

ETUDE DE L'INFLUENCE DES POLYMORPHISMES GENETIQUES SUR LE PROFIL D'EXPRESSION DES GÈNES PAR DES MÉTHODES D'EXTRACTION DE MOTIFS DE CO-RÉGULATION

Maxime Rotival¹, Tanja Zeller², Stefan Blankenberg², François Cambien¹, Laurence Tiret¹

¹INSERM - UMRS 937, Génomique Cardiovasculaire
Faculté de médecine, 6^{ème} étage,
91 boulevard de l'hôpital
75634 PARIS CEDEX 13

²Département de Médecine II, Johannes Gutenberg-University, Mainz, Allemagne

Résumé :

En génétique, l'utilisation de données haut débit (mesure simultanée de l'expression de dizaines de milliers de gènes, et de plusieurs centaines de milliers de polymorphismes) pour tester les liens entre patrimoine génétique et les différences phénotypiques à l'échelle cellulaire, conduit à une augmentation drastique du nombre de tests statistiques effectués. Dès lors, le contrôle de l'erreur de type I impose des seuils de significativité extrêmement faibles, se traduisant par une importante perte de puissance. Nous proposons une méthode visant à utiliser les phénomènes de co-régulation de l'expression des gènes pour faciliter la recherche d'effets génotypiques à grande échelle. La méthode décrite ici repose sur l'extraction non supervisée de motifs pertinents par l'analyse en composantes indépendantes et la recherche de motifs d'expression affectés par le génotype. La connaissance des polymorphismes ayant des effets à grande échelle, permet dans un deuxième temps d'affiner la recherche afin d'identifier avec précision les gènes impliqués dans les processus biologiques affectés par le polymorphisme. Une étude par simulations montre que cette approche permet d'augmenter considérablement la puissance des études d'association pour la recherche de polymorphismes affectant l'expression d'un grand nombre de gènes co-régulés. Une application à des données de la Gutenberg Heart Study est proposée.

Abstract :

In contemporary genetics, high throughput technologies allow the assessment of thousands of gene expression levels and genetic polymorphisms simultaneously. However, relating these 2 sources of data to study the influence of DNA variations on gene expression leads to a drastic increase in the number of tested hypotheses. Controlling the type I error implies to adopt very stringent significance levels, leading to an important loss of power. We propose a method aiming at using the existence of co-regulation networks at the gene expression level to make the search for large scale genotypic effect easier. The proposed method relies on the unsupervised extraction of relevant patterns using independent component analysis and the search for expression patterns that are related to the genotype. The detection of polymorphisms having large scale effects then facilitates the identification of specific genes involved in the biological pathway affected by the polymorphism. A simulation study is used to measure the increase of power induced by the proposed procedure for large scale genotypic effects. An application to real data from the Gutenberg Heart Study is proposed.

Mots-clés : Etudes d'association, Genome entier, Corrélations, grande dimension, Analyse en composantes indépendantes.

Motivation :

L'utilisation de technologies haut débit permet d'étudier à grande échelle l'influence de polymorphismes génétiques sur le développement des pathologies et sur leurs phénotypes intermédiaires. Parmi ces phénotypes intermédiaires, le transcriptome (ensemble des niveaux d'expression des gènes d'une cellule) a naturellement un rôle important puisque la transcription est la première étape conduisant à l'expression des gènes dans la cellule et que des modifications à ce niveau sont susceptibles de refléter des variations physiopathologiques à une étape très précoce. Notre travail se situe dans le contexte de l'étude la relation entre polymorphismes génétiques et expression des gènes.

Jusqu'à présent, l'étude de l'impact des polymorphismes sur le transcriptome est restée principalement cantonnée dans la littérature à la recherche de polymorphismes ayant des effets cis-régulateurs [Göring 2007] (affectant directement l'expression du gène dans lequel ou près desquels ils sont situés) ou à des études par gènes candidats [Barboux 2007]. La recherche de gènes trans-régulés (dont l'expression est affectée par un polymorphisme situé en dehors du gène) est beaucoup plus complexe car elle suppose d'avoir une connaissance a priori du système biologique dans lequel le gène intervient et de la cascade d'évènements conduisant à moduler son expression. Le développement des nouvelles technologies dites 'génomique entière' permettant de mesurer chez un grand nombre d'individus tous les gènes exprimés dans une cellule ainsi que tous les polymorphismes de la séquence d'ADN offre de nouvelles perspectives pour identifier les mécanismes de trans-régulation sans hypothèse a priori. Toutefois, la multiplicité des hypothèses testées lorsqu'on croise les 2 sources de données implique d'adopter des seuils de signification drastiques et donc une perte de puissance considérable.

Pour répondre à ce problème, nous proposons une méthode visant à extraire des 'patterns' d'expression et à tester leur association avec les polymorphismes afin de trouver des polymorphismes affectant des systèmes biologiques dans leur ensemble plutôt que des gènes isolés. L'expression des gènes est en effet dépendante d'un réseau complexe de phénomènes de co-régulation, dans lequel de larges ensembles de gènes voient leur transcription activée ou inhibée par des effets communs (par exemple celui de facteurs de transcription) ou par des cascades au sein de systèmes biologiques. Dans ce réseau, des gènes présentant des fonctions liées vont généralement être activés conjointement, favorisant ainsi l'efficacité de la réponse de la cellule aux stimuli extérieurs. On peut donc s'attendre à ce que des polymorphismes affectant le transcriptome à grande échelle au travers d'effets directs ou indirects aient des effets plus facilement détectables si l'on est capable d'identifier ces patterns d'expression.

Données étudiées

La Gutenberg Heart Study (GHS) est une étude allemande en population générale réalisée dans la région de Mainz qui vise à étudier les déterminants précoces de l'athérosclérose et de ses complications, en particulier les déterminants génétiques. Dans ce but, 3306 individus ont été génotypés à l'aide de la puce Affymetrix 6.0 et pour 1554 de ces individus, l'expression des monocytes circulants a été mesurée à l'aide de la puce Illumina Human HT12. Après les diverses opérations de filtrage (contrôle qualité, fréquence, stratification,...), 670 237 polymorphismes et 13154 gènes ont été retenus pour les analyses. Le croisement de ces 2 sources de données conduit donc en théorie à 8.8×10^9 tests statistiques.

Ces données illustrent clairement les problèmes posés par la génomique moderne, où l'utilisation des données de grande dimension devient autant un atout qu'une faiblesse du fait du grand nombre d'hypothèses à tester et des problèmes informatiques inhérents à l'utilisation de données d'une telle dimension.

Méthode proposée

Extraction des patterns de co-régulation par l'analyse en composantes indépendantes (ACI)

L'analyse en composantes indépendantes (ACI), décrite par Hyvärinen (2000) et déjà utilisée dans de nombreux domaines (traitement d'image, finances,...) a été utilisée pour la première fois en génétique par Liebermeister (2003) pour l'analyse de la régulation des gènes chez les levures au cours du cycle cellulaire. Son principe repose sur la recherche d'une décomposition simultanée des différentes variables en une combinaison de facteurs indépendants et non gaussiens, l'ACI se résumant à l'analyse en composantes principales (ACP) dans le cas gaussien. Cette décomposition se fait par la minimisation d'un critère de néguentropie, équivalent à la minimisation de l'information mutuelle entre les différents facteurs latents.

Pour nos analyses, nous utilisons l'algorithme d'ACI, 'fastICA' proposé par Hyvärinen (1999) dont la faible complexité s'adapte particulièrement bien au traitement des données de grande taille. Nous appliquons ici l'ACI à la matrice transposée des données, conformément à ce qu'a proposé Liebermeister, de sorte que les facteurs latents, au lieu de figurer des profils d'expressions d'une condition à l'autre, représentent la participation des gènes au sein d'un système. De ce fait, dans l'espace des variables, les axes de l'ACI ne sont pas nécessairement orthogonaux, et l'ACI aboutit à la formation de « méta-gènes » (combinaisons linéaires de gènes) pouvant être corrélés entre eux et plus aptes à refléter l'activation de systèmes biologiques (Frigyesi 2006). Cette utilisation de l'ACI se justifie également par la forme particulière des données génétiques (nombre de variables beaucoup plus grand que le nombre d'individus) et par le fait qu'en privilégiant les facteurs super-gaussiens elle permet de s'assurer que le système implique un faible nombre de gènes.

Réduction de la dimension et détermination du nombre d'axes.

Préalablement à l'ACI, une ACP est effectuée sur les données afin de réduire la dimension des données et par là, le temps de calcul ainsi que le nombre de tests. Pour déterminer le nombre d'axes retenus pour l'ACI, une ACP est effectuée sur des données permutées où les expressions des différents gènes sont permutées indépendamment les unes des autres (Horn 1965). Afin d'améliorer la méthode proposée par Horn, nous proposons de comparer pour l'axe k la part de variance observée sur les données réelles non pas à celle observée en l'absence complète de structure, mais à celle attendue avec une structure de dimension $k-1$. On s'arrête quand il n'y a plus de gain de variance expliquée par l'ajout d'un axe supplémentaire (figure 1).

Une vérification est faite a posteriori par l'étude de la stabilité des axes de l'ACI (corrélation maximale entre les motifs trouvés d'un appel de l'algorithme à l'autre) en fonction de la dimension retenue (figure 2). On voit nettement sur le graphique que lorsque la dimension des données est surestimée (multipliée par 2 dans l'exemple), seul un sous-ensemble des axes trouvés par l'ACI est stable d'un run à l'autre. Par un critère de points d'inflexion sur la courbe de stabilité on retrouve le nombre d'axes estimé a priori à partir du scree-plot de l'ACP.

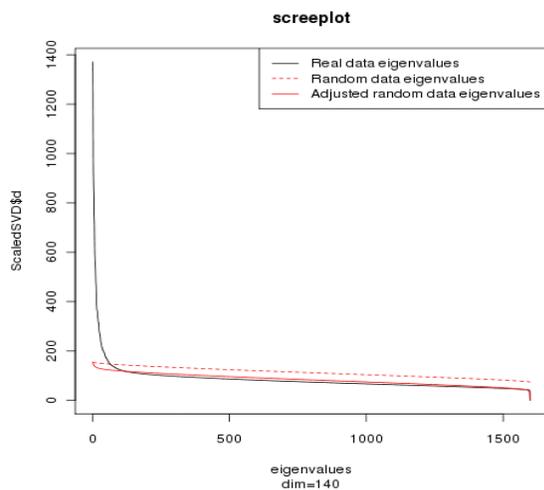


Figure 1 : le graphe montre les valeurs propres obtenues par l'ACP de la matrice d'expression (trait noir plein), ainsi que les valeurs attendues sous l'hypothèse d'absence de structure (trait rouge pointillé) et les valeurs attendues sous l'hypothèse d'absence de structure à partir de l'axe courant.

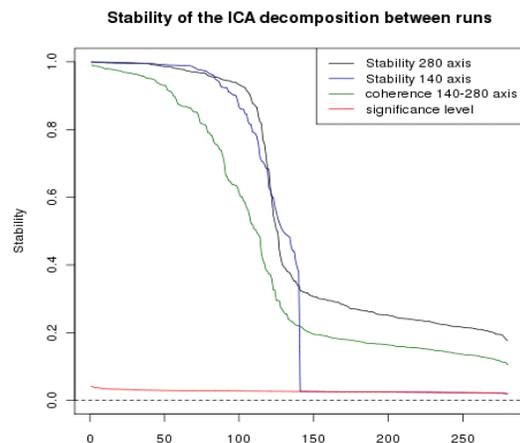


Figure 2 : le graphe 2 montre le degré de stabilité (d'un run à l'autre) des axes de l'ACI, classés par ordre décroissant. La stabilité d'un métagène est estimée par la corrélation maximale entre son profil de contributions des gènes et les profils de contributions des métagènes d'un autre run. On compare la stabilité des métagènes avec une décomposition à 140 axes à celle d'une décomposition avec le double d'axes et à une décomposition sur des données sans structure.

Utilisation des métagènes pour la détermination des gènes affectés par un polymorphisme.

Une fois les métagènes identifiés par l'ACI, leur association avec les polymorphismes est testée par une analyse de variance. A ce niveau, une correction de Bonferroni pour le nombre de polymorphismes et de métagènes testés est effectuée. L'étape suivante consiste à s'intéresser uniquement aux gènes considérés comme contribuant 'significativement' au métagène associé au polymorphisme.

Ainsi que l'ont déjà remarqué plusieurs auteurs, en l'absence de motifs, la position des gènes sur les axes de l'ACI suit approximativement une loi gaussienne, il est donc possible de définir une 'significativité' de la contribution du gène au métagène en comparant sa position sur l'axe (dans l'espace des variables) aux quantiles d'une loi normale. Il est alors possible de sélectionner un sous-ensemble de gènes exprimant 'significativement' le motif d'un métagène.

L'étape suivante est d'identifier dans ce sous-ensemble de gènes, ceux qui sont affectés par le polymorphisme associé au métagène.

Dans ce but nous proposons deux approches qui permettent de contrôler l'erreur de type I :

- Une approche simple : une fois les gènes participant significativement au métagène identifiés, on teste indépendamment l'association de chaque gène retenu avec le polymorphisme par une analyse de variance simple. On retient uniquement les gènes significativement associés après une correction de Bonferroni sur le nombre total de gènes mesurés¹. Cette approche peut sembler toutefois trop conservatrice car elle ne tient pas compte du faible nombre d'hypothèses testées par rapport au nombre total de gènes.
- Une approche 'exacte' qui consiste à estimer par des permutations la distribution sous H0 (pas d'effet du polymorphisme) des coefficients estimés sachant que le gène est trouvé significativement associé au métagène. Une p-value conditionnelle est ainsi calculée par une méthode de smooth bootstrap ou de régression non paramétrique à partir de la distribution empirique des coefficients. L'avantage de cette méthode réside dans le fait que

¹ Et non pas uniquement sur le nombre de gènes du sous-ensemble testé, la sélection des gènes liés au métagène étant susceptible de biaiser l'échantillon de gènes et donc la loi sous H0.

les p-values sont ajustées sur le biais de sélection engendré par la restriction aux gènes situés aux extrémités de l'axe. Une correction de Bonferroni sur le nombre de gènes effectivement testés suffit donc à assurer le contrôle de l'erreur de type I.

Simulations et calculs de puissance

L'efficacité des méthodes proposées a été comparée à la méthode 'naïve' consistant à tester toutes les associations possibles polymorphisme-expression et appliquer une correction simple de Bonferroni. Dans ce but des simulations ont été faites par ré-échantillonnages successifs d'un jeu de données de 3000 gènes sur 300 individus échantillonnés à partir des données réelles de la Gutenberg Heart Study. Sur ces données a été artificiellement ajouté l'effet simulé d'un facteur de transcription affectant un sous-ensemble de 50 gènes en aval et lui-même affecté par un polymorphisme en amont. Le polymorphisme simulé est donc associé à 51 des 3000 gènes simulés. Différents jeux de données ont été simulés en faisant varier la part de variance du facteur de transcription expliquée par le polymorphisme et la corrélation entre celui-ci et les gènes en aval.

Les simulations montrent que les motifs extraits par l'ACI sont plus pertinents que ceux extraits par l'ACP (meilleure correspondance avec les motifs simulés) et offrent une meilleure puissance pour la détection d'un effet des polymorphismes sur le transcriptome. Les simulations montrent également que la corrélation entre les patterns simulés et estimés apparaît pour l'ACP comme pour l'ACI être une fonction croissante du niveau de corrélation entre les gènes et du nombre de gènes exprimant le motif. Malgré cela, il apparaît que même pour des motifs caractérisés par un faible nombre de gènes (25 gènes parmi les 3000 gènes simulés) un gain de puissance significatif peut être atteint par l'utilisation des structures de co-régulation dans la recherche des effets génotypiques. Au niveau de la détection des gènes associés au polymorphisme, les simulations montrent qu'en présence d'un nombre élevé de polymorphismes à tester la puissance est augmentée par la prise en compte de la structure de co-régulation.

Applications aux données réelles

La méthode présentée a été appliquée aux données de GHS. Cent quarante métagènes ont pu être identifiés parmi lesquels 28 étaient significativement enrichis en gènes impliqués dans des voies biologiques connues. Après correction pour le nombre de polymorphismes testés, 8 des métagènes ont été trouvés associés à des polymorphismes, avec parfois plusieurs polymorphismes associés à un même métagène du fait de la redondance des polymorphismes mesurés, et plusieurs métagènes liés au même polymorphisme (effets pléiotropiques). Parmi ces 8 métagènes, deux se sont avérés être créés uniquement par de forts effets en 'cis' (gène régulé par un polymorphisme situé dans la même région chromosomique) sur des clusters de gènes localisés sur les mêmes régions chromosomiques. Ces métagènes présentent donc un intérêt moindre car ils ne révèlent pas de mécanismes de trans-régulation qui sont ceux qui nous intéressent. L'étude des 6 autres métagènes a en revanche confirmé la présence d'effets 'trans' à grande échelle sur le transcriptome et conduit à l'identification de 107 nouveaux gènes potentiellement trans-régulés par les polymorphismes identifiés.

Pour chacun des 5 polymorphismes à l'origine des associations 'trans', le tableau 1 indique le nombre de gènes trans-régulés trouvés avec la méthode naïve (tests exhaustifs sans hypothèse a priori) et en tirant parti des métagènes (avec l'approche dite 'simple').

Tableau 1 : nombre d'associations en 'trans' trouvées en fonction de la méthode utilisée.

méthode 'naïve'	méthode proposée	nombre de gènes communs aux deux méthodes
-----------------	------------------	--

polymorphisme	(tous les tests)	(tests des gènes participant aux métagènes associés uniquement)	
SNP 1	108	181	102
SNP 2	6	12	6
SNP 3	0	2	0
SNP 4	0	3	0
SNP 5	20	37	20

Lecture: on trouve 108 gènes significativement associés au SNP1 par la méthode naïve, et 181 par la méthode proposée. Seuls 6 des 108 gènes ne sont pas retrouvés, mais 75 nouveau gènes sont sélectionnés.

Conclusion

En présence de polymorphismes ayant des effets à grande échelle sur le transcriptome, l'utilisation des structures de co-régulation permet par l'extraction de motifs d'expression de faciliter la recherche de gènes associés tout en contrôlant efficacement le taux de faux positifs générés et en sélectionnant des gènes dont la cohérence biologique est plus importante. En outre, la faible complexité des méthodes proposées les rend tout à fait adaptées à l'évolution contemporaine des analyses 'génomique entier'.

Bibliographie

- [1] Liebermeister, W. (2002) *Linear modes of gene expression determined by independent component analysis*, Bioinformatics, 18(1), 51–60.
- [2] Frigyesi A., Veerla S., David Lindgren D. et Höglund M. (2006) *Independent component analysis reveals new and biologically significant structures in micro array data*, BMC bioinformatics, 7:290.
- [3] Hyvärinen, A. et Oja, E. (2000) *Independent Component Analysis: Algorithms and Applications*, Neural Networks, 13 (4-5), 411–430.
- [4] Horn, J.L. (1965) A rationale and test for the number of factors in factor analysis, Psychometrika ,30 (2), 179–185.
- [5] Hyvärinen, A. et Oja, E. (1999) Survey on *Independent Component Analysis*, Neural Computing Surveys, 2 , 94–128.
- [6] Göring, H. et al. (2007) *Discovery of Expression QTL using large-scale transcription profiling in human lymphocytes*, Nature Genetics, 39 (10), 1208–1216.
- [7] Barboux, S. et al. (2007) *Differential haplotypic expression of the interleukin-18 gene*, European Journal of Human Genetics 15, 856–863.