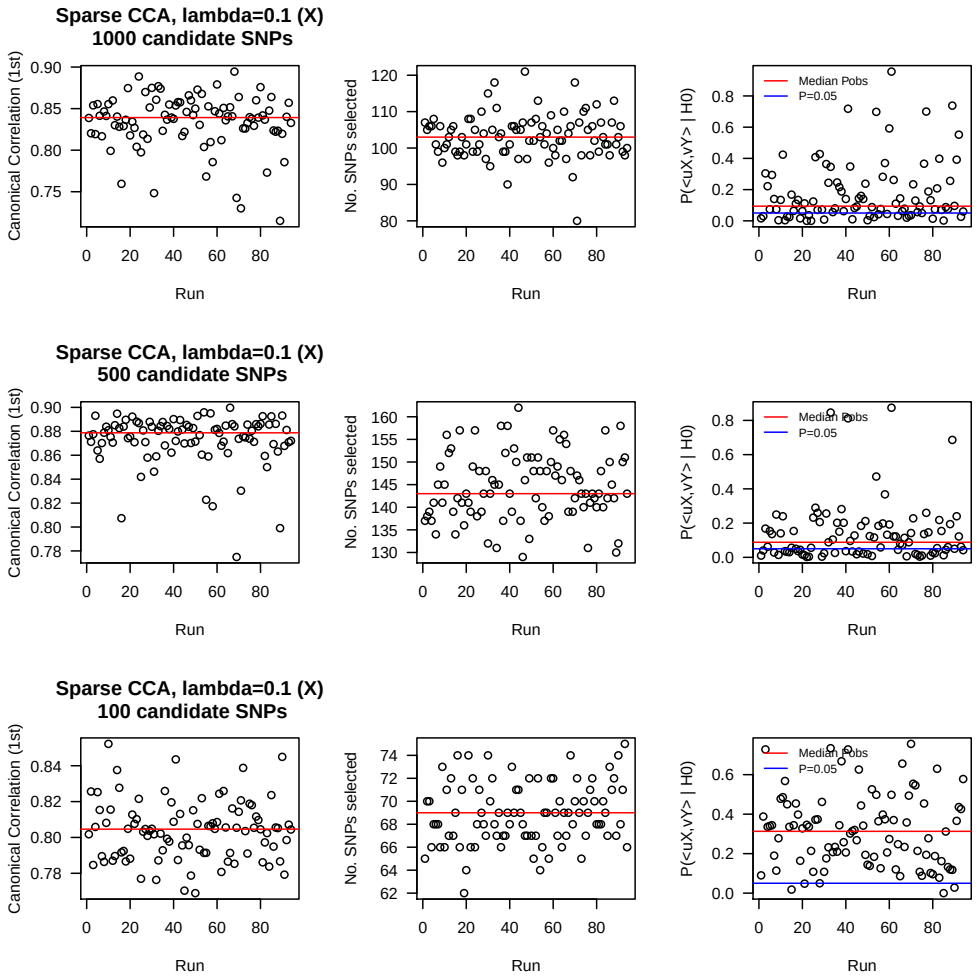


FIGURE A-10 – Correlation canonique, nombre de SNP sélectionnés et distribution du critère de test (de gauche à droite) en fonction du nombre de variables initialement fournies à l'algorithme de CCA *sparse*.

Algorithm 1 Algorithme de construction des composantes de la CCA avec pénalisation des deux blocs de variables X et Y. (Tiré de Parkhomenko et al. (2009, page 5))
$K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$ est la matrice de corrélation entre X et Y ; les paramètres de pénalisation, notés λ_u (pour X) et λ_v (pour Y), sont estimés par une procédure de validation croisée (10-fold) externe ; les vecteurs canoniques, u et v, sont initialisés aux vecteurs moyens des matrices correspondantes ; la fonction sign(x) est la fonction signe usuelle tandis que $(x)_+ = x$ si $x \geq 0$ et $(x)_+ = 0$ si $x < 0$.

Require: X centré-réduite, Y centré-réduite
  $\lambda_u \leftarrow \hat{\lambda}_u$, $\lambda_v \leftarrow \hat{\lambda}_v$
  $u^0 \leftarrow \bar{x}.$, $v^0 \leftarrow \bar{y}.$, $i = 0$
  repeat
    {Mise à jour de u}
    $u^{i+1} \leftarrow Kv^i$
    $u^{i+1} \leftarrow \frac{u^{i+1}}{\|u^{i+1}\|}$
    for $j = 1$ to p **do**
      $u_j^{i+1} \leftarrow \left(|u_j^{i+1}| - \frac{1}{2}\lambda_u\right)_+ \text{sign}(u_j^{i+1})$
    end for
    $u^{i+1} \leftarrow \frac{u^{i+1}}{\|u^{i+1}\|}$
    {Mise à jour de v}
    $v^{i+1} \leftarrow K^t u^{i+1}$
    $v^{i+1} \leftarrow \frac{v^{i+1}}{\|v^{i+1}\|}$
    for $j = 1$ to q **do**
      $v_j^{i+1} \leftarrow \left(|v_j^{i+1}| - \frac{1}{2}\lambda_v\right)_+ \text{sign}(v_j^{i+1})$
    end for
    $v^{i+1} \leftarrow \frac{v^{i+1}}{\|v^{i+1}\|}$
    $i \leftarrow i + 1$
  until convergence