

Analyse des données génétiques : approches univariées, multivariées et perspectives biomédicales



Christophe LALANNE

13 novembre 2009

Plan de l'exposé

- Type de données et enjeux épidémiologiques
- Approche exploratoire univariée « génome entier »
- Robustesse des résultats
- Approches multivariées
- Perspectives neuroimagerie × génomique

Problématique générale

Les études de « cartographie » des QTL (Quantitative Trait Loci mapping) visent à répondre aux questions suivantes :

- existe-t-il des régions chromosomales exhibant des variants génétiques qui influencent les variations d'un trait héritable ?
- peut-on identifier précisément ces variants génétiques ou polymorphismes ?
- quelle est la contribution de ces variants au niveau des variations du trait dans la population ?

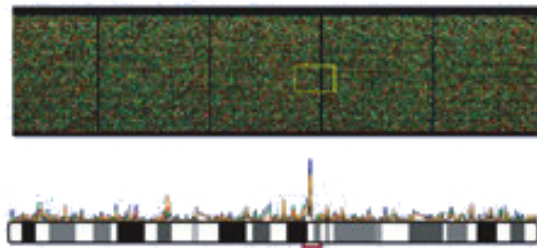
On distingue deux approches bien distinctes : les **études de liaison** (*linkage*) et les **études d'association**, ces dernières étant plus adaptées aux traits complexes (de nature polygénique).

Les études d'association (GWAS)

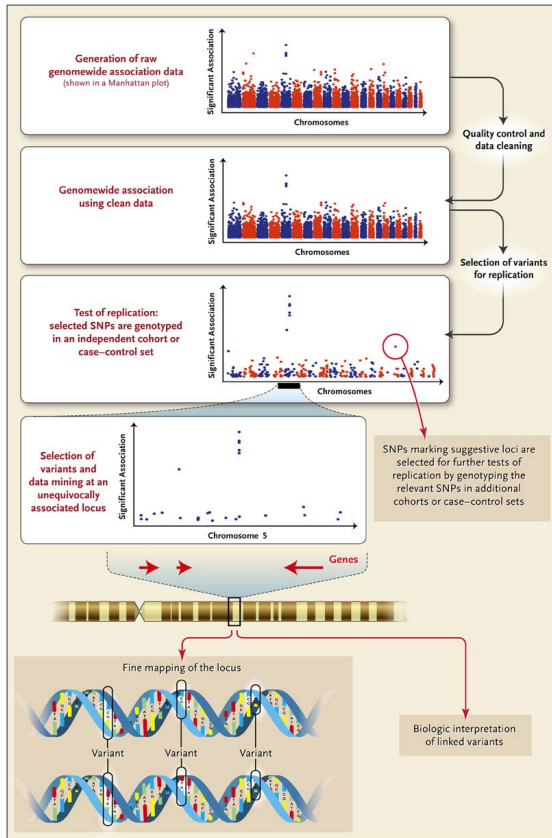
- En dehors des données d'expression (produits de gène), on dispose de plus en plus de données de génotypage, en particulier les **polymorphismes de séquence** (SNP).

AACGAT → AAGGAT

- Ces variations sur la séquence génomique sont observés de manière consistante ($\geq 1\%$ des populations) et sont susceptibles de rendre compte des différences inter-individuelles du point de vue d'un phénotype particulier (trait physique, mesure biologique, maladie).



Une démarche incrémentale



Démarche générale :

1. GWAS sur puce Illumina/Affymetrix (500k à 1000k SNP)
2. Étude de réplication sur échantillon indépendant (cohorte ou CC)
3. Analyse approfondie des locus d'intérêt aux effets « robustes »
4. Étude de liaison (aggrégation familiale)

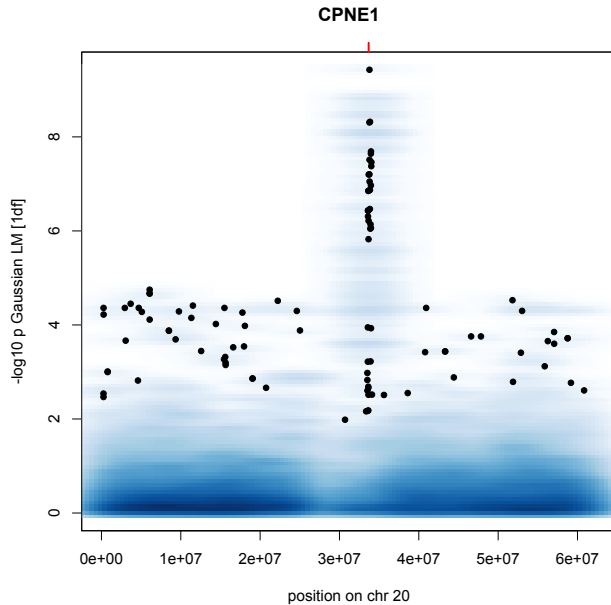
Enjeux statistiques :

- comparaisons multiples, tests corrélés
- puissance, portée généralisante
- fidélité/précision de mesure de la variable comportementale

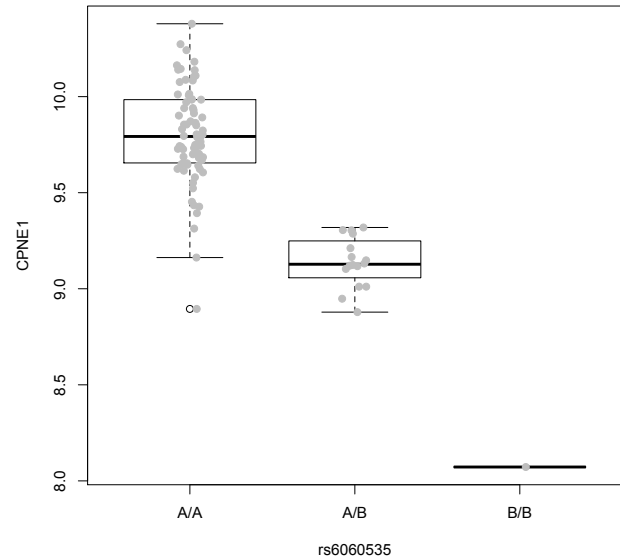
Approches statistiques univariées

- L'objectif est d'**isoler des biomarqueurs** expliquant une part significative de la variabilité observée au niveau d'un phénotype particulier (expression d'un gène, statut/score du sujet, etc.).
- Classiquement, on applique un GLM en traitant la fréquence allélique du variant comme une variable numérique, et en tenant compte des éventuels facteurs de confusion (sexe, âge, centre de génotypage).
- La fréquence allélique dépendant de la population considérée, on stratifie généralement les analyses par rapport au facteur d'ethnicité.

Illustrations (1)

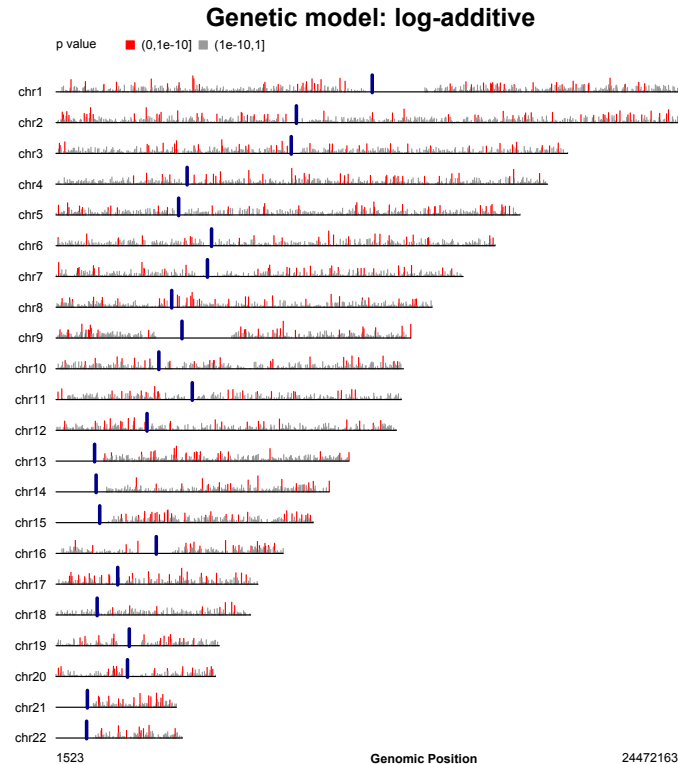


Données d'expression pour le gène
CPNE1 $\times$ données de génotypage
(CEPH) d'individus européens.



Distribution allélique (B = allèle mineur)

Illustrations (2)



Exemple à partir des données [Hapmap](#) (Gonzalez et al., 2007)

Robustesse des résultats

- La multiplicité des tests non-indépendants se traduit par une inflation du risque de conclure à tort à l'existence d'un effet (risque de première espèce α).
- Typiquement, en considérant que l'on travaille sur tous les SNP communs, on choisit un risque unitaire de $5 \cdot 10^{-8}$ (Dudbridge and Gusnanto, 2008).
- Or il s'agit d'une correction très conservatrice (FWER, Bonferroni) qui risque de masquer des effets importants sur le plan fonctionnel bien que peu exprimé du point de la taille de leur effet.
- Alternatives possibles : contrôler le risque de fausses détections (FDR) (Dudoit and van der Laan, 2008).

Perspectives épidémiologiques (1)

TABLE 4. Significant Genomewide Association Study Findings for Nonpsychiatric Disorders^a

Type of Disease or Trait	Unique Findings With $p \leq 5 \times 10^{-8}$	Number of Disorders or Traits
Autoimmune	12	3
Bone density	10	1
Cancer	37	8
Cardiovascular	5	4
Diabetes type 1	10	1
Diabetes type 2	10	1
Gastrointestinal	25	5
Lipid levels	13	3
Neurological	9	6
Physical traits	28	7
Plasma values	22	10
Other	19	10
Total	200	59

^a There is no definitive p value threshold that predicts true positive genomewide association study findings. Interpretation rests on consistency of replication and/or meta-analysis of cumulative data. A p value threshold of 5×10^{-8} has been used throughout this review, based on three estimates that assumed that all common SNPs have been tested (42–44), but other thresholds can be defended. Other approaches include false discovery rate (45) or Bayes Factor (41, 46) criteria. Shown for each category is the number of distinct findings (defined as one or more SNPs in a single chromosomal band for a specific disease or trait) with a p value $\leq 5 \times 10^{-8}$, counting only once those findings reported more than once. Of the 200 findings, 95 had p values $< 10^{-12}$, and 58 had p values $< 10^{-15}$. Some SNPs or regions have produced findings for different disorders or traits (see article). In many cases, there are additional studies or meta-analyses (not included in this tabulation of genomewide association study reports) that contain additional findings or updated significance levels. “Physical traits” includes nondisease traits such as hair and eye color and height. “Plasma values” includes studies of potentially disease-related values (other than lipids) such as C-reactive protein, glucose, and IgE. “Other” includes studies related to eye, skin, or pulmonary diseases; obesity-related traits; aging; and other traits. Data are summarized online (<http://www.genome.gov/GWastudies/> [accessed November 2008]).

Crédit: Psychiatric GWAS Consortium Coordinating Committee, 2009.

Perspectives épidémiologiques (2)

TABLE 6. Summary of Psychiatric GWAS Consortium Genomewide Association Study Samples and Characteristics of Studied Disorders^a

Disorder	Samples	Case Subjects	Comparison Subjects	Trios or Families	Prevalence Rate (%)	Heritability (%)
ADHD	6	1,418	0	2,443	4–12	70–80
Autism	6	652	6,000	4,661	0.3–0.6	90–100
Bipolar disorder	10	7,075	10,559	0	0.3–1.5	73–93
Major depressive disorder	9	12,926	9,618	0	5–18	31–42 ^b
Schizophrenia	11	9,588	13,500	650	0.2–1.1	73–90
Total	42	31,659	26,945	7,772		

^a Data shown are expected combined sample sizes for meta-analysis of genomewide association study data by the Psychiatric GWAS Consortium by the end of 2009. Data are reported for subjects of European-ancestry only; a small number of African American samples are also available for schizophrenia and bipolar disorder. The case subjects are all independent (although independence is tested using genotypes). For each disorder, comparison subjects used in more than one study are counted once; also, comparison subjects used for more than one disorder are counted once in the Total, which is therefore less than the sum of the rows for the disorders. The column for "Trios or Families" includes a sample of multiply-affected families for schizophrenia and trio or sib-pair families (with parents) for ADHD and autism. References for prevalence and heritability are as follows: ADHD (72–74), autism (75, 76), bipolar disorder (77, 78), major depressive disorder (79–82), and schizophrenia (83, 84).

^b For major depression, higher estimates have been obtained in clinical samples (85) or using repeated interviews (81).

Informations complémentaires disponibles sur les sites [GWAS synthesis](#) et [Psychiatric GWAS Consortium](#)

Crédit: Psychiatric GWAS Consortium Coordinating Committee, 2009.

Approches statistiques multivariées

On s'intéresse à la modélisation des liens entre deux blocs de données. Ce type d'études devient de plus en plus fréquent, que ce soit en génomique (Parkhomenko et al., 2007), transcriptomique (Lê Cao et al., 2009), ou dans le cadre génétique×neuroimagerie (Haroon et al., 2009).

Questions afférentes :

1. Comment sélectionner les prédicteurs d'intérêt ?
2. Quel modèle multivarié choisir ? (cas $n \ll p + q$)
3. Comment évaluer ses performances ? (généralisation)
4. Comment comparer les résultats de différents modèles ?

État de l'art sur la question multivariée

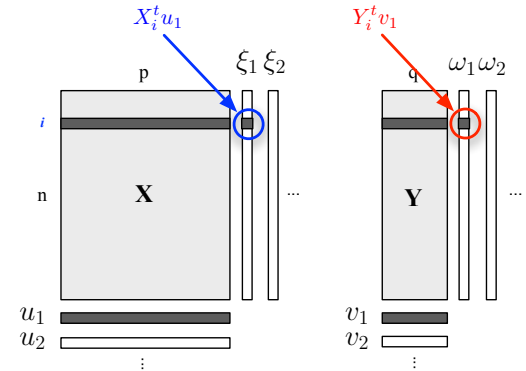
- Les méthodes multivariées comme la PLS, la CCA ou l'analyse de co-inertie ont permis de mettre en évidence des liens intéressants entre des données de transcriptomique, de génomique et de neuroimagerie, en considérant un nombre limité de variables latentes expliquant la variance de leur propre bloc ainsi qu'une part significative du lien inter-blocs.
- Toutefois, il est nécessaire d'introduire des techniques de **régularisation** de type L_1 (« lasso »), L_2 (« ridge ») ou L_1L_2 (« elasticnet ») pour pallier la grande dimensionnalité des données et aboutir à des résultats interprétables (Parkhomenko et al., 2007, 2009, Witten and Tibshirani, 2009, Lê Cao et al., 2008 and González et al., 2009).
- D'autre part, les problèmes de **sur-ajustement** et de **généralisabilité** doivent être pris en considération, en utilisant des méthodes de validation croisée (Guyon et al., 2006).

Un exemple : la régression PLS

1. Modèle de régression reposant sur l'idée de construire des combinaisons linéaires (orthogonales) des variables d'intérêt de chaque bloc tout en maximisant leur covariance :

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

2. En d'autres termes, on cherche des **variables latentes** qui rendent compte au maximum de l'information (linéairement) contenue dans le bloc X tout en permettant de prédire le bloc Y avec une erreur minimale.
3. Comme $n \ll p + q$, on peut imposer une **pénalisation sur le poids** des variables et ne retenir que celles qui ont la contribution la plus importante.



$$\hat{\beta}_{\text{LASSO}} = \arg \min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_1, \quad \text{avec } \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

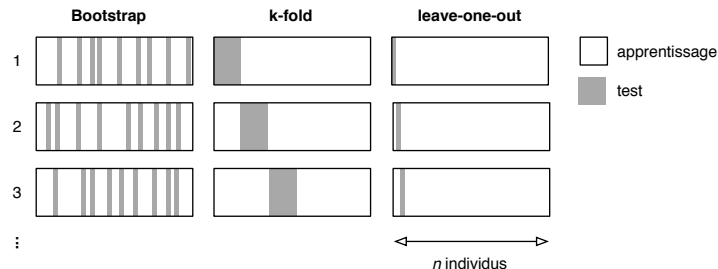
- La corrélation entre les scores factoriels sur la première dimension (ξ_1 et ω_1) reflète l'ampleur du lien X - Y .
- D'autres méthodes multivariées peuvent être considérées, e.g. sparse ou ridge CCA.

Méthodologie statistique dans le cas $n \ll p + q$

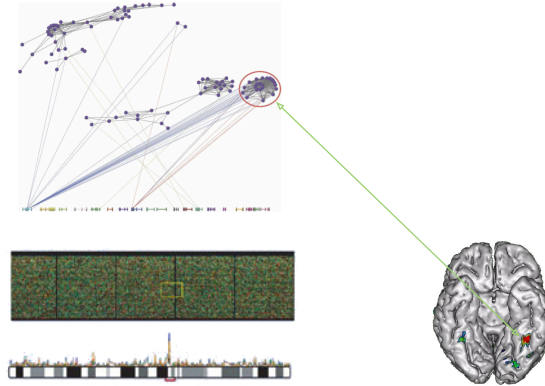
Le grand nombre de paramètres à estimer entraîne un **risque de sur-ajustement** du modèle aux données observées. En conséquence, il est indispensable d'évaluer dans quelle mesure les prédictions du modèle choisi sont généralisables à de nouvelles données (pouvoir de généralisation), e.g. Molinaro et al. (2005).

Principe de la validation croisée :

1. estimer le modèle sur les individus d'apprentissage,
2. tester le modèle sur le ou les individu(s) de test.



Génétique et neuroimagerie



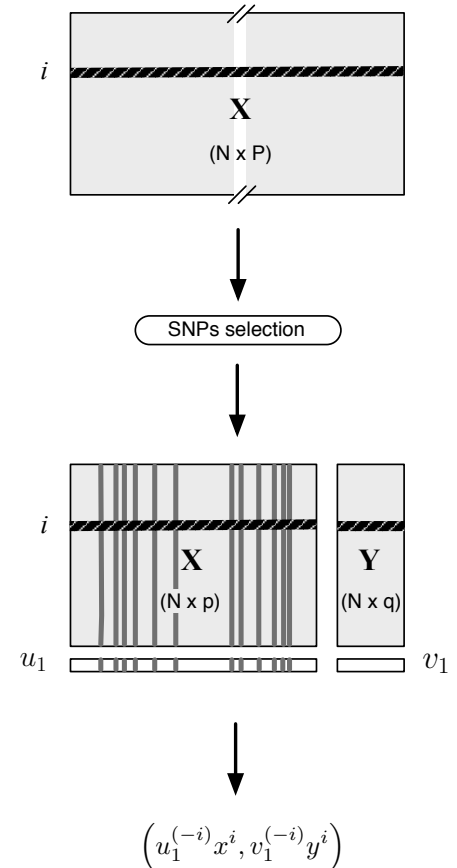
- On s'intéresse à la modélisation des liens entre des données de génotypage (haute densité) et des données de neuroimagerie (IRMf) avec l'objectif de mettre en évidence des réseaux de covariation entre les gènes et les activations cérébrales.
- À terme, cela peut permettre de prédire un phénotype clinique ou comportemental particulier à partir de ces biomarqueurs.

Crédit: Williams et al., 2007.

Méthodologie proposée

Utilisation d'une procédure 10-fold CV, enchâssée dans un schéma de ré-échantillonnage (1000 permutations),

1. analyse génome entier pour tester chaque SNP vs. un ensemble de ROI distribuées dans le cerveau (test F) ;
2. sélectionner les SNP classés selon leur p -valeur (non ajustée);
3. appliquer le modèle PLS avec des paramètres de régularisation variables sur l'échantillon d'apprentissage ;
4. calculer la statistique de test (produit vectoriel des scores factoriels, $\sim$ corrélation canonique en CCA) sur l'échantillon de test ;
5. estimer la p -valeur empirique associée à cette statistique de test sous l'hypothèse nulle.



Conclusion

- Les données de grande dimension posent des problèmes sur le plan computationnel mais également du point de l'interprétation des résultats.
- Dans une approche confirmatoire, la question de la robustesse des résultats devient cruciale, en particulier lorsque l'on étudie des traits discrets rares (faible prévalence).
- Il est donc nécessaire de développer des méthodes de validation croisée pour optimiser les paramètres des modèles, et de réaliser des méta-analyses pour confirmer les résultats acquis (Ioannidis et al., 2009).

Références

- Gonzalez, J. R., Armengol, L., Sole, X., Guino, E. and Mercader, J. M. et al. (2007). SNPAssoc: an R package to perform whole genome association studies. *Bioinformatics*, 23(5):654-655.
- Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32:227-234.
- Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genomics*. Springer.
- Psychiatric GWAS Consortium Coordinating Committee (2009). Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, 166(5):540-556.
- Parkhomenko, E., Tritchler, D. and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings*, 1(1):S119.
- Lê Cao, K.-A., Martin, P., Robert-Granié, C. and P, B. (2009). A sparse pls for variable selection when integrating omics data. *BMC Bioinformatics*, 10(34).
- Hardoon, D. R., Ettinger, U., Mourao-Miranda, J., Antonova, E. and Collier, D. et al. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neuroscience Letters*, 450:281-286.
- Parkhomenko, E., Tritchler, D. and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

- Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1).
- Lê Cao, K.-A., Rossouw, D., Robert-Granié, C. and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1).
- González, I., Déjean, S., Martin, P. G. P., Gonçalves, O. and Besse, P. et al. (2009). Highlighting relationships through regularized canonical correlations analysis: Applications to high throughput biology data. *Journal of Biological Systems*, 17(2):173-199.
- Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A., editors (2006). *Feature Extraction: Foundations And Applications* Springer-Verlag.
- Williams, R. B. H., Chan, E. K. F. and Cowley, M. J. et al. (2007). The influence of genetic variation on gene expression. *Genome Research*, 17.
- Ioannidis, J. P. A., Thomas, G. and Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, 10:318-329.