UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UTILISATION DE TECHNIQUES DE FOUILLE DE DONNÉES DANS L'ANALYSE DE DONNÉES DE GÉNOTYPAGE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR
GILLES GODEFROID

FÉVRIER 2016

DÉDICACES

Je dédie ce mémoire à mon épouse Annie et à ses deux enfants Mathilde et Guillaume, avec tout mon amour.

> Je ne suis pas d'accord avec ce que vous dites, mais je me battrai jusqu'au bout pour que vous puissiez le dire.

Evelyn Beatrice Hall

> Le chemin le plus court d'un point à un autre est la ligne droite, à condition que les deux points soient bien en face l'un de l'autre.

Pierre Dac

REMERCIEMENTS

Je tiens à remercier Monsieur PETKO VALTCHEV professeur à l'Université du Québec à Montréal pour son encadrement tout au long de ce travail, ainsi que pour ses directives et ses remarques les plus constructives, qu'il trouve dans cet ouvrage un témoignage de ma profonde reconnaissance.

Mes remerciements les plus sincères vont au professeur JOHANNE TREMBLAY Professeure titulaire, Département de médecine, Université de Montréal, pour son soutien durant mes études au 2e cycle, et sans qui cette très belle expérience à l'université du Québec à Montréal n'aurait pas été possible.

Je remercie également MARIE-PIERRE SYLVESTRE Professeure adjointe, Département de médecine sociale et préventive, École de santé publique, Université de Montréal, pour ses remarques et suggestions faites sur le projet alors qu'il était encore en chantier, et qui ont contribué grandement à l'amélioration de ce travail.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ALGORITHMES

RÉSUMÉ

Le génotypage est une technique permettant l'identification d'une variation génétique dans une localisation précise du génome d'un individu, contrairement aux techniques de séquençage qui vont décoder le génome entier d'un individu.

Les Genome-Wide Associations (GWAS) sont des analyses statistiques utilisées pour analyser ces données de génotypage afin de déterminer les variantes génétiques responsables d'une pathologie ou d'un trait phénotypique.

Ces analyses mettent en valeur l'importance de telle ou telle variation génétique au sein de la population présentant cette caractéristique (population cas), par opposition à la population ne présentant pas cette caractéristique (population contrôle).

Cette méthode statistique a donné des résultats très intéressants pour des pathologies dites mono-géniques qui ne dépendent que d'un seul gène telles que la fibrose kystique.

Cependant les résultats face à des pathologies poly-géniques telles que le diabète de type 2, ou encore le syndrome métabolique, sont plus limités (Pearson et Manolio, 2008).

En effet, alors que les GWAS ont permis d'identifier plusieurs gènes associés à la maladie, une grande partie de la composante génétique demeure encore inconnue pour ces maladies.

Cela tient sans doute de la méthode même des GWAS qui évalue l'importance des variations génétiques une par une, alors que leur action est simultanée, déclenchée par la présence conjointe de plusieurs facteurs génétiques chez un même individu.

Nous avons tenté dans ce travail de caractériser des ensembles de variations génétiques, dont la présence simultanée chez un individu permettrait de déterminer un risque accru de présenter une pathologie ou un trait phénotypique. Donc nous allons rechercher des ensembles de variations génétiques (ou motifs) dont la fréquence est significativement plus grande dans la population atteinte que dans la population saine. Pour caractériser ces motifs nous allons utiliser les techniques de fouille de données (ou data mining), plus particulièrement les techniques de recherche de motifs fréquents, à l'analyse de ces données de génotypage.

En effet la fouille de données est spécialisée dans la recherche de motifs plus ou moins fréquents au sein d'une masse de données gigantesque.

Mots clés : Data mining, Motifs fréquents fermés, Génotypage, GWAS, Syndrome poly-génique.

CHAPITRE I

PRÉSENTATION DES GWAS

1.1 Rappels de biologie

1.1.1 Hérédité et caractères génétiques

Au milieu du XIX^{me} siècle, Gregor Mendel, moine botaniste dans le monastère de Brno (Moravie), a établi les prémisses de la génétique en observant la transmission de caractères héréditaires chez les pois. (Mendel, 1865)

Ce qui lui a permis d'établir les trois lois fondatrices de la génétique :

- Loi d'uniformité des hybrides de première génération, qui dit que lorsque les parents sont de souches pures aucune forme intermédiaire n'apparaît en première génération . Le concept de l'hérédité par mélange est réfuté.
- Loi de pureté des gamètes, affirmant que les facteurs héréditaires se séparent dans les gamètes. Un gamète ne contient qu'un facteur de chaque caractère.
- Ségrégation indépendante des caractères héréditaires multiples.

L'important à retenir est qu'elles établissent que des caractères, appelés à l'époque `facteurs`, sont transmis par les gamètes en double exemplaire dont chaque exemplaire provient d'un parent. À l'époque, la cytologie en était à ses balbutiements et on a découvert par la suite que les structures porteuses de ces facteurs, que l'on a ensuite appelé gènes, étaient les chromosomes situés dans le noyau de chaque cellule et constitués chacun d'une molécule d'acide désoxyribo-

nucléique (ADN).

1.1.2 Structure de la molécule d'ADN

Le matériel génétique de la plupart des êtres vivants est supporté par la molécule d'ADN dont la structure, découverte par Watson J.D. et Crick F.H.C (Watson et Crick, 1953a), est composée de deux chaines appariées en double hélice. Ces chaines sont composées d'une structure externe faite d'une succession de sucres désoxyriboses reliés par des groupements phosphate.

Chacune de ces deux chaines comporte une succession de bases appelées nucléotides, qui sont au nombre de 4 (Watson et Crick, 1953b) :

— Adénine (A),
— Cytosine (C),
— Guanine (G),
— Thymine (T).

Ces bases sont disposées perpendiculairement aux sucres et sont reliées deux à deux avec une autre base de la chaîne opposée, l'appariement des bases se fait toujours dans cet ordre :

— Adénine avec Thymine,
— Cytosine avec Guanine.

Cette liaison est réalisée par des liaisons hydrogènes.

Un schéma de cette structure est présenté dans la figure 1.1.

Cet ensemble permet de relier les deux chaines de désoxyribose et de former une échelle, celle-ci s'enroulant sur elle-même selon son axe longitudinal formant ce que l'on appelle communément la double hélice de l'ADN.

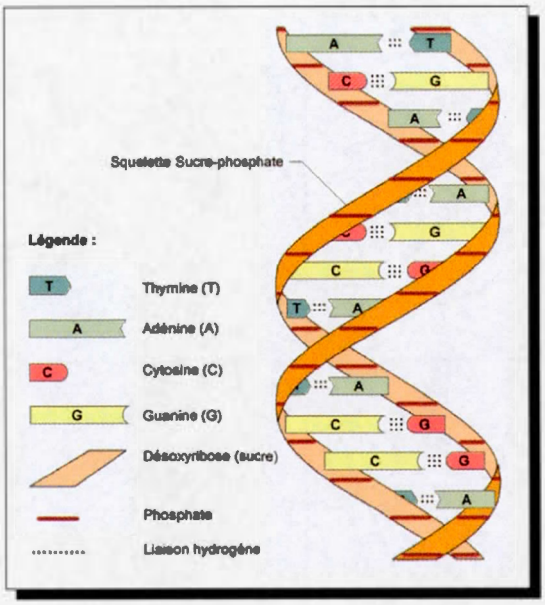

Figure 1.1 Schéma de la structure d'ADN (Jobling *et al.*, 2013)

1.1.3 Le code génétique (Martin *et al.*, 1962), (Nirenberg *et al.*, 1963)

Élucidé dans les années 60, le code génétique est la clé de voute de la synthèse de protéines et de leur régulation.

La succession de 3 nucléotides détermine ce que l'on appelle un *codon*. Le nombre de codons possibles est donc de 4^3 soit 64.

Les protéines sont composées d'une chaîne d'acides aminés qui sont au nombre de 20. Chaque codon va, lors de la synthèse des protéines, coder pour un acide aminé. Plusieurs codons codent pour le même acide aminé, par exemple les codons AGC, AGT, TCA, TCG, TCC et TCT codent tous pour l'acide aminé appelé le sérine. D'autre part 3 codons (TAA, TAG, TGA) ne codent pour aucun acide aminé ce sont les codons stop.

Ce code est représenté dans la figure 1.2.

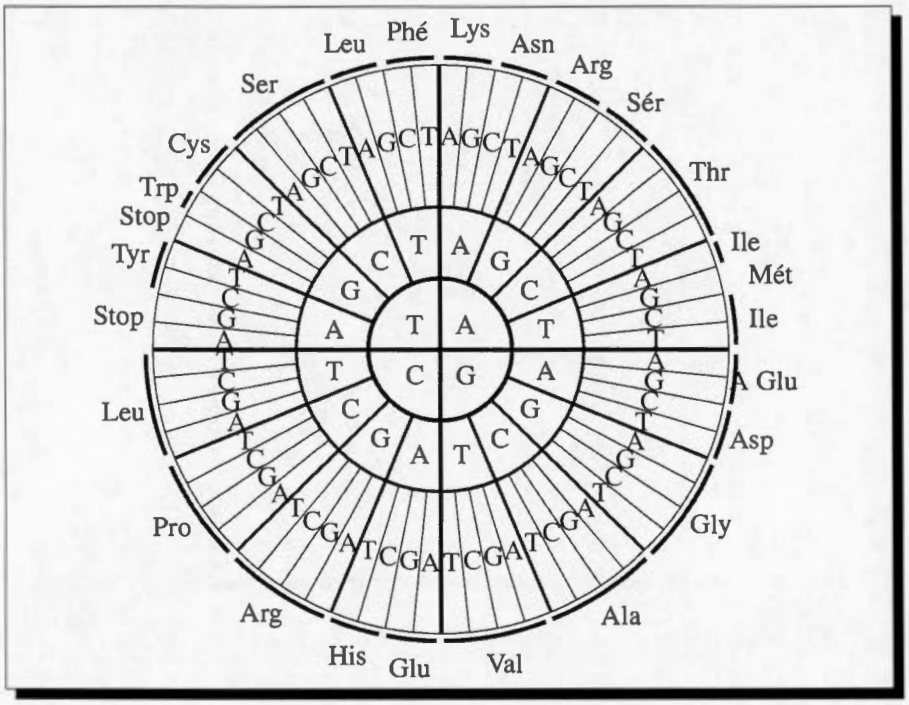

Figure 1.2 Tableau des codons et des acides aminés

1.1.4 Structure des protéines (Branden et Tooze, 1996), (Nelson *et al.*, 2008)

On distingue les structures :

- primaires, correspondant à la séquence linéaire des acides aminés,
- secondaires, dues aux interactions physico-chimiques des acides aminés entre eux, elle est le résultat des contraintes d'interaction,
- tertiaires, qui consistent en le repliement en trois dimensions des structures secondaires,
- quaternaires, consistant en la mise en relation avec d'autres molécules protéiques (ou non) permettant d'obtenir une configuration spatiale thermodynamiquement fonctionnelle.

De ces diverses structures on doit retenir que chacune d'entre elles résultent de la succession initiale des acides aminés et donc de la séquence des nucléotides sur le brin d'ADN, d'où

l'importance potentielle d'une variation de cette séquence.

1.1.5 Variations génétiques

Cette succession est sujette à des variations qui peuvent être de différentes natures :

- Changement d'une seule base dans la succession des nucléotides (Varela et Amos, 2010) ce type de variant est appelé 'Single Nucleotid Polymorphism' (SNP).
- Variabilité du nombre de copies (CNV) (Stankiewicz et Lupski, 2010) qui se répètent un certain nombre de fois à la suite et sont importantes pour l'apparition de certaines pathologies. Le nombre de copies à l'état sauvage (c'est à dire en l'absence de variation du génome) étant de 2 on parle de délétions lorsque celui-ci est inférieur à 2, ou de duplications s'il est supérieur à 2.
- Ou d'autres recombinaisons génétiques (inversion, translocation, etc..).

Nous ne nous intéresserons qu'aux variations de type SNP, très fréquentes au sein du génome humain.

Les SNPs (Brookes, 1999)

Comme mentionné précédemment, un SNP est le changement d'un seul nucléotide au sein de la succession de bases dans la chaine d'ADN. Un SNP peut être muet et ne pas s'exprimer si, étant présent dans une séquence codante, il ne change pas la structure de la protéine, mais il peut également entraîner une modification de la structure et de la fonction de la protéine. Enfin il peut entraîner une variation dans les mécanismes de régulation de l'expression d'un gène, ou d'autres entités telles que les micro-ARN. La figure 1.3 schématise un SNP dans une chaîne d'ADN.

Ces SNPs sont caractérisés par leur *fréquence minimale allélique* (MAF) qui est la fréquence de l'allèle le moins présent dans les génotypes étudiés.

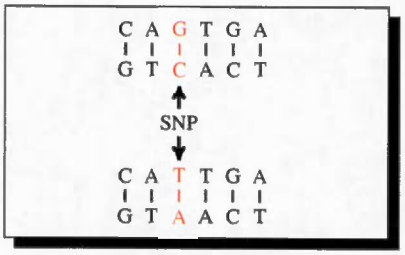

Figure 1.3 Schéma d'un SNP dans une chaine d'ADN

1.1.6 Modèles mendéliens et autres modèles d'expressions génétiques

Lorsque Mendel à dicté les trois lois fondamentales de la génétique, l'idée qui sous-tendait son raisonnement était que pour un caractère donné, un seul `facteur` était mis en oeuvre, d'où la notion de dominance et de récessivité d'un allèle par rapport à un autre.

Modèles dominants et récessifs

Dans le modèle récessif on considère que l'allèle ne s'exprime que s'il est présent dans les deux chromosomes (*homozygote*), alors que dans le modèle dominant il s'exprime également lorsqu'il est présent dans un seul des chromosomes (*hétérozygote*) .

Modèle additif

Dans ce modèle l'expression de l'allèle est considérée comme étant deux fois plus forte lorsque l'individu est homozygote.

1.1.7 Haplotypes et linkage desequilibrium (Pritchard et Przeworski, 2001)

Haplotype

On définit les haplotypes comme étant des ensembles d'allèles sur différents loci situés sur un même chromosome habituellement transmis ensemble.

Linkage desequilibrium et haploblocs

Soit l'allèle A au locus 1 et l'allèle B au locus 2 avec une fréquence respective de π_A et π_B.

Si ces allèles sont indépendants la fréquence de l'haplotype AB devrait être $\pi_A \times \pi_B$, si cette fréquence est supérieure à cette valeur, cela démontre que ces allèles tendent à être observés ensemble, on dit alors qu'ils sont en *linkage desequilibrium* (LD).

En d'autre termes, la tendance de certains allèles, situés sur des loci différents, à être liés à cause d'un nombre restreint de recombinaisons, est appelé linkage desequilibrium. (Jobling *et al.*, 2013)

De nombreuses variables statistiques ont été définies afin de mesurer la force du LD, la plus communément utilisée est appelée r^2, définie ainsi :

- Soit deux loci bialléliques sur le même chromosome, que nous noterons respectivement A a et B b,
- Soit leurs fréquences notées $\pi_A\ \pi_a\ \pi_B\ \pi_b$,
- Soit les fréquences des 4 haplotypes notées $\pi_{AB}\ \pi_{Ab}\ \pi_{aB}\ \pi_{ab}$.

Alors :

$$r^2 \equiv \frac{(\pi_{AB} - \pi_A \pi_B)^2}{\pi_A \pi_a \pi_B \pi_b}$$

La valeur de r^2 communément utilisée pour affirmer que deux SNPs ont en LD est de 0.8. Le regroupement de ces SNPs permet de définir des 'blocs' appelés *haploblocs*. La sélection d'un nombre restreint de SNPs (tagSNPs) au sein d'un haplobloc permet de caractériser ce bloc en entier.

1.2 Données de génotypage

Nous passerons les aspects techniques de l'obtention de ces données (extraction de l'ADN, réactions PCR, puces à ADN, lectures des puces, etc..), pour nous intéresser plus spécialement aux données directement issues du génotypage. Le format général de ces données est un tableau avec une ligne par individu avec son identification, et éventuellement sa filiation, et des colonnes indiquant le nombre d'exemplaires du SNP considéré 0, 1 ou 2, selon que cette variation géné-

tique est, respectivement, absente du génome de l'individu, présente sur un de ses chromosomes ou présente sur les deux chromosomes. C'est à partir de ces données que nous démarrons pour effectuer les analyses GWAS. Dans les données à notre disposition nous détectons la présence d'environ un million de SNPs (Teumer *et al.*, 2013).

1.3 Principe général des GWAS (Visscher *et al.*, 2012)

Les GWAS consistent en l'analyse de nombreuses variantes génétiques pour déterminer si la présence de certaines d'entre elles augmentent ou baissent le risque d'une maladie au sein d'une population. Classiquement les GWAS se concentrent sur la correspondance entre certains SNPs et des pathologies.

Les GWAS ont permis depuis 2005 de mettre en évidence des milliers d'associations entre des variantes génétiques et des phénotypes particuliers (Pearson et Manolio, 2008) (Manolio, 2010). Dans le cas de traits complexes, tels que des maladies métaboliques (par exemple diabète de type 2 ou l'indice de masse corporelle) ou des maladies auto-immunes (par exemple diabète de type 1, colite ulcéreuse ou maladie de Crohn), l'utilisation des GWAS a mis en lumière de nombreux SNPs depuis 2007 ainsi que le montre le tableau 1.1.

Tableau 1.1 Évolution du nombre de loci mis en évidence depuis l'utilisation des GWAS entre 2007 et 2012 d'après (Visscher *et al.*, 2012)

	Nombre de loci	
Pathologie	Avant 2007	Après 2007
Diabete de type 1	4	40
Colite ulcéreuse	3	44
Maladie de Crohn	4	67
Diabete de type 2	3	50
Indice de masse corporelle	1	30

Cependant si l'on compare la transmission de ces pathologies lors d'études familiales avec les variations au sein de la population expliquées par les GWAS, ainsi que reportée dans le tableau 1.2, on observe de grandes différences, ouvrant la voie aux critiques vis à vis des GWAS confrontés aux pathologies poly-géniques.

Tableau 1.2 Comparaison de la transmission de certains traits complexes avec les variations expliquées par les GWAS d'après (Visscher *et al.*, 2012)

Pathologie	Études familiales	SNPs GWAS
Diabète de type 1	0.9	0.6
Diabète de type 2	0.3-0.6	0.05-0.06
Indice de masse corporelle	0.4-0.6	0.01-0.02
Maladie de Crohn	0.6-0.8	0.1
Colite ulcéreuse	0.5	0.05

1.4 Conclusion

Les GWAS ont permis d'identifier de nombreuses variantes génétiques comme étant responsables de pathologies importantes.

Ces succès sont particulièrement marqués dans le cas de maladies dites mono-géniques, telle la fibrose kystique (Cantor *et al.*, 2010). Par contre elles se heurtent souvent à des écueils face à des pathologies complexes, multi-factorielles telles, que le diabète de type 2 ou l'hypertension (Manolio, 2010).

La sélection d'un ensemble de SNPs significatifs passe par l'établissement de score de risque génétique pour chaque SNPs dont les critères varient d'une étude à l'autre (Vaxillaire *et al.*, 2014) (Krarup *et al.*, 2015) (Edridge *et al.*, 2015). Ce manque d'homogénéité dans la démarche provoque un manque de méthodes standards permettant de tirer des résultats comparables d'une étude à l'autre.

Une des hypothèses possibles expliquant les résultats insatisfaisants face aux pathologies mult-factorielles (Hodge, 1994) est que ces phénotypes sont dépendants de la présence simultanée de plusieurs variantes génétiques, alors que l'analyse par GWAS considère l'importance de chaque variants individuellement, sans appréhender l'intervention de groupes de SNPs dans leur globalité.

Les techniques de *fouille de données*, bien qu'initialement développées dans d'autres sphères d'activité, semblent tout à fait adaptées pour détecter la fréquence et l'importance de regroupements d'éléments distincts au sein d'une population.

Aussi notre démarche va être de chercher, dans le cadre de phénotypes qualitatifs pour lesquels la population peut-être séparée en cas et en contrôles, à extraire des ensembles de SNPs dont la fréquence est significativement différente entre les populations cas et les populations contrôles.

CHAPITRE II

MÉTHODES D'EXTRACTION DE MOTIFS FRÉQUENTS

2.1 Notions de fouille de données

La croissance exponentielle de la masse des données stockées dans les entrepôts de données, dans des domaines aussi différents que le marketing, la génétique, l'astronomie ou le monde de la finance, ne permet plus d'opérer de manière traditionnelle en transformant directement les données en connaissances (Fayyad *et al.*, 1996). De nos jours des bases de données contenant un nombre d'enregistrement de l'ordre de 10^9 n'a plus rien d'exceptionnel (Han et Kamber, 2012).

Les fondements de la fouille de données furent établies par John Tukey en 1977 (Tukey, 1977) sous le terme de `Exploratory data analysis`. La fouille de données est une étape essentielle du processus d'extraction des connaissances dans les bases de données (Knowledge Discovery from Database) (Piateski et Frawley, 1991). Les différentes étapes en sont montrées dans la figure 2.1.

L'étape de la fouille de données correspond à l'ensemble des techniques et méthodes mises en oeuvre afin d'extraire, à partir de données préparées, des informations exploitables non accessibles par les méthodes classiques (Abbas, 2012). Ces informations sont des *règles d'associations*, des motifs fréquents ou rares, des regroupements, des anomalies et des modèles.

On peut citer une définition donnée par Frawley et al. (Frawley *et al.*, 1992) :

"Extraction of interesting (non-trivial, implicit, previously unknown and poten-

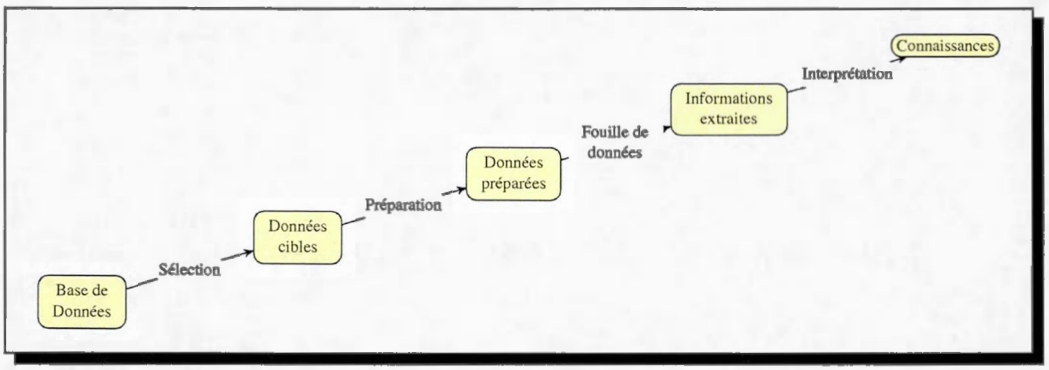

Figure 2.1 Étapes du processus d'extraction de connaissances dans les bases de données

tially useful) patterns or knowledge from huge amount of data."

Donc nous nous intéressons à l'extraction de motifs ou de connaissances intéressants :

— non triviaux,
— implicites,
— inconnus auparavant,
— potentiellement utiles,

à partir d'une masse énorme de données.

D'après (Tan *et al.*, 2006), il existerait trois grandes catégories de problèmes de fouille :

— la classification,
— le regroupement automatique,
— la découverte d'associations.

Notre démarche de recherche nous orientera vers la découverte de règles d'associations, tout au moins à sa première phase consistant en la découverte de motifs fréquents.

2.2 Définitions

Soit un ensemble $\mathcal{I} = \{i_1, i_2, \cdots, i_m\}$ constitué de m éléments (articles, items, etc.). Soit $\mathcal{D} = \{t_1, t_2, \cdots, t_n\}$ un ensemble de n transactions, où chaque transaction t_i représente un

ensemble d'items $\mathcal{I}_i$ tel que $\mathcal{I}_i \subset \mathcal{I}$ (Abbas, 2012).

Motif

Un motif est un sous-ensemble de $\mathcal{I}$, donc un ensemble d'éléments présents dans $\mathcal{I}$, la cardinalité des motifs varie de 0 ($\emptyset$) à m ($\mathcal{I}$) en entier. Un motif composé de k éléments est appelé k-motif.

Support absolu d'un motif

Le support absolu d'un motif (σ) est le nombre de transactions contenant ce motif dans la base de données $\mathcal{D}$.

Support relatif d'un motif

Le support relatif d'un motif (σ) est le pourcentage de transactions contenant ce motif dans la base de données $\mathcal{D}$ comme indiqué dans la figure 2.2

$$\sigma_{\mathcal{I}_i} = \frac{n_{\mathcal{I}_i}}{n_T} \times 100$$

$\sigma_{\mathcal{I}_i}$= Support relatif du motif $\mathcal{I}_i$
$n_{\mathcal{I}_i}$= Nombre de transactions contenant $\mathcal{I}_i$ dans $\mathcal{D}$
n_T=Nombre total de transactions dans $\mathcal{D}$

Figure 2.2 Calcul du support relatif d'un motif

Motifs fréquents (Agrawal *et al.*, 1993)

Un motif est dit fréquent si son support est supérieur ou égal à une valeur σ_{min} appelée support minimal. Cette valeur est choisie arbitrairement par l'utilisateur, en fonction de sa recherche.

Motifs fermés

Un motif est dit fermé lorsqu'aucun des motifs qui le contiennent ne possède le même support (Pasquier *et al.*, 1999). Cela permet une représentation minimale des motifs sans perte de leur support.

Motifs fréquents fermés

Ce sont des motifs fermés dont le support est supérieur ou égal au support minimal.

2.3 Étapes de la recherche de motifs fréquents

L'extraction des motifs fréquents se répartit en 3 étapes (Pasquier, 2000) :

— la sélection et la préparation des données,

— l'extraction des motifs fréquents,

— la visualisation et l'interprétation des résultats.

Sélection et préparation des données

Le but est de sélectionner à partir de la base de données un sous-ensemble de données utiles dans l'extraction des motifs fréquents et de les transformer en un *contexte d'extraction*. La transformation est utile afin d'appliquer les algorithmes d'extraction des motifs fréquents.(Chen *et al.*, 1996)

Contexte d'extraction

Un contexte d'extraction est un triplet $\mathcal{T} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, dans lequel $\mathcal{O}$ et $\mathcal{I}$ sont respectivement des ensembles finis de transactions et d'items, et $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ est une relation binaire entre les transactions et les items.

Un couple (o,i) signifie que la transaction $o \in \mathcal{O}$ est en relation avec l'item $i \in \mathcal{I}$ (Pasquier, 2000)

Soit une base de données composée de 6 transactions sur un ensemble de 5 articles ou items {A, B, C, D, E}. Un contexte d'extraction peut se représenter sous forme d'un tableau à deux dimensions où les lignes correspondent aux objets et les colonnes aux items (voir tableau 2.1).

Recherche des motifs fréquents

Cette recherche est un problème non trivial car la cardinalité de l'ensemble des motifs potentiels augmente de façon exponentielle.

Tableau 2.1 Contexte d'extraction

transactions \ articles	A	B	C	D	E
1	X	X		X	X
2		X	X		X
3	X	X		X	X
4	X	X	X		X
5	X	X	X	X	X
6		X	X	X	

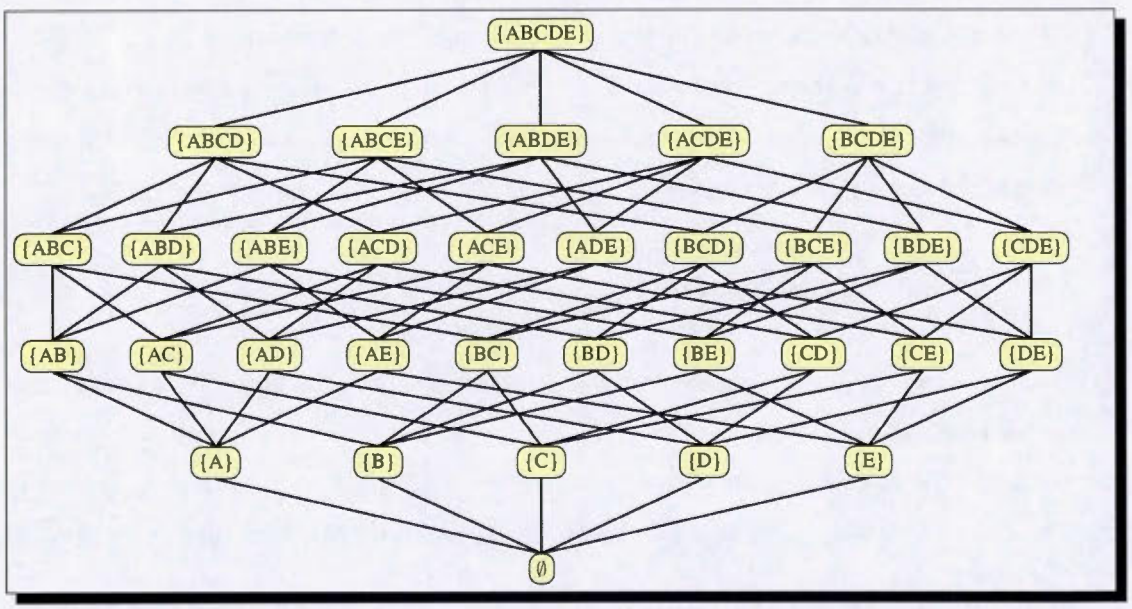

Figure 2.3 Treillis Booléen des motifs associés au contexte $\mathcal{T}$

Treillis Un treillis est un ensemble partiellement ordonné dans lequel chaque couple d'éléments admet une borne supérieure et une borne inférieure.

Treillis booléen Un treillis booléen est un treillis entièrement généré par ses noeuds de premier niveau (au dessus de l'infimum), appelés atomes. Ainsi, chaque noeud correspond à un sous-ensemble d'atomes. Ce treillis est isomorphe de tous les sous-ensembles de l'ensemble d'atomes. Par conséquent, son opérateur infimum correspond à l'intersection de ses arguments (ensembles d'atomes) et le supremum en est l'union. Un opérateur additionnel existe, le com-

plément par rapport à l'ensemble de tous les atomes : il peut être calculé pour tout élément du treillis.

Le treillis booléen de l'ensemble des motifs potentiels de l'ensemble $\mathcal{I}$ du contexte $\mathcal{T}$ est représenté dans la figure 2.3. Ainsi on se rend compte que pour 5 items on a 32 possibilités, depuis l'ensemble vide jusqu'à $\mathcal{I}$ lui même. Le nombre de motifs potentiellement fréquents, pour un ensemble de m items est de l'ordre de 2^m. La méthode triviale serait de tester chacune de ces possibilités une par une afin de déterminer leur support et d'éliminer les objets dont le support est inférieur au support minimal. Du fait des remarques précédentes concernant la taille de l'ensemble des possibilités, on a cherché à réduire le champ d'investigation. Ainsi se sont développé deux familles d'algorithmes de recherche de motifs fréquents. Les algorithmes à niveaux ou algorithmes séquentiels et les algorithmes verticaux.

2.4 Les algorithmes à niveaux

Cette famille d'algorithmes procède de manière itérative en parcourant le treillis des motifs potentiels par niveaux. Cette approche est efficace sur des données faiblement corrélées et éparses. Le chef de file de ce type d'algorithme, et le premier des algorithmes de recherche de motifs fréquents est l'algorithme APRIORI développé par Agrawal en 1993 (Agrawal *et al.*, 1993).

Étant donné la structure des bases de données visées dans cette recherche, nous nous intéresserons plus spécifiquement aux algorithmes verticaux.

2.5 Les algorithmes verticaux

Concepts de base

La terminologie utilisée provient des travaux qui ont abouti à la description de l'algorithme Eclat, chef de file de cette famille d'algorithmes (Zaki *et al.*, 1997) (Zaki, 2000) (Zaki et Hsiao, 2002) (Zaki et Gouda, 2003) (Abbas, 2012). Soit une base de données transactionnelle $\mathcal{D}$. Soit un ensemble d'items classés en ordre lexicographique $\mathcal{I}$. Chaque transaction possède un identifiant : t_{id}. Soit $\mathcal{T}$ l'ensemble de tous les t_{id}. Un ensemble de $t_{id} \subseteq \mathcal{T}$ est appelé *tidset*.

Un motif {A, B, C} est représenté par ABC et un tidset {1, 2, 3} est représenté par 123. À chaque motif *X* correspond un tidset *t(X)* représentant l'ensemble des transactions contenant ce motif comme sous-ensemble.

À un tidset *Y* correspond le motif *i(Y)* ensemble de tous les items contenus communs à toutes les transactions contenues dans le tidset.

$t(X) = \bigcap x \in Xt(x)$

$i(Y) = \bigcap y \in Xi(y)$

On utilise la notation $X\times$ *t(X)* pour désigner une paire motif-tidset, et nous l'appelons IT-paire. Le support absolu d'un motif *X* noté σ_X est égal à $|t(X)|$.

La partition par classes d'équivalence, arbre des préfixes

Une classe d'équivalence représente un ensemble de motifs possédant le même préfixe. Elle est associée à ce préfixe ainsi qu'aux items qui sont possiblement ajoutés à un préfixe afin d'obtenir de nouveaux motifs présents.

Soit une fonction $p(X, k) = X[1 : k]$ définissant le préfixe de longueur k de X, et une relation d'équivalence θ_k entre deux motifs *X* et *Y* définie par le fait que les k premiers éléments formant les préfixes de deux motifs *X* et *Y* sont les mêmes.

$$\forall X, Y \subseteq I, X\theta_k Y \Leftrightarrow p(X, k) = p(Y, k)$$

Ainsi dans le contexte d'extraction pris comme exemple précédemment pour les motifs {A, B, C} et {A, B, D}, on a {A, B, C} θ_2 {A, B, D}. Ainsi on peut définir un arbre de recherche ainsi que montré dans la figure 2.4.

La puissance de cette approche est que toute branche issue d'un noeud non fréquent n'a pas à être examinée, elle permet de partitionner l'espace initial de recherche en sous problèmes indépendants.

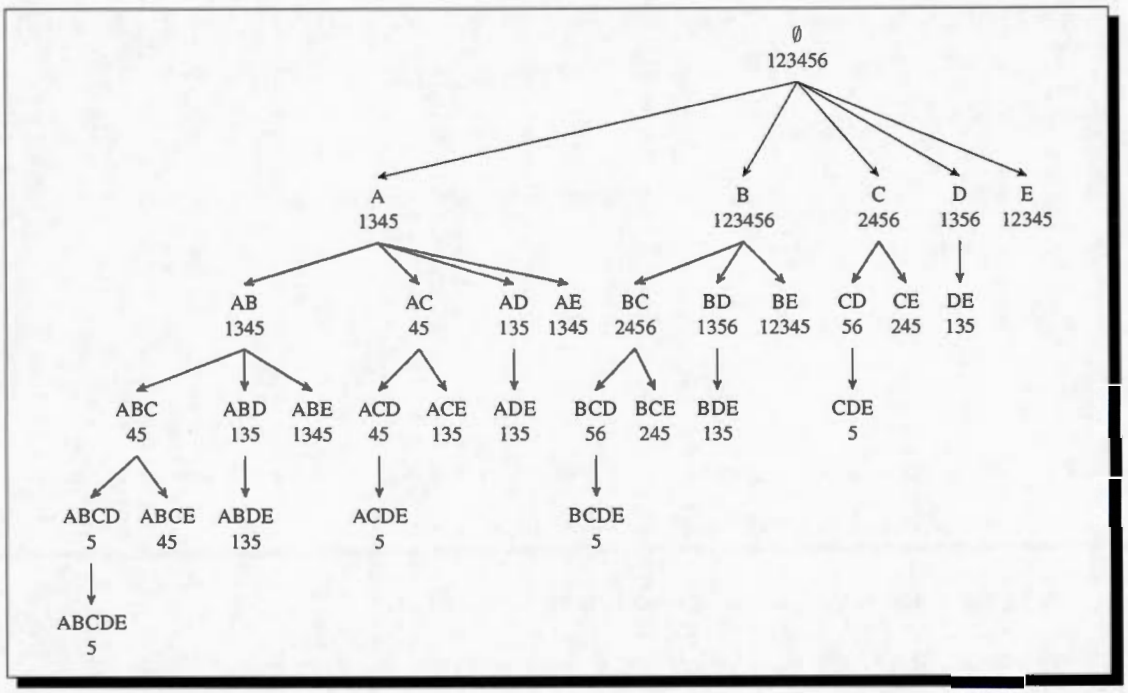

Figure 2.4 Arbre de recherche issu du contexte d'extraction du tableau 2.1

2.6 Notion sur les concepts fermés et les treillis de Galois (Valtchev *et al.*, 2003)

Afin de compléter la notion de motifs fréquents fermés, nous aborderons de manière succincte des notions d'analyse formelle des concepts, d'opérateurs de fermeture, et de connexions de Galois (Barbut et Monjardet, 1970) (Ganter et Wille, 1997) (Davey et Priestley, 2002) (Valtchev *et al.*, 2002).

Soit un contexte d'extraction $\mathcal{T} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ où $\mathcal{O}$ est un ensemble fini de transactions, $\mathcal{I}$ un ensemble fini d'items. $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ est une relation binaire entre les transactions et les items.

2.6.1 Opérateurs de fermeture (Caspard et Monjardet, 2003).

Soit un S un ensemble fini et soient $A, B \subseteq S$, la fonction ϕ est un opérateur de fermeture de S si et seulement si elle possède les trois propriétés décrites dans la figure 2.5.

1. Isotonie : $A \subseteq B \implies \phi(A) \subseteq \phi(B)$
2. Extensivité : $A \subseteq \phi(A)$
3. Idempotence : $\phi(\phi(A)) = \phi(A)$

Figure 2.5 Propriétés d'un opérateur de fermeture

2.6.2 Connexions de Galois (Wille, 1992).

Soit $\mathcal{T} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction . Pour tout $O \subseteq \mathcal{O}$ et $I \subseteq \mathcal{I}$ on définit :

$f(O) : P(\mathcal{O}) \rightarrow P(\mathcal{I})$

$f(O) = \{i \in \mathcal{I} | \forall o \in O, (o, i) \in R\}$

et

$g(I) : P(\mathcal{I}) \rightarrow P(\mathcal{O})$

$g(I) = \{o \in \mathcal{O} | \forall i \in I, (o, i) \in R\}$

Le couple de fonctions (f, g) est une connexion de Galois entre l'ensemble des parties de $\mathcal{O}$ et l'ensemble des parties de $\mathcal{I}$.

Ainsi dans le contexte d'extraction cité précédemment, f({A, B}) = {1, 3, 4, 5}, et g({1, 6}) = {B, D}.

2.6.3 Opérateurs de fermeture de Galois

Par simple application des définitions précédentes, on peut déduire que $f \circ g$ est un opérateur de fermeture de $\mathcal{O}$, et que $g \circ f$ est un opérateur de fermeture de $\mathcal{I}$ ($f \circ g(x) = f(g(x))$, f et g sont les fermetures de Galois. Étant donné la connexion de Galois (f, g), on déduit les propriétés suivantes décrites dans la figure 2.6.3.

$\forall I, I_1, I_2 \subseteq \mathcal{I}$ et $O, O_1, O_2 \subseteq \mathcal{O}$

$I_1 \subseteq I_2 \implies g(I_1) \supseteq g(I_2)$	$O_1 \subseteq O_2 \implies g(O_1) \supseteq g(O_2)$
$I \subseteq f \circ g(I)$	$O \subseteq g \circ f(O)$
$f \circ g(f \circ g(I)) = f \circ g(I)$	$g \circ f(g \circ f(O)) = g \circ f(O)$
$I_1 \supseteq I_2 \implies f \circ g(I_1) \supseteq f \circ g(I_2)$	$O_1 \supseteq O_2 \implies g \circ f(O_1) \supseteq g \circ f(O_2)$
$f \circ g(g(I)) = g(I)$	$g \circ f(f(O)) = g(O)$

$$O \subseteq g(I) \iff I \subseteq f(O)$$

Figure 2.6 Propriétés des opérateurs de fermeture d'une connexion de Galois

2.6.4 Motifs fermés

Traitée par l'analyse formelle de concepts et les connexions de Galois la notion de motif fermé devient :

Soit $I \subseteq \mathcal{I}$, I est un motif fermé $\iff f \circ g(I) = C$.

Par exemple, dans le contexte d'extraction cité précédemment, f({A, B, D, E}) = {1, 3, 5} et g({1, 3, 5}) = {A, B, D, E}, donc c({A, B, D, E}) = {A, B, D, E} donc c'est un motif fermé. Par contre f({A, D, E}) = {1, 3, 5} et g({1, 3, 5}) = {A, B, D, E}, donc c({A, B, D}) = {A, B, D, E} par conséquent c({A, B, D}) $\neq$ {A, B, D, E}, donc {A, B, D} n'est pas un motif fermé.

2.6.5 Treillis des motifs fermés

Soit $\mathcal{C}$ l'ensemble des motifs fermés dérivés d'un contexte d'extraction et la relation d'ordre partiel $\subseteq$. La paire $\mathcal{L}_\mathcal{C} = (\mathcal{C}, \subseteq)$ forme un treillis complet, car elle respecte les deux propriétés définissant un treillis complet :

1. Un ordre partiel sur le treillis : $C_1, C_2 \in \mathcal{L}_\mathcal{C}, C_1 \leq C_2 \iff C_1 \subseteq C_2$
2. $\forall \mathcal{A} = \{C_i\}_{i=1,n} \in C$ il existe un plus petit majorant $Join(C_i)$ et un plus grand minorant $Meet(C_i)$

$$Join(C_i) = f \circ g(\bigcup_{C_i \in \mathcal{A}} C_i)$$

$$Meet(C_i) = \bigcap_{C_i \in \mathcal{A}} C_i$$

Les connexions ou treillis de Galois sont des modèles mathématiques qui, appliqués à la fouille de données et plus particulièrement à la recherche de motifs fréquents fermés, permettent de conceptualiser des structures de données adaptées à la problématique (Valtchev *et al.*, 2004), et ils offrent de nombreux avantages pour la résolution des étapes, tels que :

— la construction, économique en ressource, d'une représentation compacte de toute ou d'une partie de la base de données,
— la construction facilitée d'arbres à préfixes,
— l'application possible des propriétés des Treillis de Galois (propriétés de fermeture par exemple) sur un contexte d'extraction.

2.7 Algorithme CHARM

Cet algorithme, développé par Zaki en 2002 (Zaki et Hsiao, 2002), fait partie des algorithmes verticaux.

2.7.1 Arbre de recherche, classes d'équivalence

Soit la classe [P] des préfixes tel que P=$\{l_1, l_2, \cdots, l_n\}$, où P est le nœud parent et chaque l_i est un seul item, représentant le nœud P$l_i \times t(\text{P}l_i)$. En reprenant l'exemple de la figure 2.4 la classe [A]={B, C, D, E}, chaque membre de la classe représente un enfant du nœud parent. La classe [A] représente l'ensemble des motifs ayant A comme préfixe. Il apparait qu'aucune sous-branche d'un préfixe infréquent n'a à être examiné. Ainsi par cette approche, on a âtitionné le problème initial de recherche de motifs fréquents en plusieurs sous-problèmes indépendants.

Propriétés de fermeture

La fermeture d'un motif X appelé $c(X)$ est le plus petit ensemble fermé qui contient X. Pour trouver la fermeture d'un motif X, on calcule son image dans l'espace des transactions $t(X)$, puis on calcule l'image de cette image dans l'espace des motifs $i(t(X))$. Enfin il s'ensuit que

X est un motif fermé si et seulement si $X = c(X)$.

De cette définition on déduit 4 propriétés résumées dans la figure 2.7 :

Soit $X_k \times t(X_k)$ et $X_j \times t(X_j)$ deux éléments d'une classe [P], avec $X_k \leq_f X_j$ où f est l'ordre de tri, alors :

1. Si $t(X_k) = t(X_j)$, alors $c(X_k) = c(X_j) = c(X_k \cup X_j)$
2. Si $t(X_k) \subset t(X_j)$, alors $c(X_k) \neq c(X_j)$ et $c(X_k) = c(X_k \cup X_j)$
3. Si $t(X_k) \supset t(X_j)$, alors $c(X_k) \neq c(X_j)$ et $c(X_k) = c(X_k \cap X_j)$
4. Si $t(X_k) \neq t(X_j)$, alors $c(X_k \neq c(X_j) \neq c(X_k \cup X_j)$

Figure 2.7 Propriétés de la fermeture

DÉMONSTRATION

Soit $X_k \times t(X_k)$ et $X_j \times t(X_j)$ deux éléments d'une classe [P], avec $X_k \leq_f X_j$ où f est l'ordre de tri, cela signifie que l'on est sur le même niveau de l'arbre avec un préfixe commun X, par exemple AB et AC.

1. Si $t(X_k) = t(X_j)$, alors $c(X_k) = c(X_j)$. D'autre part $t(X_k) = t(X_j)$ implique que $t(X_k \cup X_j) = t(X_k) \cap t(X_j) = t(X_k)$. Par conséquent si $t(X_k) = t(X_k \cup X_j)$, alors $i(t(X_k)) = i(t(X_k \cup X_j))$ donc $c(X_k) = c(X_k \cup X_j)$. Cette propriété implique que l'on peut remplacer chaque occurence X_k par $X_k \cup X_j$ et on peut retirer l'élément X_j des investigations ultérieures. Dans l'arbre représenté dans la figure 2.4, ce cas est représenté par AB et AE.
2. Si $t(X_k) \subset t(X_j)$, alors $t(X_k \cup X_j) = t(X_k) \cap t(X_j) = t(X_k)$, mais $t(X_k) \neq t(X_j)$, donnant ainsi $c(X_k \cup X_j) = c(X_k) \neq c(X_j)$. Donc nous pouvons remplacer chaque occurence de X_k par $X_k \cup Y_j$, puisqu'ils ont la même fermeture. Mais étant donné que $c(X_k) \neq c(X_j)$ nous ne pouvons pas enlever X_j de la classe [P]. Dans l'arbre de la figure 2.4, ce cas est représenté par AB et AC.
3. Si $t(X_k) \supset t(X_j)$, alors $t(X_k \cup X_j) = t(X_k) \cap t(X_j) = t(X_j)$, mais $t(X_k) \neq t(X_j)$, donnant $c(X_k \cup X_j) = c(X_j) \neq c(X_k)$. Donc nous pouvons remplacer chaque occu-

rence de X_j par $X_k \cup X_j$, étant donné qu'ils ont la même fermeture. Mais étant donné que $c(X_k) \neq c(X_j)$ nous ne pouvons pas enlever X_k de la classe [P]. dans l'arbre de la figure 2.4ce cas est représenté par AC et AE dans l'arbre de la figure 2.4.

4. Si $t(X_k) \neq t(X_j)$ alors $t(X_k \cup X_j) \neq t(X_k) \cap t(X_j)$, $t(X_k) \cup t(X_j) \neq t(X_k)$ et $t(X_k \cup X_j) \neq t(X_j)$, donc $c(X_k) \cup c(X_j) \neq c(X_k)$ et $c(X_k) \cup c(X_j) \neq c(X_j)$. Aucun des deux éléments ne peut être éliminé de la classe [P], X_k et X_j menant à des fermetures différentes. Ce cas de figure est représenté par BC et BD dans l'arbre de la figure 2.4.

2.7.2 Principe de l'algorithme

CHARM commence par initialiser la classe [P] des 1-motifs fréquents et leur tidsets associés. L'arbre est parcouru de haut en bas à partir de la branche gauche vers la droite, chaque nouvelle paire candidate X_k et X_j est testée par rapport à son support et aux 4 propriétés énoncées précédemment, selon les résultats :

1. soit X_k est remplacé par $X_k \cup X_j$ et X_j (ainsi que tous les noeuds enfants de cet motif) est supprimé,
2. soit X_k est remplacé par $X_k \cup X_j$ et X_j est conservé et ajouté à la classe $[P_i]$,
3. soit X_j est remplacé par $X_j \cup X_k$ et X_k est conservé et ajouté à la classe $[P_i]$,
4. soit X_k et X_j sont ajoutés à la classe $[P_i]$.

Pour chaque branche l'algorithme CHARM-EXTEND est appelé récursivement jusqu'à épuisement de la branche considérée. Les propriétés des treillis de Galois, notamment des opérateurs de fermeture, font que les motifs sélectionnés sont des motifs fermés, sans avoir à effectuer de vérifications ultérieures.

Espace de recherche L'ordre de tri de la base de données est classiquement l'ordre lexicographique. Mais l'approche la plus prometteuse est de classer les motifs selon leur support. La motivation est de pouvoir augmenter la chance d'enlever des éléments de la classe [P]. Un examen rapide des propriétés décrites ci-dessus, nous indique que les propriétés 1 et 2 sont particulièrement intéressantes. Pour la propriété 1, la fermeture des deux motifs est égale et

ainsi on peut éliminer X_j et remplacer X_k par $X_k \cup X_j$. Pour la propriété 2 nous remplacerons X_k par $X_k \cup X_j$. Plus nous aurons d'occurences des propriétés 1 et 2, moins nous aurons de niveaux à investiguer. Par contre pour les propriétés 3 et 4 nous aboutissons à une addition d'éléments dans l'ensemble des nouveaux noeuds requérant des niveaux supplémentaires de recherche. Dans la mesure où on recherche à augmenter l'apparition de cas de propriétés 1 ou 2, le tri selon un support croissant est intéressant.

Critiques L'inconvénient de cet algorithme est qu'il nécessite un espace de stockage important du fait de la nécessité de stocker les motifs et leur tidset.

2.7.3 Pseudo-code

Algorithme 1 : Algorithme CHARM

Entrées : $D, min - sup$

```
1 [P] ={X_i × t(X_i) : X_i ∈ I ∧ σ(X_i) ≥ min − sup}
2 CHARM-EXTEND ([P], C = ∅)
3 retourner C // tous les ensembles fermés
```

CHARM trie la base et appelle CHARM-EXTEND.

Algorithme 2 : Algorithme CHARM-EXTEND

Entrées : $[P], C$

```
1 pour chaque X_i × t(X_i) dans [P] faire
2     [P_i] = ∅ et X = X_i
3     pour chaque X_j × t(X_j) dans [P], avec X_j ≥_f X_i faire
4         X = X ∪X_j et Y = t(X_i) ∩ t(X_j)
5         CHARM-PROPERTY([P], [P_i])
6     si [P_i] != ∅ alors
7         CHARM-EXTEND([P_i], C)
8     Supprimer [P_i]
9     C = C ∪ X si X n'est pas inclus ;
```

CHARM-EXTEND explore de manière récursive chaque branche de l'arbre, en faisant appel à CHARM-PROPERTY.

Algorithme 3 : Algorithme CHARM-PROPERTY

Entrées : [P], $[P_i]$

```
si σ(X) ≥ minsup alors
    si t(X_i) == t(X_j) alors
        // Propriété 1
        Enlever X_j de [P]
        Remplacer tous X_i par X
    sinon si t(X_i) ⊂ t(X_j) alors
        // Propriété 2
        Remplacer tous X_i par X
    sinon si t(X_i) ⊃ t(X_j) alors
        // Propriété 3
        Enlever Xj de [P]
        Ajouter X × Y à [P_i]
    sinon
        //Propriété 4
        Ajouter X × Y à [P_i]
```

CHARM-PROPERTY, analyse les propriétés des deux motifs considérés, traite ce cas et on revient à CHARM-EXTEND.

2.7.4 Déroulement de l'algorithme

En utilisant notre contexte d'extraction, voyons tout d'abord, avec un support minimal relatif de 50% quels sont les motifs infréquents, les motifs fréquents non fermés, et enfin les motifs fréquents fermés, représentés dans la figure 2.8.

Déroulement des opérations .

Ce déroulement est schématisé dans la figure 2.9.

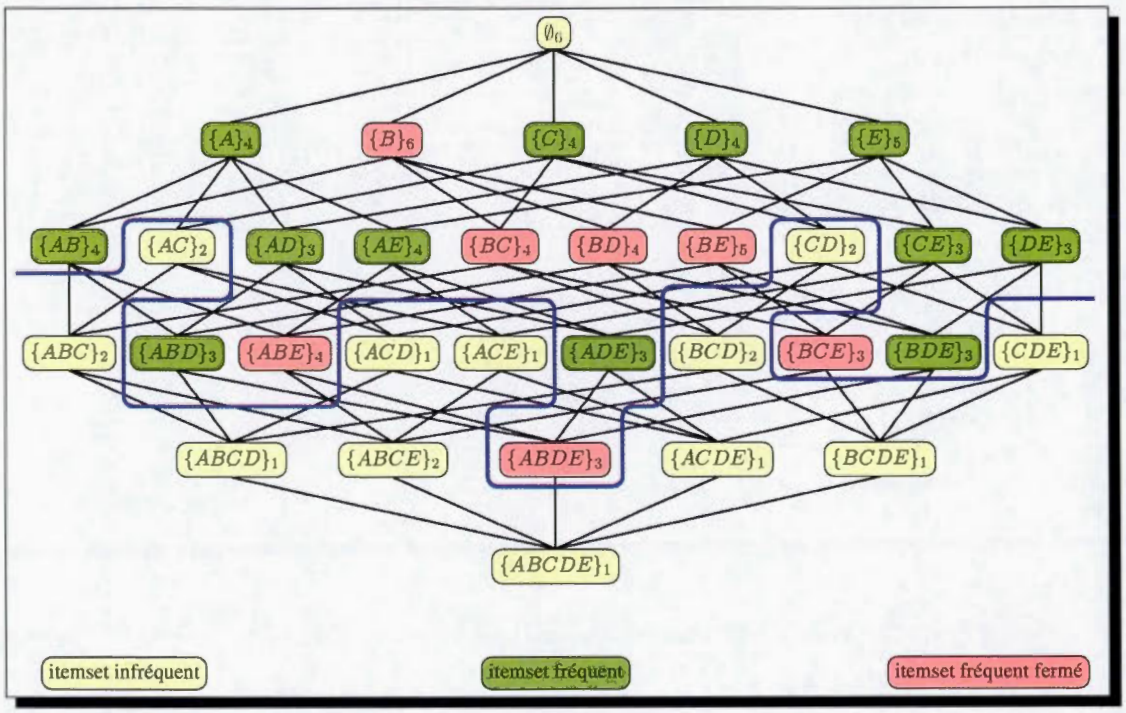

Figure 2.8 Itemsets fréquents et fermés associés au contexte $\mathcal{T}$, support minimal (σ_{min})=50%

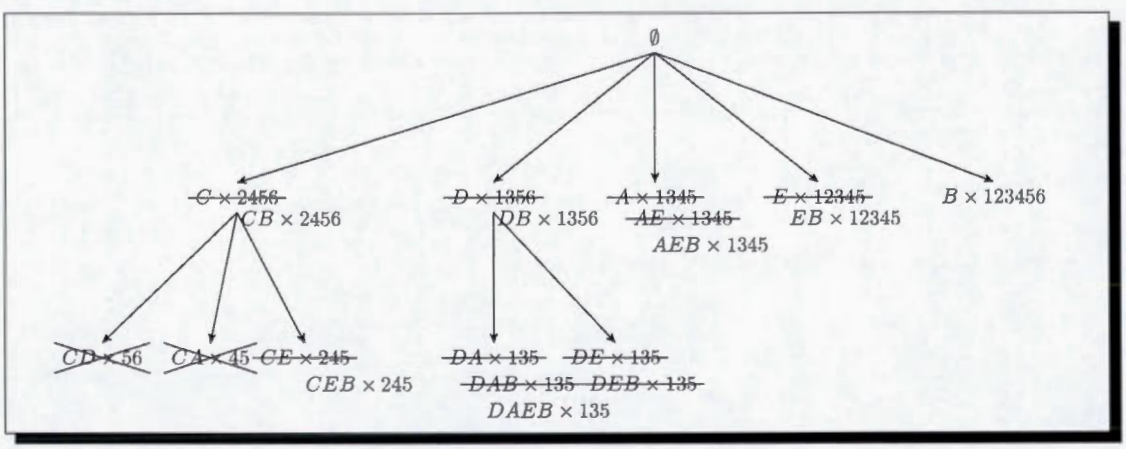

Figure 2.9 Progression de la recherche avec CHARM ($\sigma_{min} = 50\%$)

poids $w(K)$ **d'un item** K somme du support des 2-motifs fréquents le contenant :

$w(K) = \sum_{KL \in F_2} \sigma(KL)$

Ainsi, en ce qui concerne la branche A de l'arbre de recherches on a :

- AB avec un support de 4,
- AC n'est pas pris en compte car son support de 2 le classe dans les motifs non-fréquents,
- AD avec un support de 3,
- AE un support de 4.

Ce qui fait un poids total pour A $w(A) = 11$.

Au départ on trie les items selon l'ordre croissant de leur poids ce qui nous donne l'ordre suivant : C, D, A, E, B pour un poids respectif de 7, 10, 11, 15, 17.

On applique alors l'algorithme CHARM-EXTEND sur le noeud C, CD et CA sont éliminés car infréquents. Ensuite on considère C et E : $t(C) \neq t(E)$ alors on applique la 4e propriété et on ajoute CE à [C]. Enfin examinons C et B : $t(C) \subset t(B)$ la 2e propriété s'applique et on remplace C par CB et l'élément CE par CEB. Un appel récursif de CHARM-EXTEND à la classe [CE], celle-ci ne possédant qu'un seul élément CEB celui-ci est ajouté à l'ensemble des motifs fréquents fermés $\mathcal{C}$, retournons à C devenu CB la branche est complète donc on ajoute CB à $\mathcal{C}$.

Ensuite on passe à la branche D : t(D)≠t(A), donc on ajoute DA à [D], même chose pour D et E et on ajoute DE à [D]. Considérons B, t(D)⊂t(B), donc D est remplacé par DB, DA par DAB, DE par DEB ; enfin examinons le cas de DAB et DEB : t(DAB) == t(DEB) on est dans le cas où la 1re propriété s'applique DAB et DEB fusionnent et deviennent DAEB, la branche étant terminée on ajoute DB et DAEB à $\mathcal{C}$.

La branche A est analysée : t(A)⊂t(E) A devient AE et est ajouté à [A], t(A)⊂t(B) AE devient AEB, la branche A étant terminée AEB est ajouté à $\mathcal{C}$. De même pour la branche E : t(E)⊂t(B) E est remplacé par EB, la branche de E est terminée et on ajoute EB à $\mathcal{C}$. La dernière branche est B, qui est terminée donc on ajoute B à $\mathcal{C}$.

Ainsi pour résultat final nous avons l'ensemble des Motifs fréquents fermés C={CB, CEB, DB, DAEB, AEB, EB, B}, ce qui correspond aux résultats prévus dans le treillis Booléen de la figure 2.8.

2.8 Utilisations précédentes des techniques de fouille de données dans ce domaine

Contrairement aux tentatives précédentes qui utilisaient soit des recherches statistiques à partir d'un faible échantillon de SNPs (quelques dizaines ou voire quelques centaines) (Ritchie *et al.*, 2001) ce qui représente leur limite, ou des heuristiques (He *et al.*, 2009) dont on peut reprocher de rater des associations importantes, étant donné le trop grand nombre d'hypothèses testées et le nombre élevé de degré de liberté, dans l'étude (Gang *et al.*, 2012) un premier essai a été effectué de caractériser des motifs significatifs de SNPs sur des phénotypes qualitatifs (cas-contrôles).

Dans cette étude ils ont testés un nombre restreint de SNPs sur 3 cohortes différentes :

— survie courte (moins d'un an) ou longue (supérieure à trois ans) à un myélome multiple,

— rejet aigu (dans les six mois) ou non-rejet (au-delà de 8 ans) suite à une greffe de rein,

— cancer du poumon ou absence de cancer du poumon chez de gros fumeurs,

ces données sont regroupés dans le tableau 2.2, sélectionnés notamment pour l'implication des gènes qui leur sont associés dans les différents pathways biologiques impliqués dans les phénotypes étudiés.

Tableau 2.2 Détail des cohortes utilisées dans l'analyse de 2012, d'après (Gang *et al.*, 2012)

Phénotype	Nombre de SNPs	Nombre de patients	Nombre de cas	Nombre de contrôles
Survivants	2 755	143	70	73
Cancer du poumon	3 428	195	96	99
Rejet du rein	3394	271	135	136

Techniquement, ils ont utilisé l'algorithme APRIORI qui est l'algorithme de référence des algorithmes à niveau (Agrawal *et al.*, 1993), son utilisation est possible étant donné la taille réduite de l'ensemble des objets (ici des SNPs).

Les limites de cette méthode sont le nombre de SNPs non analysés, ce qui suppose la non-implication de zones non géniques dans la régulation de l'expression des gènes qui est de plus en plus remise en cause selon les dernières études.

Par rapport à notre expérimentation les principales différences sont :

— le nombre de SNPs analysés,

— la caractérisation de SNPs à risque et de SNPs protecteurs,

— la détermination d'un mode d'expression dominant ou récessif.

2.9 Conclusion

Nous avons choisi de rechercher les motifs fréquents fermés pour différentes raisons dont la réduction du nombre de résultats à traiter en éliminant les motifs intermédiaires sous-ensemble des motifs fréquents fermés, sans pour autant perdre de la significativité ou de l'exhaustivité des résultats. D'autre part l'algorithme CHARM est adapté à ce type de données, denses et importantes en taille.

Nous allons classer les SNPs en deux catégories SNPs à risque et SNPs protecteurs afin de caractériser des motifs de SNPs fréquents fermés dans une population (cas ou contrôle), d'une catégorie dont la présence est significativement différente dans la population opposée.

Dans l'objectif de la problématique qui nous intéresse nous allons successivement :

— construire un contexte d'extraction à partir des données brutes dont nous disposons,

— mettre en oeuvre l'algorithme CHARM afin de trouver des motifs fréquents fermés,

— sélectionner les motifs fréquents fermés dont la fréquence est significativement différente dans la population cas et la population contrôle.

Dans cette analyse nous nous attendons à rencontrer des difficultés, dues principalement à la taille de l'ensemble d'objets analysés, et dans l'adaptation du mode d'expression (dominant ou récessif) des SNPs.

CHAPITRE III

SCHÉMA GÉNÉRAL DE L'ANALYSE PROPOSÉE

Notre analyse va chercher à caractériser des ensembles de SNPs dont la présence peut augmenter le risque de présenter une prédisposition pour une pathologie ou une caractéristique donnée, nous regrouperons ces notions sous le terme de phénotype. Pour ce faire, nous allons émettre des hypothèses dont nous tenterons d'évaluer la validité à postériori. Nous énoncerons ces hypothèses au fur et à mesure de leur apparition. Donc nous allons chercher des ensemble de SNPs dont le support est significativement différent entre une population cas présentant ce phénotype et une population contrôle ne présentant pas ce phénotype.

Ce qui nous amène à formuler la première hypothèse :

> Il existe des motifs de SNPs permettant de classer la population cas et la population contrôle de manière significativement différente.

Figure 3.1 Hypothèse 1

3.1 Principe général

Nous ne nous intéresserons qu'à des phénotypes qualitatifs, c'est à dire permettant, une classification en population cas et population contrôle.

Odds ratio (OR) : ratio de la population cas porteuse d'un variant génétique (ici un SNP) sur la population contrôle porteuse de ce même variant.

De ce fait nous avons pour chaque SNP une valeur pour l'OR, qui va nous permettre de classer les SNPs en :

SNPs à risques dont l'OR est supérieur à 1 et

SNPs protecteurs dont l'OR est inférieur à 1.

Ce qui nous amène à la deuxième hypothèse :

> Les SNPs dont l'OR est supérieur à 1 sont des SNPs favorisant l'apparition du phénotype, alors que ceux dont l'OR est inférieur à 1 sont des variations génétiques protégeant l'individu vis à vis de ce phénotype.

Figure 3.2 Hypothèse 2

L'hypothèse 2 a pour corollaire l'hypothèse 3 :

> L'absence d'un SNP protecteur est un facteur de risque, équivalent à la présence d'un SNP à risque.

Figure 3.3 Hypothèse 3

Ensuite nous allons chercher des motifs fréquents fermés de SNPs à risques dans la population cas dont la fréquence est significativement plus faible dans la population contrôle, et inversement des motifs fréquents fermés de SNPs protecteurs dans la population contrôle dont la fréquence est significativement plus faible dans la population cas.

3.2 Données de départ

Liste des cas et des contrôles dépend du phénotype choisi, nécessite un phénotype qualitatif.

Liste des SNPs génotypés dépend du type de puce ADN utilisée par la plate-forme de génotypage. Dans notre cas il s'agit de puces Affymétrix^{TM}Genome-Wide Human SNP Array 6.0.

Fichiers .bgen, Fichiers .gen Ces fichiers sont les résultats bruts du génotypage de chaque individu. Les fichiers .bgen sont sous forme binaires afin de réduire leur taille de stockage.

Après avoir fait tourner le logiciel qctool, on obtient un fichier .gen.

Fichiers .assoc Ces fichiers, issus des GWAS effectués par le logiciel SNPTest (Marchini et Howie, 2010), contiennent des données statistiques de chaque SNPs en fonction du phénotype étudié.

3.3 Pipeline général de traitement des données

Le pipeline s'effectue dans un centre de Calcul Québec (Guillimin) qui se situe physiquement à l'Université Mc Gill à Montréal. La communication avec ce centre se fait grâce à une connexion utilisant un protocole ssh. Tout le processus depuis le pré-traitement des données jusqu'au post-traitement des résultats, se fait à partir de scripts écrit en Perl. Les scripts Perl effectuent des taches de manière séquentielle, et lorsqu'une parallélisation des tâches est possible il crée un autre script perl lancé par un script bash qu'il crée également, et lance ce script bash, parallélisant les tâches tout en se mettant en dormance le temps que ces tâches soient terminées.

Le déroulement du script peut être séparé en 3 parties distinctes :

— construction du contexte d'extraction,
— recherche des motifs fréquents fermés,
— post-traitement des résultats (recherche des motifs significativement différents dans les deux populations).

3.3.1 Construction du contexte d'extraction

Cette partie peut être schématisée par la figure 3.4

3.3.1.1 Sélection des données

À partir du fichier assoc on va sélectionner les SNPs selon certains critères :

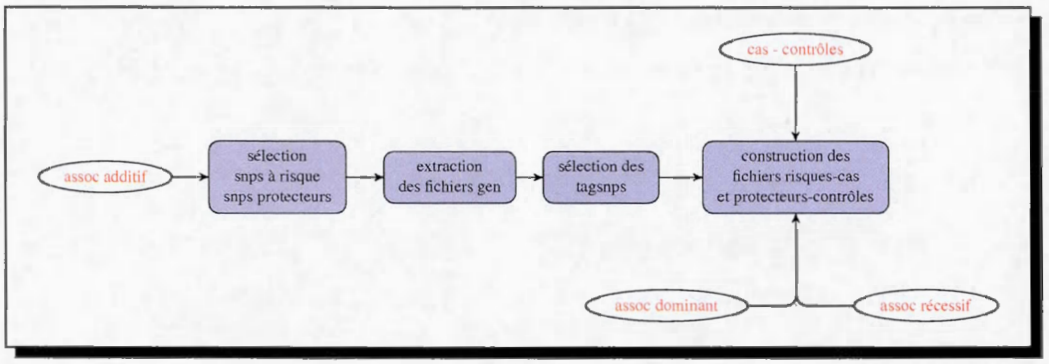

Figure 3.4 Schéma de fonctionnement du prétraitement des données

— élimination des SNPs dont le MAF est inférieur à 0,1 ou supérieur à 0,9 (un SNP trop rare ou trop fréquent dans la population totale ne saurait être discriminant entre la population cas et la population contrôle),

— sélection des SNPs dont l'OR est supérieur à 1 et dont la limite inférieure de l'intervalle de confiance de l'OR est également supérieur à 1 (sélection des SNPs dont on est sûr qu'ils sont à risque),

— sélection des SNPs dont l'OR est inférieur à 1 et dont la limite supérieure de l'intervalle de confiance de l'OR est également inférieure à 1 (sélection des SNPs dont on est sûr qu'ils sont protecteurs),

— sélection de tagsSNPs pour les SNPs à risque et les SNPs protecteurs (afin de réduire le déséquilibre entre les objets et les items dans la recherche des motifs fréquents fermés).

3.3.1.2 Préparation des données

Les fichiers gen sont sous forme de matrice, les lignes sont les SNPs et les colonnes les individus, chaque SNP est représenté par 3 colonnes :

1. colonne des homozygotes mineurs,
2. colonne des hétérozygotes,
3. colonne des homozygotes majeurs,

Un schéma de ce type de fichier est représenté dans le tableau 3.1.

Tableau 3.1 Structure d'un fichier gen

id / SNPs	1			2			...	n		
	aa	ab	bb	aa	ab	bb		aa	ab	bb
SNP1	0	1	0	0	1	0	...	1	0	0
SNP2	0	0	1	1	0	0	...	0	1	0

Cette représentation est compatible avec le modèle additif adopté pour les calculs statistiques des GWAS, mais ne correspond pas aux exigences des algorithmes de recherche de motifs fréquents et notamment de CHARM. Nous ne pouvons considérer que l'expression de tous les SNPs sélectionnés répondent au modèle dominant ou tous au modèle récessif.

Par conséquent nous avons émis une quatrième hypothèse :

> La valeur de p-value d'un SNP la plus faible selon le modèle récessif ou dominant, détermine son mode d'expression.

Figure 3.5 Hypothèse 4

À partir de fichiers assoc issus du logiciel SNPTest que l'on a fait tourner en forçant le modèle dominant et ensuite le modèle récessif, pour un même SNP on compare la valeur de p-value pour ces deux modèles et la valeur la plus faible déterminera son mode d'expression.

Donc pour un SNP *récessif* seuls les homozygotes mineurs sont marqués comme étant porteurs, alors que pour un SNP *dominant* les hétérozygotes et les homozygotes mineurs sont considérés comme étant porteurs.

Cette partie du pipeline aboutit à la construction des fichiers risques-cas et protecteurs-contrôles sous la forme de fichiers rcf dont nous montrons la structure dans la figure 3.6.

3.3.2 Traitement des données

Pour l'analyse de ces données, les SNPs à risque et les SNPs protecteurs seront traités séparément.

[*RelationalContext*]
Default Name
[*BinaryRelation*]
Name_of_dataset
1 | 2 | 3 | 4 | 5 | 6 | ← *id de chaque ligne séparés par un pipe*
a | c | d | t | w | ←*header de chaque colonne*
1 1 0 1 1
0 1 1 0 1
1 1 0 1 1
1 1 1 0 1
1 1 1 1 1
0 1 1 1 0
[*ENDRelationalContext*]

Figure 3.6 Structure d'un fichier rcf

L'algorithme utilisé est Charm qui recherche les motifs fermés de SNPs à risque dans les populations de cas et les motifs fermés de SNPs protecteurs dans les populations contrôle. Nous ferons varier le support minimal à 60 à 85 % par pas de 5 %.

Le fichier de sortie se présente sous la forme présentée dans la figure 3.7 :

1re ligne : path du fichier source,

2e ligne : nombre de lignes du fichier source (dans notre cas nombre de patients dans la population considérée),

3e ligne : nombre d'attributs (ici nombre de SNPs à risques ou protecteurs suivant la population considérée),

4e ligne : nombre d'attributs non vides (nombre de SNPs non vides dans la population),

5e ligne : nombre moyen d'attributs non vides par ligne (nombre moyen de SNPs présents par patient),

6e ligne : densité des données ($\frac{3^e ligne}{5^e ligne} \times 100$),

7e ligne : support minimal absolu et relatif,

8e ligne : algorithme choisi,

lignes jusqu'à l'avant dernière : liste des motifs fréquents fermés, avec la liste des SNPs formant le motif et le support de ce motif,

dernière ligne : nombre total de motifs fréquents fermés trouvés.

```
# Database file name : Documents/Maitrise/retinopathy/temporary_files/risques_cas.rcf
# Database file size : 1,681,461 bytes
# Number of lines : 766
# Total number of attributes : 1,086
# Number of non empty attributes : 1,086
# Number of attributes in average : 606,98
# Density : 55,89 %
# min_supp : 460, i.e. 60,05 %
# Chosen algorithm : dCharm [like Charm (v4), with diffsets]

{rs2607767} (464) +
{rs4455271, rs7624202, rs7613445, rs7632236} (462) +
{rs7624202, rs7613445, rs7632236} (463) +
{rs4455271, rs7624202, rs7632236} (463) +
⋮
{rs12485273, rs2541732, rs1027081, rs9798973, rs9647385, rs4260465, rs9682872} (462) +

# FCIs : 802,041
```

Figure 3.7 Structure d'un fichier de sortie de Coron alg :charm

Il s'agit maintenant de déterminer parmi ces motifs lesquels sont significativement différents entre la population cas et la population contrôle.

3.3.3 Post-traitement des données

Afin de tester si les populations porteuses d'un motif donné sont significativement différentes nous allons utiliser un test χ^2.

Ici nous n'avons qu'un degré de liberté, donc, en prenant exemple sur les SNPs à risque, nous calculons le support absolu théorique (P_{the}) du motif testé dans la population contrôle, d'aprés le support observé dans la population cas, puis nous calculons le support réel (P_{obs}) dans la population contrôle, et nous effectuons :

$$\chi^2 = \frac{(P_{the} - P_{obs})^2}{P_{obs}}$$

Notre hypothèse 0 est que la population porteuse de ce motif parmi les cas et celle porteuse parmi les contrôles sont identiques, étant donné que nous n'avons qu'un seul degré de liberté pour ce test, si le test χ^2 est supérieur à 3.48 (valeur donnée par la table de χ^2), alors les deux populations sont significativement différentes avec une certitude de 95 % .

Une autre condition à ajouter avant de sélectionner un motif est que la fréquence de la population porteuse parmi les contrôles soit inférieure à celle parmi les cas pour un motif de SNPs à risque, et inversement pour un motif de SNPs protecteurs.

Cette partie du pipeline est schématisée dans la figure 3.8

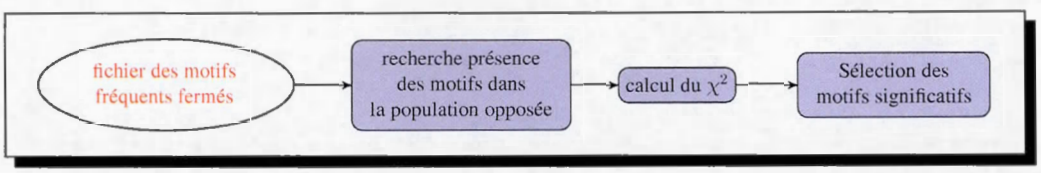

Figure 3.8 Schéma de fonctionnement du post-traitement des données

Dans l'interprétation des résultats nous allons considérer, en ce qui concerne les SNPs protecteurs, que les individus positifs sont ceux qui ne sont pas porteurs de ces motifs.

Par conséquent nous considèrerons que, pour un motif donné, l'absence d'un seul SNP le classe dans la population contrôle pour les SNPs à risque et dans la population cas pour les motifs de SNPs protecteurs.

3.4 Conclusion

Ce pipeline de traitement de données à été appliqué sur des données de génotypage à notre disposition, selon un phénotype que nous avons choisi selon sa solidité de définition et la qualité des résultats obtenus par l'analyse GWAS.

Nous allons tout d'abord présenter l'étude ADVANCE utilisée, le phénotype choisi et enfin nous ferons une analyse critique des résultats obtenus.

CHAPITRE IV

ÉTUDE D'UN EXEMPLE : PHÉNOTYPE RÉTINOPATHIE

4.1 Présentation de l'étude ADVANCE

ADVANCE est une étude clinique prospective effectuée dans 215 centres répartis sur 20 pays d'Asie, Australie, Europe et Amérique du Nord, s'étendant sur une période d'environ 5 ans (Patel, 2007), concernant 11 140 patients atteints de diabète de type 2, afin d'étudier l'effet d'une médication particulière sur l'hypertension artérielle et d'une surveillance accrue de la glycémie sur l'apparition de complications (principalement rénales et cardio-vasculaires).

Cette étude a obtenu l'approbation des comités d'éthique de chaque institution, pour chaque nation. Chaque patient a donné son accord éclairé pour sa participation.

Après une période de démarrage de 6 semaines, les patients ont été répartis aléatoirement en 4 groupes ; cette phase est représentée dans la figure 4.1.

Parmi ces 11 140 patients nous avons à notre disposition le génotypage de 3 449 patients de type caucasiens.

Les caractéristiques de ces deux groupes de sujets sont consignées dans le tableau 4.1.

4.2 Phénotype rétinopathie

La définition de ce phénotype est :

— antécédents de rétinopathie **ou**

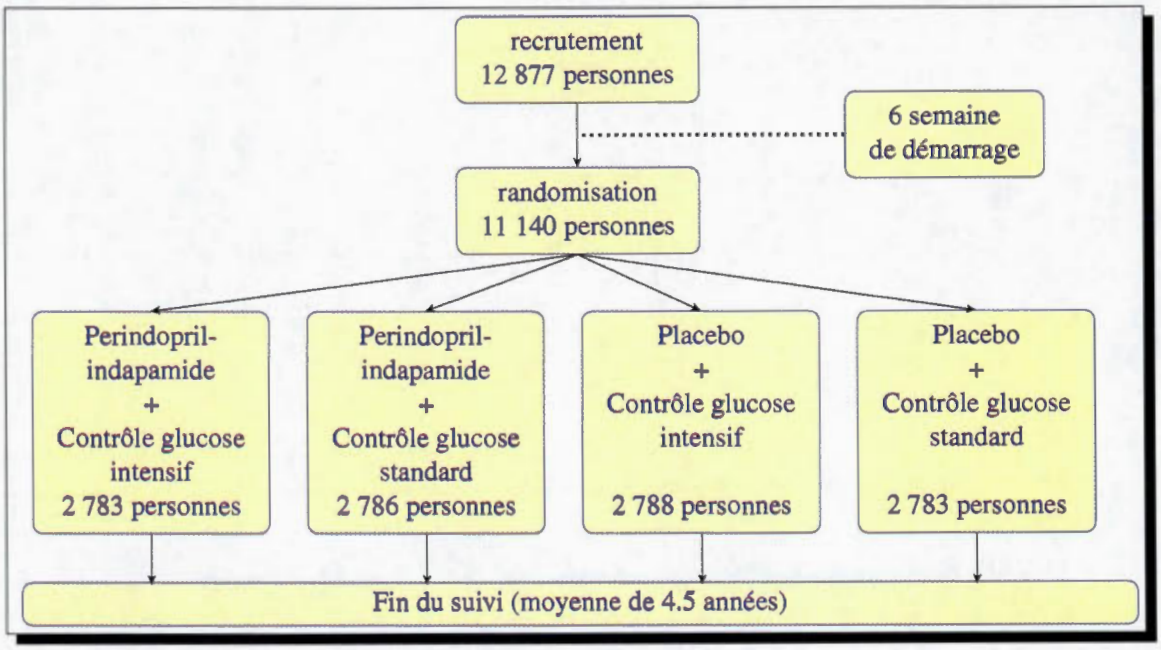

Figure 4.1 Schéma général de l'étude ADVANCE(Patel, 2007)(Committee, 2001)

— rétinopathie proliférative **ou**
— chirurgie ophtalmique au laser **ou**
— œdème maculaire **ou**
— cécité.

Le choix de ce phénotype relève de plusieurs considérations :

— d'une part l'existence d'un syndrome néphropathie, rétinopathie, neuropathy chez les diabétiques de type 2 (Raptis et Viberti, 2001), l'état de la rétine étant le 'reflet' de la fonction rénale,
— la néphropathie étant au centre des recherches menées au sein de l'équipe des Drs Tremblay et Hamet dont je fais partie,
— les caractéristiques des résultats des GWAS issus des analyses faites pour ce phénotype.

4.3 Prétraitement des données

4.3.1 Séparation Cas-Contrôles

Le phénotype rétinopathie au départ de l'étude présente sur les 3 449 patients gènotypés :

— 766 cas,

— 2 647 contrôles,

— 36 données manquantes,

donc sur les 3 413 personnes pour lesquels on possède les informations la population cas représente 22.4 % et la population contrôle 77.6 %.

4.3.2 Sélection des SNPs à risques et protecteurs

La micro Chip Affymétrix^{TM}Genome-Wide Human SNP Array 6.0, comporte 931 946 SNPs testés. SNPTEST, lors de son analyse, applique un premier filtre et enlève les SNPs dont la MAF est inférieure à 1 % dans la population totale, ce qui réduit le nombre de SNPs à 721 339. La première sélection portant sur l'analyse des OR de chaque SNPs aboutit à une première sélection de 18 892 SNPs protecteurs et 20 289 SNPs à risque. La détermination de tagSNPs par Haploview, et l'élimination des SNPs dont la fréquence est supérieure à 90 % dans la population totale, sélectionnent 5 475 SNPs protecteurs et 6 121 SNPs à risques. Ces chiffres sont collectés dans le tableau 4.2.

Tableau 4.1 Caractéristiques de la population d'ADVANCE et de la population génotypée(Patel, 2007)(Zoungas *et al.*, 2009)

	Population Totale			Population Génotypée		
	Hommes	Femmes	Total	Hommes	Femmes	Total
Effectif	6 405 (57%)	4 735 (43%)	11 140	2 219 (64%)	1 230 (36%)	3 449
Âge (moyenne en année)			66	66.7	66.9	66.8
groupe 1	1 595	1 198	2 783	558	288	846
groupe 2	1 618	1 198	2 786	575	323	898
groupe 3	1 607	1 181	2 788	554	289	843
groupe 4	1 595	1 188	2 783	532	330	862

Tableau 4.2 Évolution du nombre de SNPs sélectionnés au cours du prétraitement des données

	SNPs à risques	SNPs protecteurs
	Nombre Total	
	931 946	
SNPtest	721 339	
Sélection des SNPs à risque et des protecteurs	17 890	17 232
Tag SNPs	6 121	5 475

4.4 Traitement des données

Divers essais ont été effectués faisant varier la valeur du support de 60% à 85% par pas de 5%.

Pour chaque support, le nombre de motifs trouvés et le nombre maximal de SNPs dans un même motif pour les SNPs à risques dans la population cas et pour les SNPs protecteurs dans la population contrôle sont regroupée dans le tableau 4.3

Tableau 4.3 Évolution du nombre de motifs fréquents fermés et du nombre maximal de SNPs dans un motif pour les SNPs à risque et les SNPs protecteurs en fonction du support

	risques		protecteurs	
Support	Nb motifs	Nb max de SNPs	Nb motifs	Nb max de SNPs
85	411	2	571	2
80	950	3	2 701	3
75	61 306	4	151 004	4
70	365 174	5	2 095 412	5
65	11 853 644	6	40 919 633	6
60	+ 97 000 000	8	+ 138 000 000	7

Avec un support minimal de 60 %, nous n'avons réussi à obtenir qu'un résultat partiel, dû à une limite technique du serveur à notre disposition. En effet, le logiciel coron que j'utilise est un programme Java, et la taille de la machine virtuelle (que j'ai augmenté jusqu'à 100 Go) ne me permet d'obtenir que des résultats partiels, d'où le signe + placé devant ces résultats signifiant que les résultats réels sont supérieurs aux chiffres mentionnés. Par conséquent, j'ai décidé de traiter les résultats avec un support minimal de 65 %.

4.4.1 Post-traitement des donnéees

Un test de χ^2 permet de caractériser les motifs dont la présence est significativement différente dans la population cas et contrôle, le seuil de confiance du test à 95%. Les résultats sont consignés dans les tableaux 4.4 et 4.5.

Tableau 4.4 Évolution du nombre de motifs significatifs selon le support et le seuil de confiance du test chi_carré pour les SNPs à risques

	motifs		motifs significatifs à 95%	
Support	nombre	max SNPs	nombre	max SNPs
85	411	2	0	NA
80	950	3	0	NA
75	61 306	4	0	NA
70	365 174	5	0	NA
65	11 853 644	6	1	4
60	+ 97 000 000	8	NA	NA

Tableau 4.5 Évolution du nombre de motifs significatifs selon le support pour les SNPs protecteurs

	motifs		motifs significatifs à 95%	
Support	nombre	max SNPs	nombre	max SNPs
85	571	2	0	NA
80	2 701	3	0	NA
75	151 004	4	0	NA
70	2 095 412	5	0	NA
65	40 919 633	6	799	3
60	+ 138 000 000	7	NA	NA

Au vu de ces résultats, nous avons décidé d'analyser les motifs significatifs pour un support de 65 % comme support minimal des motifs fréquents et un seuil de confiance de 95 % pour le test de χ^2.

4.4.2 Analyse des résultats

4.4.2.1 Sensibilité et spécificité

Pour quantifier la validité d'un motif nous allons calculer la sensibilité et la spécificité paramètres classiques des études épidémiologiques (Fawcett, 2006) de ce motif pour la prédiction de la présence de ce phénotype selon les formules présentées dans le tableau 4.6.

Tableau 4.6 Sensibilité et spécificité

	cas	contrôles
Test positif	VP	FP
Test négatif	FN	VN

La sensibilité est définie par $\frac{VP}{VP+FN}$

et la spécificité par $\frac{VN}{VN+FP}$

4.4.2.2 Les motifs de SNPs à risques

L'analyse des SNPs à risque aboutit à plus de 11 853 644 motifs fréquent fermés avec un support minimal de 65 % variant de 1 à 8 SNPs, dont 1 seul présente un score de χ^2 supérieur à 3.48 (seuil de confiance de 95 %). La répartition des positifs et négatifs pour le motif de SNPs à risques significatif est indiqué dans le tableau 4.7.

Tableau 4.7 Répartition des cas et des contrôles selon la présence du motif à risque.

	cas	contrôles
Test positif	505	1 663
Test négatif	261	984

Ce qui aboutit à une sensibilité de 65,93 % et une spécificité de 37,17 %.

4.4.2.3 Les motifs de SNPs protecteurs

L'analyse des SNPs protecteurs avec un support minimal de 65 % indique 40 919 633 motifs fréquents fermés, dont la taille varie de 1 à 6 SNPs.

799 de ces motifs sont significatifs à 95 % au test de χ^2 avec un nombre maximal de 3 SNPs.

Dans la mesure où l'on cherche à caractériser les facteurs génétiques favorisant l'apparition de tel ou tel phénotype, pour les SNPs protecteurs les vrais positifs vont être les individus n'étant pas porteurs du motif considéré dans la population cas et par conséquent les vrais négatifs sont les patients porteurs de ce motif dans la population contrôle.

Ainsi pour le motif ayant obtenu le score le plus élevé au test de χ^2 , nous avons 2 014 patients contrôles qui en sont porteurs, et 492 parmi les cas.

Ce qui nous donne la répartition décrite dans le tableau 4.8 :

Tableau 4.8 Répartition des cas et des contrôles selon la présence du motif MP1.

	cas	contrôles
Test positif	274	633
Test négatif	492	2 014

Ce qui correspond à une sensibilité de 35,77 % et une spécificité de 23,91 %

4.4.3 Score GV

Reprenons le cas du motif significatif pour les SNPs à risques :

Dans la population de départ le pourcentage de cas est de $\frac{766 \times 100}{3413} = 22.44\%$.

Dans la population porteuse de ce motif à risque significatif on a :

cas 505

controles 1 663

total 2 168

Soit un pourcentage de cas de $\frac{505 \times 100}{2168} = 23,29\%$.

Afin de quantifier cet `enrichissement` de la population en cas on peut définir un score nommé GV = pourcentage de cas dans la population porteuse du motif ÷pourcentage de cas dans la population de départ.

Ce score est défini dans la figure 4.2.

GV score GV
PT population totale
PC population cas
VP vrais positifs
FP faux positifs

$$GV = \frac{\frac{VP}{VP+FP}}{\frac{PC}{PT}} \iff GV = \frac{VP \times PT}{(VP+FP) \times PC}$$

Figure 4.2 Définition du score GV

Afin de souligner l'importance d'un tel enrichissement, imaginons une étude clinique nécessitant une population de X patients atteint d'un phénotype.

Posons :

— P_{T_1} population totale nécessaire à l'étude sans l'utilisation de notre recherche,
— P_{T_2} population totale nécessaire à l'étude avec l'utilisation de notre recherche,
— X nombre de cas requis,
— P_{Cas} nombre de cas,
— P_{Cont} nombre de contrôles,
— VP vrais positifs,
— FP faux positifs.

$$P_{T_1} = \frac{X \times (P_{Cas} + P_{Cont})}{P_{cas}}$$

$$P_{T_2} = \frac{X \times (VP + FP)}{VP}$$

$$\text{Soit } E = \frac{P_{T_1} - P_{T_2}}{P_{T_1}}$$

$$E = \frac{X(\frac{P_{Cas}+P_{Cont}}{P_{Cas}} - \frac{VP+FP}{VP})}{X \times \frac{P_{Cas}+P_{Cont}}{P_{Cas}}}$$

$$\begin{cases} E = 1 - \frac{P_{Cas} \times (VP+FP)}{(P_{Cas}+P_{Cont}) \times VP} \\ GV = \frac{VP \times (P_{Cas}+P_{Cont})}{(VP+FP) \times P_{Cas}} \end{cases} \implies E = 1 - \frac{1}{GV}$$

Nous appellerons E le facteur d'économie.

Les caractéristiques du motif à risque significatif sont indiquées dans le tableau 4.9

Tableau 4.9 Tableau des caractéristiques du motif à risque

nombre de SNPs	VP	FP	χ^2	Sensibilité(%)	Spécificité(%)	GV	E(%)
4	505	1 663	3,86	65,93	37,17	1,038	3,66

Les caractéristiques des 10 motifs protecteurs, dont le score GV est le plus élevé, sont définies dans le tableau 4.10

Tableau 4.10 Tableau des 10 motifs protecteurs ayant les meilleurs scores GV

nombre de SNPs	VP	FP	χ^2	Sensibilité(%)	Spécificité(%)	GV	E(%)
3	274	633	14,15	35,77	23,91	1,346	25,73
3	274	634	14,67	35,77	23,95	1,345	25,65
3	266	635	11,62	34,73	23,99	1,316	24,00
3	262	636	10,44	34,20	24,03	1,300	23,10
3	250	610	9,16	32,64	23,04	1,296	22,82
3	258	633	9,61	33,68	23,91	1,291	22,51
3	245	602	8,47	31,98	22,74	1,289	22,43
3	259	637	9,58	33,81	24,06	1,288	22,38
3	253	623	9,03	33,01	23,54	1,287	22,31
3	256	631	9,23	33,42	23,84	1,286	22,26

4.4.4 Validation

Le moyen idéal de valider ces résultats serait de disposer d'une population génotypée comparable (100% de diabétiques de type 2) et génotypée selon la même technologie, et de confronter

nos résultats à cette nouvelle cohorte.

Malheureusement je ne disposais pas d'une telle base de patients, cependant nous avons pu effectuer un début de validation selon une nouvelle approche.

Dans le cadre de l'étude Advance, nous disposons de données s'étalant sur 5 ans, les analyses effectuées l'ont été sur les populations cas et contrôles au départ de l'analyse.

Par conséquent on peut déterminer combien de personnes, qui se sont révélées des cas 5 ans après, avaient été détectées d'après les données de départ.

Au cours des 5 ans qu'a duré l'étude Advance, 41 personnes au départ indemnes de rétinopathie, ont présenté ce phénotype.

Pour le motif à risques le nombre de cas nouveaux est de 25, alors que pour les 10 motifs protecteurs possédant le score GV le plus élevé ce nombre varie de 8 à 13.

De cette constatation nous pouvons conclure que le score GV ne suffit pas à lui seul de déterminer la validité d'un motif détecté nous devons également considérer les valeurs de sensibilité et de spécificité, qui sont dans les cas des motifs à SNPs protecteurs notoirement faibles.

4.4.5 Discussion

4.4.5.1 Comparaison avec les résultats obtenus avec l'analyse GWAS du phénotype rétinopathie

Le motif de SNPs à risque

Ce motif comporte 4 SNPs qui sont des tagSNPs, voici les SNPs auxquels ils sont associés :

- rs6009099 n'est associé qu'à lui-même,
- rs7562149 est associé à 3 autres SNPs rs360804,rs11682530,rs360801,
- rs7913401 n'est associé qu'à lui-même,
- rs918457 n'est associé qu'à lui-même.

Donc nous avons 7 SNPs représentés par ce motif, regardons ce que nous avons dans nos résul-

tats de GWAS et dans la littérature :

- rs6009099 situé sur le chromosome 22 dans un intron du gène TBC1D22A dont le rôle n'a jamais été remarqué pour son activité dans la rétinopathie, ou une pathologie rénale, on signale ce gène pour son rôle dans la longévité (Yashin *et al.*, 2010),
- rs7562149 situé chromosome 2 dans un intron du gène EHBP1 dont le dont le rôle n'a jamais été remarqué pour son activité dans la rétinopathie, ou une pathologie rénale, on signale ce gène dans trois études dont une porte sur le taux de lipides sanguin (Consortium *et al.*, 2013),et deux études portant sur le cancer de la prostate (Eeles *et al.*, 2009) (Gudmundsson *et al.*, 2008),
- rs360804 situé chromosome 2 dans un intron du gène EHBP1,
- rs11682530 situé chromosome 2 dans un intron du gène EHBP1,
- rs360801 situé chromosome 2 dans un intron du gène EHBP1,
- rs7913401 situé chromosome 10 entre le gène KIAA1217 signalé dans une étude portant sur les performances cognitives (Cirulli *et al.*, 2010), et le gène OTUD1 cité dans une étude sur l'apparition précoce de l'obésité extrême (Wheeler *et al.*, 2013), d'autre part, c'est le seul gène qui a été remarqué avec une p-value dans les GWAS que nous avons effectués, pour les phénotypes de triglycérides, et le déclin d'eGFR qui est un phénotype rénal,
- rs918457 situé chromosome 5 à proximité du gène LOC102467226, dont le rôle est inconnu.

Les motifs de SNPs protecteurs

Il est à noter que malgré les valeurs élevées du score GV, les faibles valeurs de sensibilité et de spécificité, corroborées par le petit nombre de nouveaux cas qui avaient été détectés, montrent le faible intérêt que représentent ces résultats.

Aussi nous n'irons pas plus loin dans l'examen des motifs significatifs, pour les SNPs protecteurs.

4.4.5.2 Discussion sur les hypothèses émises

Rappelons les 4 hypothèses émises :

Hypothèse 1 : Il existe des motifs de SNPs permettant de classer la population cas et la population contrôle de manière significativement différente.

Hypothèse 2 : Les SNPs dont l'OR est supérieur à 1 sont des SNPs favorisant l'apparition du phénotype, alors que ceux dont l'OR est inférieur à 1 sont des variations génétiques protégeant l'individu vis à vis de ce phénotype.

Hypothèse 3 : L'absence d'un SNP protecteur est un facteur de risque, et est équivalent à la présence d'un SNP à risque.

Hypothèse 4 : La valeur de p-value d'un SNP la plus faible selon le modèle récessif ou dominant, détermine son mode d'expression.

On a pu effectivement caractériser des motifs de SNPs permettant de classer la population cas et la population contrôles de manière significativement différente.

Donc **l'hypothèse 1 est vérifiée**.

Les piètres résultats obtenus en ce qui concerne les SNPs protecteurs n'ont pas permis de valider les hypothèses 2 et 3. Cependant les problèmes techniques auxquels nous avons été confrontés ne permettent pas de rejeter définitivement ces hypothèses.

Les hypothèses 2 et 3 ne sont pas vérifiées sans pour autant être rejetées.

La détermination du mode d'expression en employant les valeurs des p-value pour le mode récessif et dominant a permis d'obtenir des résultats conformes aux objectifs fixés.

L'hypothèse 4 est vérifiée

4.4.5.3 Critique

Dans la sélection des SNPs à risque et des SNPs protecteurs, on élimine systématiquement les SNPs dont l'intervalle de confiance permet d'assurer qu'ils sont bien à risque ou protecteurs, ce filtre élimine beaucoup de SNPs dont le rôle est certainement à risque ou protecteurs, ce filtre est donc très certainement trop drastique et nous fait ignorer des SNPs jouant un rôle important.

La détermination du mode d'expression utilisant le mode récessif ou dominant pourrait être faite différemment en utilisant les OR des homozygotes et des hétérozygotes dans les populations cas et contrôles.

En effet pour un SNP à risque pouvant être assimilé à un SNP récessif l'OR des homozygotes sera très proche de l'OR de l'allèle au complet, alors que c'est l'addition de l'OR des hétérozygotes et des homozygotes qui sera très proche de l'OR de l'allèle. Ce modèle est difficile à mettre en œuvre car beaucoup de SNPs sont intermédiaires et des critères de sélection seront à établir.

Tous ces développements possibles auront pour but d'affiner les résultats sans remettre en cause la validité des résultats obtenus par l'analyse présente.

CONCLUSION

L'utilisation d'algorithmes de recherche de motifs fréquents fermés pour l'analyse de données de génotypage nous a fourni des résultats encourageant permettant d'obtenir des motifs suffisamment contrastant entre les populations cas et contrôles pour permettre un enrichissement significatif des cas dans la population positive, aboutissant à de potentielles utilisations dans le domaine de la recherche clinique.

Cet enrichissement est confirmé par les données dont on dispose 5 ans après le début de l'étude.

L'analyse de cet enrichissement est quantifié par un score que nous avons appelé GV.

Les voies d'investigation futures pourraient permettre de trouver un moyen différent de convertir les données de génotypage en données binaires, de valider l'utilisation des tag SNPs, d'utiliser les capacités de prévision de l'utilisation simultanée de différents motifs à risques et de motifs protecteurs, ce qui implique l'utilisation de techniques d'analyse de motifs disjoints ou non.

Cependant, d'ores et déjà, on peut constater l'avantage de ces méthodes sur les procédés classiques de GWAS. En effet le choix des SNPs nécessaires pour obtenir une prévision donnée est :

- standardisée par l'utilisation du score GV
- nécessite moins de SNPs par notre méthode
- correspond mieux au mode d'expression supposé de syndromes polygéniques

Perspectives

L'utilisation des techniques de fouilles de données dans l'analyse des données de génotypage est très peu exploitée jusqu'à maintenant (Gang *et al.*, 2012), et la caractérisation de motifs de SNPs significatifs est une première dans ce domaine.

Les améliorations à apporter à cette démarche sont multiples.

- Détermination du mode d'expression des SNPs.
- Paramétrage de la détermination des tagsSNPs, afin de réduire le nombre de SNPs candidats.
- etc..

D'autre part le développement de nouveaux algorithmes permettant la mise en évidence de motifs à contraste dans deux populations différentes. Cette démarche à déjà été explorée avec la mise au point de divers algorithmes tels que STUCCO (Bay et Pazzani, 1999), CP_Tree (Ramamohanarao *et al.*, 2005), ou l'algorithme CIGAR (Hilderman et Peckham, 2005), mais de nombreuses adaptations doivent être apportées à ces algorithmes afin d'être applicables à notre domaine.

Il semblerait également possible d'élargir de telles techniques aux données quantitatives, avec l'algorithme *Fuzzy Data Mining Algorithm* (Hong *et al.*, 1999)

Enfin, on doit investiguer afin de mettre au point un algorithme tenant compte du modèle quantitatif, plutôt que d'essayer d'appliquer un modèle `dominant-récessif`. Ces voies semblent prometteuses pour cette nouvelle approche de la compréhension de la composante génétique des pathologies polygéniques.

ANNEXE A

DESCRIPTION DÉTAILLÉE DU PIPELINE MIS EN PLACE.

Les scripts sont inclus dans la deuxième partie des Annexes.

Le script traitement.pl, que l'on peut consulter en Annexe B.1 se divise en deux parties :

- Pré-traitement des données, c'est le nettoyage des données et la construction du contexte d'extraction,
- Traitement des données, où on applique l'algorithme CHARM sur les données préparées.

Ce script est lancé sur le serveur Guillimin de Calcul Québec par la commande bash traitement.sh Annexe B.2. Ce script génère plusieurs scripts perl, lancés par des commandes bash, ou directement des commandes bash lançant des logiciels particuliers.

La partie post-traitement des données, qui consiste en la recherche des motifs fréquents fermés significativement différents entre les populations cas et contrôles, s'effectue par plusieurs scripts, chaque sortie de CHARM nécessitant un traitement particulier.

A.1 Construction du contexte d'extraction

A.1.1 Fonction cas_contrôles

Identifie les populations cas et les populations contrôles.

A.1.2 Détermination des SNPs à risque et des SNPs protecteurs

Les fichiers de départ sont des fichiers .assoc issus de SNPtest qui sont les résultats des GWAS effectués à partir du phénotype concerné. Ces fichiers sont au nombre de 3 :

- GWAS effectué en suivant le modèle dominant
- GWAS effectué en suivant le modèle récessif
- GWAS effectué suivant le modèle additif

Des deux premiers on extrait la p-value, on les compare pour chaque SNP, si la p-value du modèle dominant est inférieure à celle du modèle récessif, on considère que son mode d'expression est plus proche du modèle dominant, et dans le cas contraire il sera plus proche du modèle récessif.

Dans le cas d'un SNP `récessif` on ne considèrera comme étant porteurs que les individus homozygotes de l'allèle mineur.

Par contre pour un SNP `dominant` on considèrera également les hétérozygotes.

Du troisième fichier .assoc issu du GWAS du modèle additif, on va extraire divers informations :

- le nom du SNP
- le maf (minimal allele frequency) de l'allèle mineur
- l'OR de l'allèle chez les homozygotes
- la borne minimale de l'intervalle de confiance à 95% de l'OR chez les homozygotes
- la borne maximale de l'intervalle de confiance à 95% de l'OR chez les homozygotes
- l'OR de l'allèle
- la borne minimale de l'intervalle de confiance à 95% de l'OR de l'allèle
- la borne maximale de l'intervalle de confiance à 95% de l'OR de l'allèle
- la valeur de la value

Les SNPs sont sélectionnés et distribués selon certains critères paramétrables :

- le maf (minor allele frequency) c'est à dire la fréquence de l'allèle mineur (celui que l'on considère) dans la population étudiée doit être supérieure à une certaine valeur seuil ici nous avons choisi 1%,

— si l'odds-ratio (OR) chez la population homozygote de l'allèle est supérieur à 1 et que la valeur inférieure de l'intervalle de confiance à 95 % est également supérieure à 1, alors le SNP est considéré comme étant à risque,

— inversement si l'OR dans la population homozygote est inférieure à 1 et que la valeur supérieure de l'intervalle de confiance à 95 % est également inférieure à 1, alors le SNP est considéré comme étant protecteur.

Ces opérations sont effectués par les fonctions dominant_recessif, modele, data et construction_fichiers.

A.1.3 QCTOOL

Ce logiciel mis au point par Jonathan Marchini et Gavin Band de l'université d'Oxford Royaume Uni, est disponible gratuitement sur le site de cette université, à l'adresse suivante : http://www.well.ox.ac.uk/ gav/-qctool/#overview.

Son utilisation dans le pipeline de pré-traitement des données permet de transformer les fichiers .bgen en fichiers .gen pour les listes des SNPs à risque et ceux protecteurs.

Cette opération est effectuée par la fonction qctool, qui écrit le script bash parallele_pretraitement_1.sh, à l'annexe B.3 qui parallélise qctool pour les 22 chromosomes et successivement pour les SNPs à risques et les SNPs protecteurs.

Le script traitement.pl est maintenu en dormance jusqu'à l'écriture des 44 fichiers .log.

A.1.4 Sélection de TagSNPs

Ensuite on cherche à établir une liste de tagSNPs issus de la liste risques et de la liste protecteurs. En effet les SNPs en LD vont apparaître (par définition) très fréquemment ensemble, et vont faire partie de motifs fréquents, mais la mise en évidence de ces motifs ne va pas fournir de connaissances supplémentaires, car on connaissait déjà leur existence.

Aussi on va chercher à ne prendre qu'un seul représentant tagSNP par haploblock. Ce processus

s'effectue avec le logiciel Haploview (Barrett *et al.*, 2005) qui nécessite en entrée un format de fichier particulier : fichier .ped et un fichier .info. Ces fichiers sont séparés par chromosome, il existe donc 22 fichiers .ped et 22 fichiers .info, pour chacune des catégories à risques et protecteurs.

Ce fichiers vont être obtenus après traitement par le script obtention_fichiers_ped_info.pl qui est écrit par traitement.pl, et lancé par parallele_pretraitement_2.sh construit également par traitement.pl.

Ces scripts sont disponibles respectivement aux annexes B.4 et B.5.

A.1.4.1 obtention_fichiers_ped_info.pl

Les fichiers `.info` comportent 2 colonnes :

- L'identité des SNPs
- Leur position dans le chromosome.

Par contre les fichiers .ped ont une structure plus complexe comme indiqué dans le tableau A.1

Chaque ligne correspond à un individu.

Les fichiers .ped contiennent 8 colonnes,

- identifiant du pedigree : identifie la famille du patient, dans notre cas on mettra le même identifiant que celui du patient,
- identifiant du patient,
- identifiant du père, 0 si inconnu
- identifiant de la mère, 0 si inconnu
- sexe 1= homme, 2= femme, -9 si ignoré,
- statut vis à vis du phénotype 0 = inconnu, 1 = contrôle, 2 = cas,
- marqueurs génotypés avec 2 colonnes par marqueurs (une par allèle) notées A, C, G ou T.

Tableau A.1 Structure d'un fichier ped

ped-id	ind-id	pere-id	mere-id	sex	status	SNP_1		SNP_2		$\cdots$	SNP_n	
1	1	0	0	-9	1	A	T	G	A	$\cdots$	T	G
2	2	0	0	-9	2	C	G	T	G	$\cdots$	T	G
$\vdots$												
k	k	0	0	-9	1	C	A	C	C	$\cdots$	G	A

A.1.4.2 Haploview

Le traitement par Haploview permet de caractériser les différents haploblocs (cf figure A.1) définis par les SNPs à risques et SNPs protecteurs pour chaque chromosome.

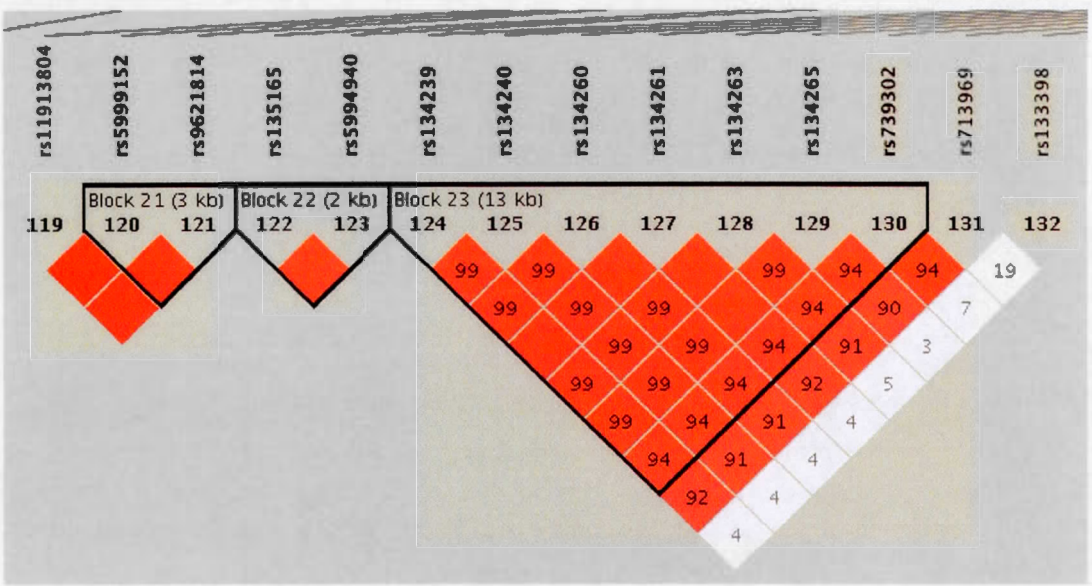

Figure A.1 représentation d'haploblocs par Haploview

Ensuite Haploview peut analyser ces blocs et déterminer une série de SNPs représentant des autres SNPs d'un bloc (ce bloc peut ne contenir qu'un seul SNP dans ce cas ce SNP est son propre tag SNP).

Ce sont ces tagSNPs qui vont donner les listes définitives de SNPs à risque et de SNPs protecteurs.

Tableau A.2 Structure du fichier prov

	SNP_1	SNP_2	$\cdots$	SNP_n
id_1	1	0		1
id_2	0	0		1
$\vdots$				
id_n	0	1		0

Le programme Haploview est lancé par le script haplo.sh consultable à l'annexe B.6

Les indications et le schéma de ce script sont formulés pour les SNPs à risques afin de simplifier ces explications, le pendant existe pour les SNPs protecteurs.

A.1.5 Fonction concatenation

Les fichiers .gen de chaque chromosome sont concaténés en un seul fichier par type de SNPs (risques ou protecteurs).

A.1.6 Fonction transformation

Partant du fichier issu des fichiers .gen où chaque SNP est représenté par 3 colonnes (voir le tableau 3.1), on va transformer cette information en une seule colonne.

Si le SNP est dominant, seront considérés comme porteurs d'un SNP les individus homozygotes et hétérozygotes, par contre si le SNP est récessif ne seront considérés porteurs que les individus homozygotes.

Ces résultats sont consignés dans le fichier risques_prov.txt, dont la structure est représentée dans le tableau A.2.

A.1.7 Fonction tag

À partir des tag SNPs sélectionnés par Haploview, le fichier risques_prov.txt est transformé en risques_prov2.txt où ne sont présents que les tag SNPs

A.1.8 Fonction elimination

Élimination des SNPs les plus fréquents dans la population, en effet les SNPs très fréquents dans une population ne pourraient être discriminants entre deux sous-ensembles de cette population, ce seuil est à ajuster en fonction de la cardinalité des populations cas et contrôles.

Dans notre cas nous avons choisi un seuil de 90 %.

Le fichier risques_d.txt est ainsi créé, ainsi que le fichier risques_retenus.txt, liste définitive des SNPs à risques retenus.

A.1.9 Fonction separation_cas_controles

Le fichier risques_prov2.txt est scindé en deux fichiers selon que l'individu appartient à la population cas ou la population contrôle. Cette séparation s'effectue en préparant un fichier comportant des commandes en `awk` extrayant les colonnes adéquates afin de construire les fichiers risques_cas_d.txt et risques_controles_d.txt

A.1.10 Fonction transposition_matrice

Les matrices sont jusqu'à maintenant sous la forme :

lignes SNPs

colonnes individus

or le but est d'analyser des matrices sous la forme :

lignes individus

colonnes SNPs

Ainsi que le détaille la figure A.2.

C'est cette transposition qu'effectue la fonction transposition_matrice.

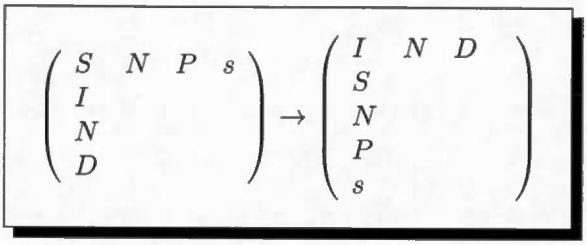

Figure A.2 Transposition des matrices

Elle prepare un fichier execute renfermant des commandes awk permettant cette transposition puis l'éxecute. Ce qui permet d'aboutir aux fichiers :

— risques_cas.txt

— risques_controles.txt

Le script de transposition se trouve à l'annexe B.7

A.1.11 Fonction construction_rcf

Met en forme le fichier risques_cas.txt, pour obtenir le fichier risques_cas.rcf

La structure du fichier risques_cas.rcf est détaillée dans la figure 3.6

A.1.12 Fonction fichiers_definitifs

Cette fonction ajoute l'id de chaque patient au début de chaque ligne et le nom de chaque SNPs en tant que header de chaque colonne.

A.2 Traitement des données

La fouille de données à proprement parler s'effectue sur la plate-forme Coron (Szathmary, 2006) (système open-source gratuit) disponible sur internet à l'URL : http://coron.loria.fr/site/index.php

Coron est lancé par la commande bash coron.sh présente à l'annexe B.8 On obtient donc les motifs fréquents fermés avec un support de 60%, 65%, 70%, 75%, 80% et 85% pour les SNPs à risque dans la population cas (ainsi que pour les SNPs protecteurs dans la population contrôle).

```
rs4679233 rs4532322 rs7832456 rs9816344    460
rs4679233 rs4532278 rs7812476 rs12    589
⋮
```

Figure A.3 Structure du fichier motis_4_snps

Maintenant il convient de trouver lesquels présentent une différence significative de répartition dans la population contrôle pour ce même phénotype.

C'est l'objectif de la troisième partie du pipeline, post-traitement des résultats.

A.3 Post-traitement des résultats

Nous prendrons pour exemple le fichier de sortie des SNPs à risques avec un support relatif de 65%.

Le schéma général de cette partie du pipeline est représenté dans la figure 3.8.

A.3.1 Transformation du fichier de sortie.

Le fichier de sortie output_risques_65.txt comporte tous les motifs fréquents fermés trouvés dans la base de données.

Ces motifs sont de taille variable et ne sont pas classés par taille, cette fonction va classer les motifs par taille dans des fichiers `motifs_n_snps.txt` où n désigne la taille de ces motifs.

Un exemple de fichier de sortie est donné dans la figure A.3.

Chaque ligne commence par le motif en lui même suivi par le nombre d'individus porteur de ce motif dans la population étudié (ici SNPs à risques dans la population cas).

Ce script Perl transformation_output.pl est à l'annexe B.9, et la commande bash le lançant post_traitement_1.sh est à l'annexe B.10

A.3.2 Décompte du nombre de patients porteurs d'un motif dans la population opposée

Il s'agit de dénombrer dans la population contrôle le nombre d'individus porteurs du motif considéré. Le script itemset_contraste.pl à l'annexe B.11, lancé par la commande bash post_traitement_2.pl à l'annexe B.12 effectue cette tâche il produit un fichier éxécutable pour 10 000 motifs (ce chiffre de 10 000 est ajustable afin de ne pas obtenir de fichiers trop lourds à éxécuter, tout en respectant la limite de 5 000 tâches envoyées simultanément au centre de calcul Guillimin).

Un extrait d'un script execute est en annexe B.13

Ensuite le script perl prep_parallel_rech.pl lancé par la commande bash post_traitement_3.sh ,que l'on va retrouver respectivement aux annexes B.14 et B.15, va écrire la commande bash post_traitement_4.sh qui lance les scripts execute. Un exemple de script post_traitement_4.sh est en annexe B.16

A.3.3 Calcul du test χ^2

Cette fonction calcule un test chi carré entre la population effectivement observée sur la population contrôle (pour les motifs de SNPs à risques) et la population théorique.

Ce test ne comporte qu'un degré de liberté, aussi il n'est considéré significatif à 95% que si il est supérieur à 3,48. Ne sont retenus que les motifs dont le test chi carré est supérieur à ce seuil et dont la population observée est inférieure à la population théorique. Enfin le script sépare les motifs significatifs en fonction de la longueur des motifs.

Ce script chi_2.pl est à l'annexe B.17 et le script bash post_traitement_5.sh le lançant à l'annexe B.18.

ANNEXE B

SCRIPTS

```
use strict;
use warnings;

  #$ARGV[0] = phenotype

my%mode;
my%tires;
my$nb=0;
my%rsid;

# Séparation des cas et des controles

sub cas_controles{
  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/retinopathy.
      sample");
  open(CAS, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/cas.txt");
  open(CONT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/controles.
      txt");
  chomp(my@tab=<IN>);

  for(my$i=2; $i<@tab; $i++){
    chomp(my@ligne=split/ /, $tab[$i]);
    if($ligne[3]==0){
      print CONT "$i;$ligne[3]\n";
    }elsif($ligne[3]==1){
      print CAS "$i; $ligne[3]\n";
    }
  }

  close(IN);
  close(CAS);
  close(CONT);
}

# Extraction de données pertinentes des fichiers dominant et récessifs

sub dominant_recessif{

  my$nom=$_[0];
  print "## ".localtime()."\n";
  print "## Construction du fichier ".$nom."\n";
  open(OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute");
  print OUT "grep ^s /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/".
      $ARGV[0]."_".$nom.".assoc >/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."
      /temporary_files/temp.txt\n";
  print OUT "awk -F \" \" '{print \$1 \" \" \$40 }' /gs/scratch/ggodefroid/
      Maitrise/".$ARGV[0]."/temporary_files/temp.txt >/gs/scratch/ggodefroid/
      Maitrise/retinopathy/temporary_files/".$nom.".txt\n";
  print OUT "rm /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
      temp.txt";
  close(OUT);
  system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute");

}
```

```
48 # Détermination du modèle applicable pour chacun des SNPs
49
50 sub modele{
51   open(DOM, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
         dominant.txt");
52   open(REC, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
         recessif.txt");
53   open(MOD, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
         mode.txt");
54   chomp(my@dom=<DOM>);
55   chomp(my@rec=<REC>);
56   my%dominant;
57   foreach(@dom){
58     chomp(my@ligne=split/ /,$_);
59     if($ligne[1] > 0){
60       $dominant{$ligne[0]}=$ligne[1];
61     }
62   }
63   foreach(@rec){
64     chomp(my@ligne=split / /,$_);
65     if($ligne[1] > 0 && exists $dominant{$ligne[0]}){
66       if($ligne[1]>$dominant{$ligne[0]}){
67         $mode{$ligne[0]}='d';
68         print MOD "$ligne[0]\td\n";
69       }
70       else{
71         $mode{$ligne[0]}='r';
72         print MOD "$ligne[0]\tr\n";
73       }
74     }
75   }
76
77 }
78
79 # Extraction des données pertinentes pour les SNPs sélectionnés
80
81 sub data{
82   open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
83
84   print OUT "grep ^id /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/".
         $ARGV[0].".assoc >/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
         temporary_files/temp.txt\n";
85   print OUT "grep ^s /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/".
         $ARGV[0].".assoc >>/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
         temporary_files/temp.txt\n";
86   print OUT "awk -F \" \" '{ print \$2 \" \" \$24 \" \" \$37 \" \" \$38 \" \"
          \$39 \" \" \$40 }' /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
         temporary_files/temp.txt >/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
         temporary_files/tmp.txt\n";
87   print OUT "rm /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
         temp.txt";
88   close(OUT);
```

```
89    system("chmod 0777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1"
          );
90    system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
91    system("rm /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
92  }
93  # Construction des fichiers risques.txt et protecteurs.txt
94
95  sub construction_fichiers{
96    open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/tmp.
          txt");
97    open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          temp.txt");
98    open(OUT1, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          risques.txt");
99    open(OUT2, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          protecteurs.txt");
100
101   chomp(@tab=<IN>);
102   chomp(my@header=split/ /, $tab[0]);
103   print OUT "$tab[0]\n";
104   for(my$i=1; $i<@tab; $i++){
105     chomp(my@ligne=split / /, $tab[$i]);
106     if(exists $mode{$ligne[0]}){
107       if($ligne[2] < 1 && $ligne[4]<1 && $ligne[1]>0.1){
108         print OUT "$tab[$i]\n";
109         print OUT2 "$ligne[0]\n";
110       }elsif($ligne[2]>1 && $ligne[3]>1 && $ligne[1] > 0.1){
111         print OUT "$tab[$i]\n";
112         print OUT1 "$ligne[0]\n";
113       }
114     }
115   }
116   close(IN);
117   close(OUT1);
118   close(OUT2);
119 }
120
121 # Intervention de qctool par la construction du fichier pretraitement_1.sh, et la mise en attente du
        script tant que toutes les tâches ne sont pas terminées
122
123 sub qctool{
124   system("mkdir /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
          fichiers_gen/");
125   system("mkdir /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
          fichiers_gen/risques");
126   system("mkdir /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
          fichiers_gen/protecteurs");
127
128   open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/pretraitement_1
          .sh");
129
130   print OUT "for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
          19 20 21 22;
```

```
131     print OUT "for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
            19 20 21 22;
132   do
133   echo \"#!/bin/bash
134           #PBS -l nodes=1:ppn=8,walltime=01:00:00
135           #PBS -A cvn-715-aa
136           #PBS -j oe
137           #PBS -V
138           #PBS -N qctool_protec_\$i
139   /sb/project/cvn-715-aa/bin/qctool -g /sb/project/cvn-715-aa/bgenfiles/4
          _result-files/ADV_hg19_with_X_imputed_imputed_chr\$i.bgen -og /gs/
          scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/gen/
          protecteurs_chr\$i.gen -incl-snpids /gs/scratch/ggodefroid/Maitrise/".
          $ARGV[0]."/temporary_files/protecteurs.txt\"|qsub
140   done
141
142   for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
143   do
144   echo \"#!/bin/bash
145           #PBS -l nodes=1:ppn=8,walltime=01:00:00
146           #PBS -A cvn-715-aa
147           #PBS -j oe
148           #PBS -V
149           #PBS -N qctool_risq_\$i
150   /sb/project/cvn-715-aa/bin/qctool -g /sb/project/cvn-715-aa/bgenfiles/4
          _result-files/ADV_hg19_with_X_imputed_imputed_chr\$i.bgen -og /gs/
          scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/gen/
          risques_chr\$i.gen -incl-snpids /gs/scratch/ggodefroid/Maitrise/".$ARGV
          [0]."/temporary_files/risques.txt\"|qsub
151   done";
152
153   close OUT;
154
155   system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
          pretraitement_1.sh");
156   system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/pretraitement_1
          .sh");
157
158
159   my@fic=</home/ggodefroid/qctool*>;
160
161   do{
162     @fic=</home/ggodefroid/qctool*>;
163     sleep(5);
164   }while(scalar(@fic)<44);
165 }
166
167 # Obtention des fichiers .ped et .info pour utiliser Haploview
168
169 sub ped_info{
170   open(OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
          obtention_fichiers_ped_info.pl");
```

```
171     print OUT "use strict;
172   use warnings;
173
174   # \$ARGV[0]=risques ou protecteurs
175   # \$ARGV[1]=Nř du chromosomeă
176
177   # Suppression de la première colonne du fichier .gen pour pouvoir être
        traité par gtool
178
179 open (IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
      gen/\".\$ARGV[0].\"_chr\".\$ARGV[1].\".gen\");
180 open (OUT, \">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
      gen/\".\$ARGV[0].\"_\".\$ARGV[1].\".gen\");
181
182 chomp(my\@tab=<IN>);
183
184 foreach(\@tab){
185   chomp(my\@ligne=split / /, \$_);
186   for(my\$i=1; \$i<\@ligne; \$i++){
187     print OUT \"\$ligne[\$i] \";
188   }
189   print OUT \"\n\";
190 }
191
192 close(IN);
193 close(OUT);
194
195   # Création d'un tableau de hashage : clé = nom affymetrix valeur = nom
        officiel
196
197 open(IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/rs_snpid.txt
      \");
198 my\%rs;
199 chomp(my\@rsid=<IN>);
200
201 foreach(\@rsid){
202   chomp(my\@ligne=split/ /,\$_);
203   \$rs{\$ligne[1]}=\$ligne[0];
204 }
205
206 close(IN);
207
208   # Remplacement du nom affymétrix par le nom officiel des SNPs
209
210 open(IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/gen
      /\".\$ARGV[0].\"_\".\$ARGV[1].\".gen\");
211 open(OUT, \">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
      gen/\".\$ARGV[0].\"_chr_\".\$ARGV[1].\".gen\");
212
213 chomp(\@tab=<IN>);
214 foreach(\@tab){
215   chomp(my\@ligne=split / /, \$_);
216   print OUT \"\$rs{\$ligne[0]} \$rs{\$ligne[0]}\";
```

```
217     for(my\$i=2; \$i<\@ligne; \$i++){
218       print OUT \" \$ligne[\$i]\";
219     }
220
221       print OUT \"\n\";
222   }
223
224   close(IN);
225   close(OUT);
226
227     # Lancement de gtool pour obtenir les fichiers .ped et .map</IN>
228   system(\"/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/gtool_v0.7.5
          _x86_64_dynamic/gtool -G --g /gs/scratch/ggodefroid/Maitrise/".$ARGV
          [0]."/temporary_files/gen/\".\$ARGV[0].\"_chr_\".\$ARGV[1].\".gen --s /
          gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/".$ARGV[0].".sample --
          ped /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/\".\
          $ARGV[0].\"_chr\".\$ARGV[1].\".ped --map /gs/scratch/ggodefroid/
          Maitrise/".$ARGV[0]."/temporary_files/\".\$ARGV[0].\"_chr\".\$ARGV
          [1].\".map --snp\");
229
230     # Changement des fichiers .map en fichiers .info (Suppression des
            colonnes 1 et 3)
231   open(IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files
          /\".\$ARGV[0].\"_chr\".\$ARGV[1].\".map\");
232   open(OUT, \">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          haploview/\".\$ARGV[0].\"_chr\".\$ARGV[1].\".info\");
233
234   chomp(\@tab=<IN>);
235
236   foreach(\@tab){
237     chomp(my\@ligne=split/\t/, \$_);
238     print OUT \"\$ligne[1]\t\$ligne[3]\n\";
239   }
240
241     # Changement de la colonne 6 dans le fichier .ped afin de pouvoir le
            faire accepter par Haploview
242   my\%status;
243   open(IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/retinopathy.
          sample\");
244   chomp(my\@sample=<IN>);
245   for(my\$i=2; \$i<\@sample; \$i++){
246     chomp(my\@ligne=split / /,\$sample[\$i]);
247     if(\$ligne[3]==0){
248       \$status{\$ligne[0]}=1;
249     }elsif(\$ligne[3]==1){
250       \$status{\$ligne[0]}=2;
251     }else{
252       \$status{\$ligne[0]}=0;
253     }
254   }
255   close(IN);
256
257   open(IN, \"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files
          /\".\$ARGV[0].\"_chr\".\$ARGV[1].\".ped\");
```

```
258 open(OUT, \">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        temp\".\$ARGV[0].\$ARGV[1].\".ped\");
259
260 chomp(\@tab=<IN>);
261 foreach(\@tab){
262   chomp(my\@ligne=split/\t/,\$_);
263   for(my\$i=0; \$i<\@ligne; \$i++){
264     if(\$i==5){
265       print OUT \"\$status{\$ligne[0]}\t\";
266     }else{
267       print OUT \"\$ligne[\$i]\t\";
268     }
269   }
270   print OUT \"\n\";
271 }
272 rename(\"/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/temp
        \".\$ARGV[0].\$ARGV[1].\".ped\",\"/gs/scratch/ggodefroid/Maitrise/".
        $ARGV[0]."/temporary_files/haploview/\".\$ARGV[0].\"_chr\".\$ARGV
        [1].\".ped\");";
273
274 close(OUT);
275
276 open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/obtention_ped.
        sh");
277
278 print OUT "for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
        19 20 21 22;
279 do
280 echo \"#!/bin/bash
281   #PBS -l nodes=1:ppn=8,walltime=1:00:00
282     #PBS -A cvn-715-aa
283     #PBS -j oe
284     #PBS -V
285     #PBS -N ped_risques_\$i
286 perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
        obtention_fichiers_ped_info.pl risques \$i \"|qsub
287 done
288
289 for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
290 do
291 echo \"#!/bin/bash
292   #PBS -l nodes=1:ppn=8,walltime=1:00:00
293     #PBS -A cvn-715-aa
294     #PBS -j oe
295     #PBS -V
296     #PBS -N ped_protecteurs_\$i
297 perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
        obtention_fichiers_ped_info.pl protecteurs \$i \"|qsub
298 done";
299
300 system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
        obtention_ped.sh");
301
302 system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/obtention_ped.
        sh");
```

```
303   @fic=</home/ggodefroid/ped_*>;
304
305   do{
306     sleep(5);
307     @fic=</home/ggodefroid/ped_*>;
308   }while(scalar(@fic<44));
309
310 }
311
312 # Intervention de Haploview
313
314 sub haplo{
315   open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/haplo.sh");
316
317   print OUT "for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18
          19 20 21 22;
318   do
319   echo \"#!/bin/bash
320     #PBS -l nodes=1:ppn=8,walltime=1:00:00
321       #PBS -A cvn-715-aa
322       #PBS -j oe
323       #PBS -V
324       #PBS -N haplo_risques_\$i
325   java -jar /sb/project/cvn-715-aa/bin/Haploview.jar -nogui -pedfile /gs/
          scratch/ggodefroid/Maitrise/retinopathy/temporary_files/haploview/
          risques_chr\$i.ped -info /gs/scratch/ggodefroid/Maitrise/retinopathy/
          temporary_files/haploview/risques_chr\$i.info -out /gs/scratch/
          ggodefroid/Maitrise/retinopathy/temporary_files/tags/risques_chr\$i -
          pairwiseTagging \"|qsub
326   done
327
328   for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
329   do
330   echo \"#!/bin/bash
331     #PBS -l nodes=1:ppn=8,walltime=1:00:00
332       #PBS -A cvn-715-aa
333       #PBS -j oe
334       #PBS -V
335       #PBS -N haplo_protecteurs_\$i
336   java -jar /sb/project/cvn-715-aa/bin/Haploview.jar -nogui -pedfile /gs/
          scratch/ggodefroid/Maitrise/retinopathy/temporary_files/haploview/
          protecteurs_chr\$i.ped -info /gs/scratch/ggodefroid/Maitrise/
          retinopathy/temporary_files/haploview/protecteurs_chr\$i.info -out /gs/
          scratch/ggodefroid/Maitrise/retinopathy/temporary_files/tags/
          protecteurs_chr\$i -pairwiseTagging\"|qsub
337   done";
338
339   close(OUT);
340
341   system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/haplo.sh")
          ;
342   system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/haplo.sh");
343
344   @fic=</home/ggodefroid/haplo*>;
```

```
345 do{
346    @fic=</home/ggodefroid/haplo*>;
347    sleep(5);
348  }while(scalar(@fic<44));
349
350 }
351
352
353 # Concaténation des fichiers .gen en un fichier . txt pour risques et pour protecteurs
354
355 sub concatenation{
356  my$nom=$_[0];
357  open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $ARGV[0]."_".$nom.".txt");
358  for(my$i=1; $i<10; $i++){
359    open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          gen/".$nom."_chr0".$i.".gen");
360    chomp(my@god=<IN>);
361    foreach(@god){
362     print OUT $_."\n";
363    }
364    close(IN);
365    open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV
366  }
367  for(my$i=10; $i<23; $i++){[0]."/temporary_files/gen/".$nom."_chr".$i.".gen"
        );
368    chomp(my@god=<IN>);
369    foreach(@god){
370     print OUT $_."\n";
371    }
372    close(IN);
373  }
374  close(OUT);
375 }
376
377 # Transformation du fichier issu des fichiers .gen selon le nom officiel du SNP et le mode d'
      expression du SNP
378
379 sub transformation{
380  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/rs_snpid.txt");
381  chomp(my@rs=<IN>);
382  foreach(@rs){
383    chomp(my@ligne=split / /,$_);
384    $rsid{$ligne[1]}=$ligne[0];
385  }
386  close(IN);
387  my$nom=$_[0];
388  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/mode
        .txt");
389  chomp(my@mod=<IN>);
390  foreach(@mod){
391    chomp(my@ligne=split /\t/, $_);
392    if($ligne[1] eq 'd'){
393      $mode{$ligne[0]}=1;
```

```
394     }else{
395       $mode{$ligne[0]}=0;
396     }
397   }
398   close(IN);
399
400
401   print "## Transformation du fichier ".$nom.".\n";
402   open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
          $ARGV[0]."_".$nom.".txt");
403   open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
          $nom."_provisoire.txt");
404   chomp(my@protec=<IN>);
405   foreach(@protec){
406     chomp(my@ligne=split/ /,$_);
407     print OUT "$rsid{$ligne[2]}";
408     for(my$i=6; $i<@ligne; $i++){
409       if(($i)%3==0){
410         if($ligne[$i]==1){
411           print OUT " 0";
412         }
413         elsif($ligne[$i+1]==1){
414           if(exists $mode{$ligne[2]}){
415             print OUT " $mode{$1igne[2]}";
416           }
417         }
418         elsif($ligne[$i+2]==1){
419           print OUT " 1";
420         }
421       }
422     }
423     print OUT "\n";
424   }
425   close(IN);
426   close(OUT);
427 }
428
429 # Rassemblement des tagSNPs dans un même fichier
430
431 sub tag{
432   my$nom=$_[0];
433   my%tag;
434   for(my$i=1; $i<10; $i++){
435     open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
            tags/".$nom."_chr0".$i.".TAGS");
436     chomp(my@tab=<IN>);
437     my$trouve=0;
438     for(my$j=0; $j<@tab; $j++){
439       if($trouve==1){
440         chomp(my@ligne=split /\t/, $tab[$j]);
441         $tag{$ligne[0]}=1;
442       }
443       elsif($tab[$i]=~/Test/){
444         $trouve=1;
```

```
445       }
446     }
447     close(IN);
448   }
449   for(my$i=10; $i<23; $i++){
450     open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          tags/".$nom."_chr".$i.".TAGS");
451     chomp(my@tab=<IN>);
452     my$trouve=0;
453     for(my$j=0; $j<@tab; $j++){
454       if($trouve==1){
455         chomp(my@ligne=split /\t/, $tab[$j]);
456         $tag{$ligne[0]}=1;
457       }
458       elsif($tab[$i]=~/Test/){
459         $trouve=1;
460       }
461     }
462     close(IN);
463   }
464   open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $nom."_provisoire.txt");
465   open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $nom."_provisoire_2.txt");
466   chomp(my@gen=<IN>);
467   foreach(@gen){
468     chomp(my@ligne=split/ /,$_);
469     if(exists $tag{$ligne[0]}){
470       print OUT "$_\n";
471     }
472   }
473 }
474
475
476 # Élimination des SNPs les plus fréquents
477
478 sub elimination{
479   my$nom=$_[0];
480   print "## Elimination des snps les plus fréquents du fichier ".$nom."\n";
481   print "## ".localtime()."\n";
482   open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $nom."_provisoire_2.txt");
483   open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $nom."_d.txt");
484   open(RET, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
        $nom."_retenus.txt");
485
486   chomp(my@prov=<IN>);
487   foreach(@prov){
488     chomp(my@ligne=split / /, $_);
489     my$pop=scalar(@ligne)-1;
490     my$nb=0;
491     for(my$i=1; $i<@ligne; $i++){
492       if($ligne[$i]==1){
493         $nb++;
```

```
494       }
495     }
496     if($nb<=0.9*$pop){
497       print OUT "$_\n";
498       print RET "$ligne[0]\n";
499     }
500   }
501   close(IN);
502   close(OUT);
503 }
504
505 # Séparation des cas et des controles pour les fichiers risques et protecteurs
506
507 sub separation_cas_controles{
508   my$nom=$_[0];
509   my$id=$_[1];
510   print "## Écriture du fichier ".$nom."-".$id.".\n";
511   open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/data/".$id.".txt");
512   open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
513
514   chomp(my@tab=<IN>);
515   chomp(my@dep=split/;/, $tab[0]);
516   print OUT "awk -F \" \" '{ print \$$dep[0] ";
517
518   foreach(my$i=1; $i<@tab; $i++){
519     chomp(my@ligne=split /;/, $tab[$i]);
520     print OUT "\" \" \$$ligne[0] ";
521   }
522   print OUT "}' /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        ".$nom."_d.txt >/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
        temporary_files/".$nom."_".$id."_d.txt\n";
523
524   close(IN);
525   close(OUT);
526
527   system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1")
        ;
528   system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
529 }
530
531 sub transposition_matrice{
532
533   my$nom=$_[0];
534   my$id=$_[1];
535
536   print "## transposition de la matrice ".$nom."_".$id."\n";
537   print "## ".localtime()."\n";
538   open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/transposition.
        sh");
539   print OUT "awk -F \" \" '{\n";
540   print OUT "for (f = 1; f <= NF; f++)\n";
541   print OUT "a[NR, f] = \$f\n";
542   print OUT "}\n";
543   print OUT "NF > nf { nf = NF }\n";
544   print OUT "END {\n";
```

```
545 print OUT "for (f = 1; f <= nf; f++)\n";
546   print OUT "for (r = 1; r <= NR; r++)\n";
547   print OUT "printf a[r, f] (r==NR ? RS : FS)\n";
548 print OUT "}' /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/"
      .$nom."_".$id."_d.txt > /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
      temporary_files/".$nom."_".$id.".txt\n";
549   close (OUT);
550
551   system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
        transposition.sh");
552   system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/transposition.
        sh");
553
554 }
555
556 sub construction_rcf{
557
558   my$nom=$_[0];
559   if($nom eq 'risques'){
560     print "## Construction du fichier risques_cas.rcf\n";
561     print "## ".localtime()."\n";
562     open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          results/risques_cas.rcf");
563     open(IN0, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/cas.txt");
564     open(IN1, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          risques_cas.txt");
565     open(IN2, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          risques_retenus.txt");
566   }else{
567     print "## Construction du fichier protecteurs_controles.rcf\n";
568     print "## ".localtime()."\n";
569     open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          results/protecteurs_controles.rcf");
570     open(IN0, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/data/controles.
          txt");
571     open(IN1, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          protecteurs_controles.txt");
572     open(IN2, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          protecteurs_retenus.txt");
573   }
574   print OUT "#$nom\n";
575   print OUT "[Relational Context]\n\n";
576   print OUT "[Binary Relation]\n\n";
577   chomp(my@individus=<IN0>);
578   for(my$i=1; $i<=@individus; $i++){
579     if($i<@individus){
580       print OUT "$i | ";
581     }else{print OUT "$i\n";}
582   }
583   close(IN0);
584   chomp(my@snp=<IN2>);
585   for(my$i=0; $i<@snp; $i++){
586     if($i<scalar(@snp)-1){
587       print OUT "$snp[$i] | ";
588     }
```

```
589     else{
590       print OUT "$snp[$i]\n";
591     }
592   }
593   chomp(my@data=<IN1>);
594   chomp(my@snps=split / /, $data[0]);
595   for(my$j=0; $j<@data; $j++){
596     chomp(my@ligne=split / /, $data[$j]);
597     for(my$i=0; $i<@ligne; $i++){
598       if($i!=(@ligne-1)){
599         print OUT "$ligne[$i] ";
600       }else{
601         print OUT "$ligne[$i]\n";
602       }
603     }
604   }
605   print OUT "[END Relational Context]";
606   close(IN1);
607   close(OUT);
608 }
609
610 sub fichiers_definitifs{
611   my $nom=$_[0];
612   my$id=$_[1];
613
614   open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
          $nom."_".$id.".txt");
615   open (OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          results/temp.txt");
616
617   chomp(my@tab=<IN>);
618   for(my$i=0; $i<@tab; $i++){
619     chomp(my@ligne=split/ /, $tab[$i]);
620     if ($i==0){
621       for(my$j=0; $j<@ligne; $j++){
622         print OUT "$j ";
623       }
624       print OUT "\n";
625     }else{
626       for(my$j=0; $j<@ligne; $j++){
627         print OUT "$ligne[$j] "
628       }
629       print OUT "\n";
630         }
631   }
632   close(IN);
633   close(OUT);
634   rename("/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
          results/temp.txt","/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
          temporary_files/results/".$nom."_".$id.".txt");
635 }
636
637 # Execution du logiciel pour les risques et les protecteurs à différents supports
638
639 sub coron{
```

```
640 open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/coron.sh");
641 print OUT "for i in 60 65 70 75 80 85;
642   do
643   echo \"#!/bin/bash
644           #PBS -l nodes=2:ppn=16,walltime=15:00:00
645           #PBS -A cvn-715-aa
646           #PBS -j oe
647           #PBS -V
648           #PBS -N coron_protecteurs_\$i_
649   java -Xmx25000m -jar /sb/project/cvn-715-aa/bin/coron-pg-bin.jar /gs/
          scratch/ggodefroid/Maitrise/retinopathy/temporary_files/results/
          protecteurs_controles.rcf \$i% -names -alg:dcharm -of:/gs/scratch/
          ggodefroid/Maitrise/retinopathy/temporary_files/results/
          output_protecteurs_\$i.txt\"|qsub
650   done
651
652   for i in 60 65 70 75 80 85;
653   do
654   echo \"#!/bin/bash
655           #PBS -l nodes=2:ppn=16,walltime=15:00:00
656           #PBS -A cvn-715-aa
657           #PBS -j oe
658           #PBS -V
659           #PBS -N coron_risques_\$i_
660   java -Xmx25000m -jar /sb/project/cvn-715-aa/bin/coron-pg-bin.jar /gs/
          scratch/ggodefroid/Maitrise/retinopathy/temporary_files/results/
          risques_cas.rcf \$i% -names -alg:dcharm -of:/gs/scratch/ggodefroid/
          Maitrise/retinopathy/temporary_files/results/output_risques_\$i.txt\"|
          qsub
661   done";
662
663   system("chmod 777 /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/coron.sh")
          ;
664   system("bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/coron.sh");
665
666   system("rm /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute");
667   system("rm /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute1");
668
669 }
670 system("rm -r /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files")
        ;
671 system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files")
        ;
672 system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        gen");
673 system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        haploview");
674 system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        tags");
675 system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
        results");
676
677 print "\n\n############################################################\n";
678 print "## ".localtime()."\n";
679 print "## Phenotype ".$ARGV[0]."\n";
```

```
680 print "## Determination des cas et des controles\n";
681
682 cas_controles();
683
684 print "## Determination du modele dominant ou recessif\n";
685
686 dominant_recessif('recessif');
687 dominant_recessif('dominant');
688 modele();
689
690 print "## Construction et execution du fichier de commande shell extrayant
        les données pertinentes pour les snps protecteurs et à risque\n";
691
692 data();
693
694 print "## Intervention de qctool\n";
695
696 qctool();
697
698
699 print "## Construction des fichiers ped et info\n";
700
701 ped_info();
702
703 print "## Intervention de Haploview\n";
704
705 haplo();
706
707 print "Concaténation des fichiers .gen en un fichier .txt pour risques et
        pour protecteurs\n";
708
709 concatenation('protecteurs');
710 concatenation('risques');
711
712 print "## Transformation du fichier issu des fichiers .gen selon le nom
        officiel du SNP et le mode d'expression du SNP\n";
713
714 transformation('protecteurs');
715 transformation('risques');
716
717 print "## Rassemblement des tagSNPs dans un même fichier\n";
718
719 tag('risques');
720 tag('protecteurs');
721
722 print "## Elimination des SNPs les plus fréquents\n";
723
724 elimination('risques');
725 elimination('protecteurs');
726
727 print "## Separation des cas et des contrôles\n";
728
729 separation_cas_controles('risques','cas');
730 separation_cas_controles('risques','controles');
731 separation_cas_controles('protecteurs','cas');
732 separation_cas_controles('protecteurs','controles');
```

```
733 print "## Transposition des matrices\n";
734 transposition_matrice('risques','cas');
735 transposition_matrice('risques','controles');
736 transposition_matrice('protecteurs','cas');
737 transposition_matrice('protecteurs','controles');
738
739 print "## Construction des fichiers au format rcf\n";
740 construction_rcf('risques');
741 construction_rcf('protecteurs');
742
743 print "## Construction des fichiers définitifs\n";
744 fichiers_definitifs('risques','controles');
745 fichiers_definitifs('protecteurs','cas');
746 fichiers_definitifs('risques','cas');
747 fichiers_definitifs('protecteurs','controles');
748
749
750 print "# Execution de Coron\n";
751 coron();
752
753 print "## C'est fini\n";
```

Figure B.1 traitement.pl

```
1 echo "#!/bin/bash
2         #PBS -l nodes=1:ppn=8,walltime=12:00:00
3         #PBS -A cvn-715-aa
4         #PBS -j oe
5         #PBS -V
6         #PBS -N traitement_
7 perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/traitement.pl
      retinopathy"|qsub
```

Figure B.2 traitement.sh

```
for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
        #PBS -l nodes=1:ppn=8,walltime=03:00:00
        #PBS -A cvn-715-aa
        #PBS -j oe
        #PBS -V
        #PBS -N qctool_protec_$i
/sb/project/cvn-715-aa/bin/qctool -g /sb/project/cvn-715-aa/bgenfiles/4
    _result-files/ADV_hg19_with_X_imputed_imputed_chr$i.bgen -og /gs/scratch/
    ggodefroid/Maitrise/retinopathy/temporary_files/protecteurs_chr$i.gen -
    incl-snpids /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    protecteurs.txt"|qsub
done

for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
        #PBS -l nodes=1:ppn=8,walltime=03:00:00
        #PBS -A cvn-715-aa
        #PBS -j oe
        #PBS -V
        #PBS -N qctool_risq_$i
/sb/project/cvn-715-aa/bin/qctool -g /sb/project/cvn-715-aa/bgenfiles/4
    _result-files/ADV_hg19_with_X_imputed_imputed_chr$i.bgen -og /gs/scratch/
    ggodefroid/Maitrise/retinopathy/temporary_files/risques_chr$i.gen -incl-
    snpids /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques.txt"|qsub
done
```

Figure B.3 parallele_pretraitement_1.sh

```
use strict;
use warnings;

  # $ARGV[0]=risques ou protecteurs
  # $ARGV[1]=Nº du chromosomeð

  # Suppression de la première colonne du fichier .gen pour pouvoir être traité par gtool

open (IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/gen/"
    .$ARGV[0]."_chr".$ARGV[1].".gen");
open (OUT, ">/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/gen
    /".$ARGV[0]."_".$ARGV[1].".gen");

chomp(my@tab=<IN>);
```

```
13 foreach(@tab){
14   chomp(my@ligne=split / /, $_);
15   for(my$i=1; $i<@ligne; $i++){
16     print OUT "$ligne[$i] ";
17   }
18   print OUT "
19 ";
20 }
21
22 close(IN);
23 close(OUT);
24
25   # Création d'un tableau de hashage : clé = nom affymetrix valeur = nom officiel
26
27 open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/data/rs_snpid.txt");
28 my%rs;
29 chomp(my@rsid=<IN>);
30
31 foreach(@rsid){
32   chomp(my@ligne=split/ /,$_);
33   $rs{$ligne[1]}=$ligne[0];
34 }
35
36 close(IN);
37
38   # Remplacement du nom affymétrix par le nom officiel des SNPs
39
40 open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/gen/".
       $ARGV[0]."_".$ARGV[1].".gen");
41 open(OUT, ">/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/gen/
       ".$ARGV[0]."_chr_".$ARGV[1].".gen");
42
43 chomp(@tab=<IN>);
44   print OUT "$rs{$ligne[0]} $rs{$ligne[0]}";
45   for(my$i=2; $i<@ligne; $i++){
46     print OUT " $ligne[$i]";
47   }
48   print OUT " ";
49 }
50
51 close(IN);
52 close(OUT);
53
54   # Lancement de gtool pour obtenir les fichiers .ped et .map</IN>
55 system("/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/gtool_v0.7.5
       _x86_64_dynamic/gtool -G --g /gs/scratch/ggodefroid/Maitrise/retinopathy/
       temporary_files/gen/".$ARGV[0]."_chr_".$ARGV[1].".gen --s /gs/scratch/
       ggodefroid/Maitrise/retinopathy/data/retinopathy.sample --ped /gs/scratch
       /ggodefroid/Maitrise/retinopathy/temporary_files/".$ARGV[0]."_chr".$ARGV
       [1].".ped --map /gs/scratch/ggodefroid/Maitrise/retinopathy/
       temporary_files/".$ARGV[0]."_chr".$ARGV[1].".map --snp");
```

```
56  # Changement des fichiers .map en fichiers .info (Suppression des colonnes 1 et 3)
57  open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/".
        $ARGV[0]."_chr".$ARGV[1].".map");
58  open(OUT, ">/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
        haploview/".$ARGV[0]."_chr".$ARGV[1].".info");
59
60  chomp(@tab=<IN>);
61
62  foreach(@tab){
63    chomp(my@ligne=split/ /, $_);
64    print OUT "$ligne[1] $ligne[3]
65  ";
66  }
67
68    # Changement de la colonne 6 dans le fichier .ped afin de pouvoir le faire accepter par Haploview
69  my%status;
70  open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/data/retinopathy.
        sample");
71  chomp(my@sample=<IN>);
72  for(my$i=2; $i<@sample; $i++){
73    chomp(my@ligne=split / /,$sample[$i]);
74    if($ligne[3]==0){
75      $status{$ligne[0]}=1;
76    }elsif($ligne[3]==1){
77      $status{$ligne[0]}=2;
78    }else{
79      $status{$ligne[0]}=0;
80    }
81  }
82  close(IN);
83
84  open(IN, "/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/".
        $ARGV[0]."_chr".$ARGV[1].".ped");
85  open(OUT, ">/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/temp
        ".$ARGV[0].$ARGV[1].".ped");
86
87  chomp(@tab=<IN>);
88  foreach(@tab){
89    chomp(my@ligne=split/ /,$_);
90    for(my$i=0; $i<@ligne; $i++){
91      if($i==5){
92        print OUT "$status{$ligne[0]} ";
93      }else{
94        print OUT "$ligne[$i] ";
95      }
96    }
97    print OUT "
98  ";
99  }
100 rename("/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/temp".
        $ARGV[0].$ARGV[1].".ped","/gs/scratch/ggodefroid/Maitrise/retinopathy/
        temporary_files/haploview/".$ARGV[0]."_chr".$ARGV[1].".ped");
```

Figure B.4 obtention_fichiers_ped_info.pl

```
for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
  #PBS -l nodes=1:ppn=8,walltime=1:00:00
     #PBS -A cvn-715-aa
     #PBS -j oe
     #PBS -V
     #PBS -N pretraitement_4_risques_$i
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
    obtention_fichiers_ped_info.pl retinopathy risques $i "|qsub
done

for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
  #PBS -l nodes=1:ppn=8,walltime=1:00:00
     #PBS -A cvn-715-aa
     #PBS -j oe
     #PBS -V
     #PBS -N pretraitement_4_protecteurs_$i
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
    obtention_fichiers_ped_info.pl retinopathy protecteurs $i "|qsub
done
```

Figure B.5 parallele_pretraitement_2.sh

```
for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
  #PBS -l nodes=1:ppn=8,walltime=1:00:00
     #PBS -A cvn-715-aa
     #PBS -j oe
     #PBS -V
     #PBS -N pre_3_risques_$i
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
    obtention_fichiers_ped_info.pl retinopathy risques $i "|qsub
done
```

```
for i in 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22;
do
echo "#!/bin/bash
  #PBS -l nodes=1:ppn=8,walltime=1:00:00
     #PBS -A cvn-715-aa
     #PBS -j oe
     #PBS -V
     #PBS -N haplo_protecteurs_$i
java -jar /sb/project/cvn-715-aa/bin/Haploview.jar -nogui -pedfile /gs/
    scratch/ggodefroid/Maitrise/retinopathy/temporary_files/haploview/
    protecteurs_chr$i.ped -info /gs/scratch/ggodefroid/Maitrise/retinopathy/
    temporary_files/haploview/protecteurs_chr$i.info -out /gs/scratch/
    ggodefroid/Maitrise/retinopathy/temporary_files/tags/protecteurs_chr$i -
    pairwiseTagging"|qsub
done
```

Figure B.6 haplo.sh

```
awk -F " " '{
  for (f = 1; f <= NF; f++)
  a[NR, f] = \$f
  }
  NF > nf { nf = NF }
  END {
  for (f = 1; f <= nf; f++)
  for (r = 1; r <= NR; r++)
  printf a[r, f] (r==NR ? RS : FS)
  }' /gs/scratch/ggodefroid/Maitrise/retinpopathy/temporary_files/
      risques_cas_d.txt > /gs/scratch/ggodefroid/Maitrise/retinopathy/
      temporary_files/risques_cas.txt
```

Figure B.7 transposition.sh

```
1
2  for i in 60 65 70 75 80 85;
3  do
4  echo "#!/bin/bash
5         #PBS -l nodes=2:ppn=16,walltime=15:00:00
6         #PBS -A cvn-715-aa
7         #PBS -j oe
8         #PBS -V
9         #PBS -N coron_protecteurs_$i_
10 java -Xmx200000m -jar /sb/project/cvn-715-aa/bin/coron-pg-bin.jar /gs/
       scratch/ggodefroid/Maitrise/retinopathy/temporary_files/results/
       protecteurs_controles.rcf $i% -names -alg:dcharm -of:/gs/scratch/
       ggodefroid/Maitrise/retinopathy/temporary_files/results/
       output_protecteurs_$i.txt"|qsub
11 done
12
13 for i in 60 65 70 75 80 85;
14 do
15 echo "#!/bin/bash
16        #PBS -l nodes=2:ppn=16,walltime=15:00:00
17        #PBS -A cvn-715-aa
18        #PBS -j oe
19        #PBS -V
20        #PBS -N coron_risques_$i_
21 java -Xmx200000m -jar /sb/project/cvn-715-aa/bin/coron-pg-bin.jar /gs/
       scratch/ggodefroid/Maitrise/retinopathy/temporary_files/results/
       risques_cas.rcf $i% -names -alg:dcharm -of:/gs/scratch/ggodefroid/
       Maitrise/retinopathy/temporary_files/results/output_risques_$i.txt"|qsub
22 done
```

Figure B.8 coron.sh

```
use strict;
use warnings;

# $ARGV[0] = phenotype
# $ARGV[1] = risques ou protecteurs
# $ARGV[2] = support
system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/"
    .$ARGV[1]);
system("mkdir /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/"
    .$ARGV[1]."/output_".$ARGV[2]);

open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
    output_".$ARGV[1]."_".$ARGV[2].".txt");

chomp(my@tab=<IN>);

print "## ÃL'criture des fichiers par nombre de snps\n";
for(my$i=10; $i<(@tab-4); $i++){
  my$substring=substr($tab[$i],1,-1);
  chomp(my@tab1=split /\(/, $tab[$i]);
  chomp(my@nb=split/\)/, $tab1[1]);
  chomp(my@coco=split /}/,$substring);
  chomp(my@snp=split/, /,$coco[0]);
  my$long=scalar(@snp);

  open(OUT,">>/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
      $ARGV[1]."/output_".$ARGV[2]."/motifs_".$long."_snps.txt");
  foreach(@snp){
    print OUT "$_ ";
  }
  print OUT "\t$nb[0]\n";
  close(OUT);
}
close(IN);
```

Figure B.9 transformation_output.pl

```
echo "#!/bin/bash
       #PBS -l nodes=1:ppn=8,walltime=01:00:00
       #PBS -A cvn-715-aa
       #PBS -j oe
       #PBS -V
       #PBS -N post_1_
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/transformation_output.pl
     retinopathy protecteurs 65"|qsub
```

Figure B.10 post_traitement_1.sh

```
use strict;
use warnings;

# $ARGV[0] = phenotype
# $ARGV[1] = risques ou protecteurs
# $ARGV[2] = support
# $ARGV[3] = nombre de snps par motifs

my$nb=0;
my$inc= 0;
my%snp;
  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
      $ARGV[1]."_retenus.txt");
  chomp(my@num=<IN>);
  for(my$i=0; $i<@num; $i++){
    my$nb=$i+1;
    $snp{$num[$i]}=$nb;
  }
  close(IN);

  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
      $ARGV[1]."/results/output_".$ARGV[2]."/motifs_".$ARGV[3]."_snps.txt");
  chomp(my@motif=<IN>);
  foreach(@motif){
    $nb++;
    if($nb%10000==0){
      $inc++;
    }
    open(EXE, ">>/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute".
        $ARGV[3].$inc);
       chomp(my@snps=split/ /,$_);

      print EXE "awk -F \" \" '{ print ";
       for(my$j=0; $j<scalar(@snps)-1; $j++){
        print EXE "\$$snp{$snps[$j]} \" \" ";
       }
    if($ARGV[1]eq 'risques'){
         print EXE " }' /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
             temporary_files/risques/risques_controles.txt >/gs/scratch/
             ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".$ARGV[1]."/
             output_".$ARGV[2]."/temp".$ARGV[3].$inc.".txt\n";
    }elsif($ARGV[1]eq 'protecteurs'){
     print EXE " }' /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
         temporary_files/protecteurs/protecteurs_controles.txt >/gs/scratch/
         ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".$ARGV[1]."/
         output_".$ARGV[2]."/temp".$ARGV[3].$inc.".txt\n";
    }
       print EXE "NB1=`grep -v 0 /gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."
           /temporary_files/".$ARGV[1]."/output_".$ARGV[2]."/temp".$ARGV[3].
           $inc.".txt | wc -l`\n";
      print EXE "echo \$NB1 >>/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/
          temporary_files/".$ARGV[1]."/output_".$ARGV[2]."/tmp/tmp".$ARGV[3].
          $inc.".txt\n";
  close(EXE);
}
```

Figure B.11 itemset_contraste.pl

```
for m in 1 2 3 4 5 6 7;
do
echo "#!/bin/bash
       #PBS -l nodes=1:ppn=8,walltime=02:00:00
       #PBS -A cvn-715-aa
       #PBS -j oe
       #PBS -V
       #PBS -N post_2_$m
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/itemset_contraste.pl
    retinopathy risques 65 $m "|qsub
done
```

Figure B.12 post_traitement_2.sh

```
awk -F " " '{ print $167 " " $168 " " $169 " " $170 " " $171 " " $172 " "
    $709 " " $711 " " $713 " " $714 " " $718 " " $828 " " $829 " " $830 " "
    $831 " " $832 " " $931 " " $932 " " $933 " " $934 " " $935 " " $936 " "
    }' /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/risques/
    risques_controles.txt >/gs/scratch/ggodefroid/Maitrise/retinopathy/
    temporary_files/risques/output_55/temp220.txt
NB1=`grep -v 0 /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/temp220.txt | wc -l`
echo $NB1 >>/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/tmp/tmp22.txt
awk -F " " '{ print $167 " " $168 " " $169 " " $170 " " $171 " " $172 " "
    $637 " " $638 " " $639 " " $640 " " $641 " " $828 " " $829 " " $830 " "
    $831 " " $832 " " $931 " " $932 " " $933 " " $934 " " $935 " " $936 " "
    }' /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/risques/
    risques_controles.txt >/gs/scratch/ggodefroid/Maitrise/retinopathy/
    temporary_files/risques/output_55/temp220.txt
NB1=`grep -v 0 /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/temp220.txt | wc -l`
echo $NB1 >>/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/tmp/tmp22.txt
awk -F " " '{ print $167 " " $168 " " $169 " " $170 " " $171 " " $172 " "
    $637 " " $638 " " $639 " " $640 " " $641 " " $709 " " $711 " " $713 " "
    $714 " " $718 " " $931 " " $932 " " $933 " " $934 " " $935 " " $936 " "
    }' /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/risques/
    risques_controles.txt >/gs/scratch/ggodefroid/Maitrise/retinopathy/
    temporary_files/risques/output_55/temp220.txt
NB1=`grep -v 0 /gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/temp220.txt | wc -l`
echo $NB1 >>/gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
    risques/output_55/tmp/tmp22.txt
```

Figure B.13 extrait d'un fichier execute

```
use strict;
use warnings;

open (OUT, ">/sb/project/cvn-715-aa/Maitrise_Gilles/scripts/
    post_traitement_4.sh");

chomp(my@exe=</sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute*>);

print OUT "for m in ";

foreach(@exe){
  my$nom=substr $_,54;
  print OUT"$nom ";
}

print OUT ";\n";

print OUT "do\n";
print OUT "echo \"#!/bin/bash\n";
print OUT " #PBS -l nodes=1:ppn=8,walltime=2:00:00\n";
print OUT " #PBS -A cvn-715-aa\n";
print OUT " #PBS -j oe\n";
print OUT " #PBS -V\n";
print OUT " #PBS -N L\$m\n";
print OUT "bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute\$m
    \"|qsub\n";
print OUT "sleep 0.2\n";
print OUT "done\n";

close(OUT);
```

Figure B.14 prep_parallel_rech.pl

```
echo "#!/bin/bash
       #PBS -l nodes=1:ppn=8,walltime=01:00:00
       #PBS -A cvn-715-aa
       #PBS -j oe
       #PBS -V
       #PBS -N post_traitement_3
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/prep_parallel_rech.pl"|
    qsub
```

Figure B.15 post_traitement_3.sh

```
for m in 10 100 101 102 103 104 105 110 111 112 120 121 130 140 150 160 170
    180 190 20 21 22 23 24 30 31 310 311 312 313 314 315 316 317 318 319 32
    320 321 322 323 324 325 326 327 328 329 33 330 34 35 36 37 38 39 40 41
    410 411 412 413 414 415 416 417 418 419 42 420 421 422 423 424 425 426
    427 428 429 43 430 431 432 44 45 46 47 48 49 50 51 510 511 512 513 514
    515 516 517 518 519 52 520 521 522 53 54 55 56 57 58 59 60 61 610 611 612
     613 614 615 616 62 63 64 65 66 67 68 69 70 71 710 711 712 713 714 715
    716 72 73 74 75 76 77 78 79 80 81 810 811 812 813 82 83 84 85 86 87 88 89
     90 91 92 93 94 95 96 97 98 ;
do
echo "#!/bin/bash
 #PBS -l nodes=1:ppn=8,walltime=2:00:00
   #PBS -A cvn-715-aa
   #PBS -j oe
   #PBS -V
   #PBS -N post_4_$m
bash /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/execute$m "|qsub
sleep 0.2
done
```

Figure B.16 post_traitement_4.sh

```
use strict;
use warnings;

 # $ARGV[0] : phenotype
 # $ARGV[1] : risques ou protecteurs
 # $ARGV[2] : support
 # $ARGV[3] : Nombre maximal de SNPs

 mkdir("/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".$ARGV
     [1]."/output_60/tmp_def");

 for(my$i=2; $i<=$ARGV[3]; $i++){
  if($ARGV[1] eq 'risques'){
   my@fic=</gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
       risques/output_60/tmp/tmp$i*>;
   open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files
       /".$ARGV[1]."/output_".$ARGV[2]."60/temp_def/tmp".$i.".txt");
   foreach(@fic){
    open(IN, $_);
    chomp(my@tab=<IN>);
    for(my$j=0; $j<@tab; $j++){
     print OUT $tab[$j]."\n";
    }
    close(IN);
   }
   close(OUT);
```

```
    }
    if($ARGV[1] eq 'protecteurs'){
     my@fic=</gs/scratch/ggodefroid/Maitrise/retinopathy/temporary_files/
         protecteurs/output_60/tmp/tmp$i*>;
     open(OUT, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files
         /".$ARGV[1]."/output_".$ARGV[2]."/temp_def/tmp".$i.".txt");
     foreach(@fic){
      open(IN, $_);
      chomp(my@tab=<IN>);
      for(my$j=0; $j<@tab; $j++){
       print OUT $tab[$j]."\n";
      }
      close(IN);
     }
    close(OUT);
    }
  }

  open(OUT98, ">/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/
      ".$ARGV[1]."/chi_carre_dcharm_".$ARGV[2]."_".$ARGV[1]."_98.csv");

print "#############################################\n";
for(my$j=2; $j<=$ARGV[3]; $j++){
  print "## ".localtime()."\n";
  print "## motifs de ".$j." motifs\n";
  open(IN2, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
      $ARGV[1]."/output_".$ARGV[2]."/motifs_".$j."_snps.txt");
  open(IN, "/gs/scratch/ggodefroid/Maitrise/".$ARGV[0]."/temporary_files/".
      $ARGV[1]."/output_".$ARGV[2]."/tmp_def/tmp".$j.".txt");

  chomp(my@tab=<IN>);
  chomp(my@motif=<IN2>);


  for(my$i=0; $i<@tab; $i++){
   my$freq_theo=0;
   chomp(my@ligne=split/\t/, $motif[$i]);
   if($ARGV[1]eq 'risques'){
    $freq_theo=($ligne[1]*2647/766);
   }elsif($ARGV[1]eq 'protecteurs'){
    $freq_theo=($ligne[1]*766/2647);
   }
   my$chi_2=($tab[$i]-$freq_theo)*($tab[$i]-$freq_theo)/$freq_theo;
   if($chi_2>=5.41 && $freq_theo > $tab[$i]){
    print OUT98 "$j\t$motif[$i]\t$tab[$i]\t$freq_theo\t$chi_2\n";
   }
  }
}

close(IN);
close(IN2);
close(OUT98);
```

Figure B.17 chi_2.pl

```
echo "#!/bin/bash
       #PBS -l nodes=1:ppn=8,walltime=01:00:00
       #PBS -A cvn-715-aa
       #PBS -j oe
       #PBS -V
       #PBS -N post_5
perl /sb/project/cvn-715-aa/Maitrise_Gilles/scripts/chi_2.pl retinopathy
    risques 65 7 3.84"|qsub
```

Figure B.18 post_traitement_5.sh

BIBLIOGRAPHIE

Abbas, H. (2012). *EXPANSION DE LA REPRÉSENTATION SUCCINCTE DES GÉNÉRATEURS MINIMAUX*. (Mémoire de maîtrise). Université du Québec à Montréal.

Agrawal, R., Imieliński, T. et Swami, A. (1993). Mining association rules between sets of items in large databases. Dans *ACM SIGMOD Record*, volume 22, 207–216. ACM.

Barbut, M. et Monjardet, B. (1970). Ordre et classification, algèbre et combinatoire,(2 tomes), paris, hachette, 1970. *Zbl0267, 6001*.

Barrett, J. C., Fry, B., Maller, J. et Daly, M. J. (2005). Haploview : analysis and visualization of ld and haplotype maps. *Bioinformatics*, *21*(2), 263–265. `http://dx.doi.org/10.1093/bioinformatics/bth457`. Récupéré de `http://bioinformatics.oxfordjournals.org/content/21/2/263.abstract`

Bay, S. D. et Pazzani, M. J. (1999). Detecting change in categorical data : Mining contrast sets. Dans *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, 302–306., New York, NY, USA. ACM. `http://dx.doi.org/10.1145/312129.312263`. Récupéré de `http://doi.acm.org.proxy.bibliotheques.uqam.ca:2048/10.1145/312129.312263`

Branden, C. et Tooze, J. (1996). *Introduction à la structure des protéines*. De Boeck Supérieur.

Brookes, A. J. (1999). The essence of snps. *Gene*, *234*(2), 177 – 186. `http://dx.doi.org/10.1016/S0378-1119(99)00219-X`

Cantor, R. M., Lange, K. et Sinsheimer, J. S. (2010). Prioritizing gwas results : A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, *86*(1), 6–22.

Caspard, N. et Monjardet, B. (2003). The lattices of closure systems, closure operators, and implicational systems on a finite set : a survey. *Discrete Applied Mathematics*, *127*(2), 241–269.

Chen, M.-S., Han, J. et Yu, P. (1996). Data mining : an overview from a database perspective. *Knowledge and Data Engineering, IEEE Transactions on*, *8*(6), 866–883. `http://dx.doi.org/10.1109/69.553155`

Cirulli, E. T., Kasperavičiūtė, D., Attix, D. K., Need, A. C., Ge, D., Gibson, G. et Goldstein, D. B. (2010). Common genetic variation and performance on standardized cognitive tests. *European Journal of Human Genetics*, *18*(7), 815–820.

Committee, A. M. (2001). Study rationale and design of advance : action in diabetes and vascular diseaseŮpreterax and diamicron mr controlled evaluation. *Diabetologia*, *44*(9), 1118–1120.

Consortium, G. L. G. *et al.* (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, *45*(11), 1274–1283.

Davey, B. A. et Priestley, H. A. (2002). *Introduction to lattices and order*. Cambridge university press.

Edridge, C., Dunkley, A., Bodicoat, D., Rose, T., Gray, L., Davies, M. et Khunti, K. (2015). Hypoglycaemia and diabetes. *Prevalence*, *63*, P410.

Eeles, R. A., Kote-Jarai, Z., Al Olama, A. A., Giles, G. G., Guy, M., Severi, G., Muir, K., Hopper, J. L., Henderson, B. E., Haiman, C. A. *et al.* (2009). Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nature genetics*, *41*(10), 1116–1121.

Fawcett, T. (2006). An introduction to {ROC} analysis. *Pattern Recognition Letters*, *27*(8), 861 – 874. {ROC} Analysis in Pattern Recognition, `http://dx.doi.org/http://dx.doi.org/10.1016/j.patrec.2005.10.010`. Récupéré de `http://www.sciencedirect.com/science/article/pii/S016786550500303X`

Fayyad, U., Piatetsky-Shapiro, G. et Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54.

Frawley, W. J., Piatetsky-Shapiro, G. et Matheus, C. J. (1992). Knowledge discovery in databases : An overview. *AI magazine*, *13*(3), 57.

Gang, F., Majda, H., Wen, W., Haoyu, Y., Michael, S., Timothy R., C., William S., O., Brian, V. N. et Vipin, K. (2012). High-order snp combinations associated with complex diseases : Eefficient discovery, statistical power and functional iinteraction. *Plos One*, *7*(4).

Ganter, B. et Wille, R. (1997). Applied lattice theory : Formal concept analysis. Dans *In General Lattice Theory, G. GrŁtzer editor, BirkhŁuser*. Preprints.

Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J. T., Manolescu, A., Gudbjartsson, D., Agnarsson, B. A., Sigurdsson, A., Benediktsdottir, K. R., Blondal, T. *et al.* (2008). Common sequence variants on 2p15 and xp11. 22 confer susceptibility to prostate cancer. *Nature genetics*, *40*(3), 281–283.

Han, J. et Kamber, M. (2012). *Data Mining : Concepts and Techniques, 3rd ed.* Morgan Kaufmann Publishers.

He, H., Oetting, W. S., Brott, M. J. et Basu, S. (2009). Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC medical genetics*, *10*(1), 1.

Hilderman, R. J. et Peckham, T. (2005). A statistically sound alternative approach to mining contrast sets. Dans *Proceedings of the 4th Australia Data Mining Conference (AusDM-05)*, 157–172. Citeseer.

Hodge, S. E. (1994). What association analysis can and cannot tell us about the genetics of complex disease. *American journal of medical genetics*, *54*(4), 318–323.

Hong, T.-P., Kuo, C.-S. et Chi, S.-C. (1999). A fuzzy data mining algorithm for quantitative values. Dans *Knowledge-Based Intelligent Information Engineering Systems, 1999. Third International Conference*, 480–483. http://dx.doi.org/10.1109/KES.1999.820227

Jobling, M., Hurles, M. et Tyler-Smith, C. (2013). *Human evolutionary genetics : origins, peoples & disease*. Garland Science.

Krarup, N., Borglykke, A., Allin, K., Sandholt, C., Justesen, J., Andersson, E., Grarup, N., Jørgensen, T., Pedersen, O. et Hansen, T. (2015). A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6,041 danish individuals. *Atherosclerosis*.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, *363*(2), 166–176. http://dx.doi.org/10.1056/NEJMra0905980

Marchini, J. et Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, *11*(7), 499–511.

Martin, R. G., Matthaei, J. H., Jones, O. W. et Nirenberg, M. W. (1962). Ribonucleotide composition of the genetic code. *Biochemical and biophysical research communications*, *6*(6), 410–414.

Mendel, G. (1865). Versuche über plflanzenhybriden. Dans *Verhand- lungen des naturforschenden Vereines in Brünn*, 3–47.

Nelson, D. L., Lehninger, A. L. et Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Nirenberg, M. W., Matthaei, J. H., Jones, O. W., Martin, R. G. et Barondes, S. H. (1963). Approximation of genetic code via cell-free protein synthesis directed by template rna. Dans *Fed. Proc*, volume 22, 55–61.

Pasquier, N. (2000). *Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données*. (Thèse de doctorat). Université Blaise Pascal-Clermont-Ferrand II.

Pasquier, N., Bastide, Y., Taouil, R. et Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, *24*(1), 25 – 46. http://dx.doi.org/http://dx.doi.org/10.1016/S0306-4379(99)00003-4. Récupéré de http://www.sciencedirect.com/science/article/pii/

S0306437999000034

Patel, A. (2007). Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the {ADVANCE} trial) : a randomised controlled trial. *The Lancet*, *370*(9590), 829 – 840. http://dx.doi.org/http://dx.doi.org/10.1016/S0140-6736(07)61303-8. Récupéré de http://www.sciencedirect.com/science/article/pii/S0140673607613038

Pearson, T. et Manolio, T. (2008). How to interpret a genome-wide association study. *JAMA*, *299*(11), 1335–1344. http://dx.doi.org/10.1001/jama.299.11.1335. Récupéré de +http://dx.doi.org/10.1001/jama.299.11.1335

Piateski, G. et Frawley, W. (1991). *Knowledge discovery in databases*. MIT press.

Pritchard, J. K. et Przeworski, M. (2001). Linkage disequilibrium in humans : Models and data. *The American Journal of Human Genetics*, *69*(1), 1 – 14. http://dx.doi.org/10.1086/321275

Ramamohanarao, K., Bailey, J. et Fan, H. (2005). Efficient mining of contrast patterns and their applications to classification. Dans *Intelligent Sensing and Information Processing, 2005. ICISIP 2005. Third International Conference on*, 39–47. http://dx.doi.org/10.1109/ICISIP.2005.1619410

Raptis, A. et Viberti, G. (2001). Pathogenesis of diabetic nephropathy. *Experimental and Clinical Endocrinology & Diabetes*, *109*(Suppl 2), S424–S437.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. et Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, *69*(1), 138 – 147. http://dx.doi.org/http://dx.doi.org/10.1086/321276. Récupéré de http://www.sciencedirect.com/science/article/pii/S0002929707614530

Stankiewicz, P. et Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, *61*(1), 437–455. http://dx.doi.org/10.1146/annurev-med-100708-204735

Szathmary, L. (2006). Symbolic data mining methods with the coron platform. *These dŠinformatique, Université Henri Poincaré–Nancy*, *1*.

Tan, P.-N., Steinbach, M., Kumar, V. *et al.* (2006). *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

Teumer, A., Ernst, F. D., Wiechert, A., Uhr, K., Nauck, M., Petersmann, A., Völzke, H., Völker, U. et Homuth, G. (2013). Comparison of genotyping using pooled dna samples (allelotyping) and individual genotyping using the affymetrix genome-wide human snp array 6.0. *BMC genomics*, *14*(1), 506.

Tukey, J. W. (1977). Exploratory data analysis. 39–41.

Valtchev, P., Missaoui, R. et Godin, R. (2004). Formal concept analysis for knowledge discovery and data mining : The new challenges. In P. Eklund (dir.), *Concept Lattices*, volume 2961 de *Lecture Notes in Computer Science* 352–371. Springer Berlin Heidelberg

Valtchev, P., Missaoui, R., Godin, R. et Meridji, M. (2002). Generating frequent itemsets incrementally : two novel approaches based on galois lattice theory. *Journal of Experimental & Theoretical Artificial Intelligence*, *14*(2-3), 115–142.

Valtchev, P., Rouane, M. H., Huchard, M. et Roume, C. (2003). Extracting formal concepts out of relational data. Dans *Proceedings of the 4th Intl. Conference Journées de lŠInformatique Messine (JIMŠ03) : Knowledge Discovery and Discrete Mathematics, Metz (FR)*, 3–6.

Varela, M. A. et Amos, W. (2010). Heterogeneous distribution of snps in the human genome : Microsatellites as predictors of nucleotide diversity and divergence. *Genomics*, *95*(3), 151 – 159. `http://dx.doi.org/10.1016/j.ygeno.2009.12.003`

Vaxillaire, M., Yengo, L., Lobbens, S., Rocheleau, G., Eury, E., Lantieri, O., Marre, M., Balkau, B., Bonnefond, A. et Froguel, P. (2014). Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective desir study. *Diabetologia*, *57*(8), 1601–1610. `http://dx.doi.org/10.1007/s00125-014-3277-x`. Récupéré de `http://dx.doi.org/10.1007/s00125-014-3277-x`

Visscher, P. M., Brown, M. A., McCarthy, M. I. et Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, *90*(1), 7–24.

Watson, J. et Crick, F. (1953a). A structure for deoxyribose nucleic acid. *Nature*, *421*(6921), 397–3988.

Watson, J. D. et Crick, F. H. (1953b). Genetical implications of the structure of deoxyribonucleic acid. *Narture*, *171*, 964–967.

Wheeler, E., Huang, N., Bochukova, E. G., Keogh, J. M., Lindsay, S., Garg, S., Henning, E., Blackburn, H., Loos, R. J., Wareham, N. J. *et al.* (2013). Genome-wide snp and cnv analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nature genetics*, *45*(5), 513–517.

Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers & mathematics with applications*, *23*(6), 493–515.

Yashin, A. I., Wu, D., Arbeev, K. G. et Ukraintseva, S. V. (2010). Joint influence of small-effect genetic variants on human longevity. *Aging (Albany NY)*, *2*(9), 612.

Zaki, M. J. (2000). Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, *12*(3), 372–390.

Zaki, M. J. et Gouda, K. (2003). Fast vertical mining using diffsets. Dans *Proceedings of*

the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 326–335. ACM.

Zaki, M. J. et Hsiao, C.-J. (2002). Charm : An efficient algorithm for closed itemset mining. Dans *SDM*, volume 2, 457–473. SIAM.

Zaki, M. J., Parthasarathy, S., Ogihara, M., Li, W. *et al.* (1997). New algorithms for fast discovery of association rules. Dans *KDD*, volume 97, 283–286.

Zoungas, S., de Galan, B. E., Ninomiya, T., Grobbee, D., Hamet, P., Heller, S., MacMahon, S., Marre, M., Neal, B., Patel, A., Woodward, M., Chalmers, J. et on behalf of the ADVANCE Collaborative Group (2009). Combined effects of routine blood pressure lowering and intensive glucose control on macrovascular and microvascular outcomes in patients with type 2 diabetes : New results from the advance trial. *Diabetes Care*, *32*(11), 2068–2074. `http://dx.doi.org/10.2337/dc09-0959`. Récupéré de `http://care.diabetesjournals.org/content/32/11/2068.abstract`