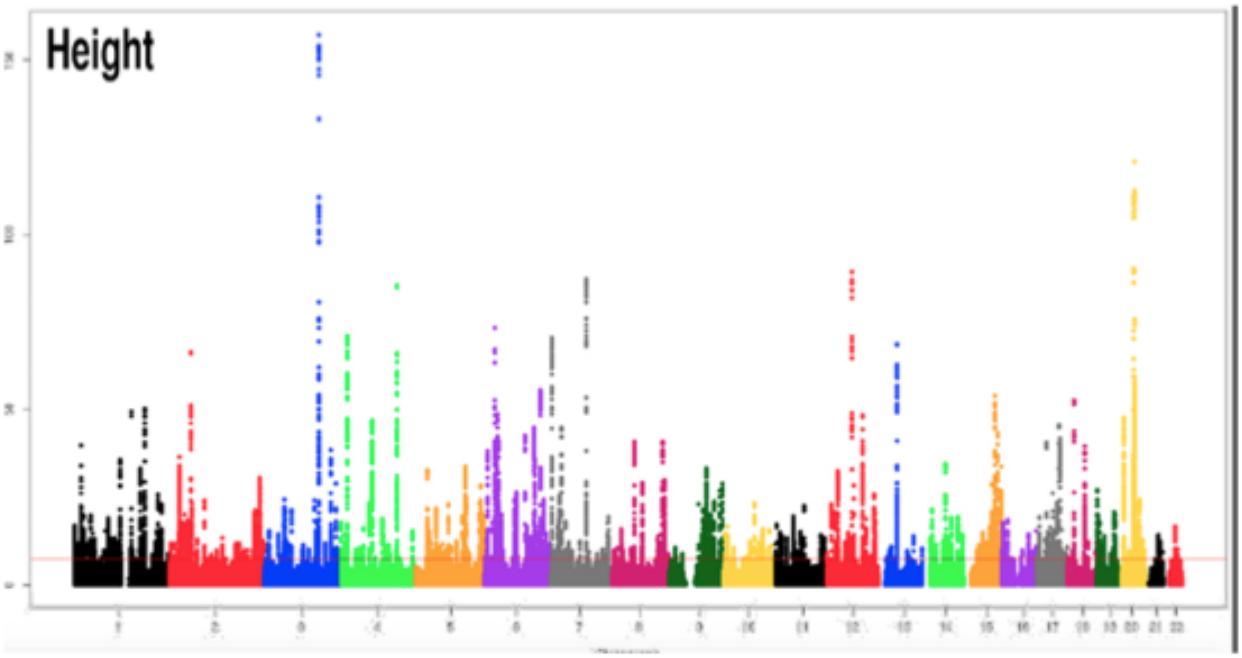
Prioritizing genes via network enrichment

Introduction

Notre projet consistait à analyser des réseaux de gène grâce à la méthode de prioritisation. Pour ce faire nous avons utilisé différentes méthodes d'enrichissement qui seront détaillées dans les paragraphes suivants.
Le but final était d'associer certains gènes à un phénotype. Cette analyse complexe nécessite quelques explications sur les concepts clés utilisés lors de notre travail.

La première définition importante à comprendre est le SNP. Ce sont des polymorphisme de l'ADN dans le quel deux chromosomes diffèrent par une seul paire de base. Ils sont responsables de la variation génétique dans le génome humain et peuvent créer une différence de sensibilité à certaines maladies.
De plus, toute notre analyse repose sur des données récoltées à partir d'un GWAS. Cette méthode consiste à prélever de l'ADN chez des patients malades et des patients sains pour les comparer et voir s'il existe des SNPs associés à des maladies. Pour chaque SNP, on regarde si la fréquence allélique est significativement différente entre les patients contrôles et les patients sains. Ensuite, on calcule une p-value qui est en lien avec le degré de variation des SNPs.

L'utilité de cette démarche est de pouvoir faire un Manhattan plot, un graphique représentant les SNPs (1 points = 1 SNP) des gène sur les différents chromosomes en fonction de leur p-value. Ces dernières sont représentées en échelle logarithmique afin de faciliter la compréhension du graphique. On comprend donc que les SNPs les plus hauts sont hautement significatifs.

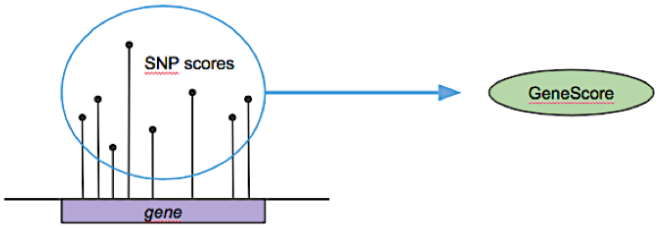


Picture source : David Lamparter

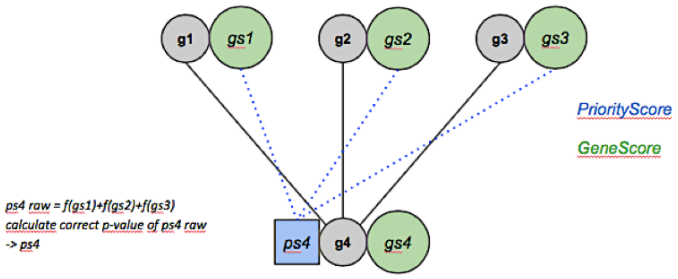
Lors de notre analyse nous ne travaillerons pas avec des SNP score mais avec des GeneScores et des PriorityScores.

Méthode

Nous utilisons les p-values associées aux SNPs obtenues par les études GWAS, que l'on va réduire pour chaque gène auquel les SNPs sont correspondants. Il s'agit d'une p- value que l'on appelle le **GeneScore**. Ce score est obtenu par une moyenne ajustée des SNPScores en tenant compte du déséquilibre de liaison, responsable de la dépendance des p-values quand les SNPs se trouvent proches sur le génome.



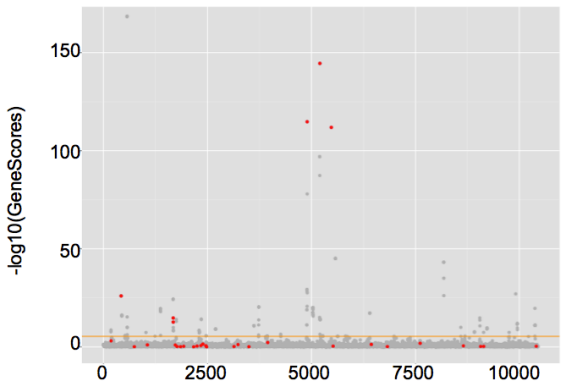
Ainsi, on peut mettre en évidence les gènes importants, ceux qui auront une p-value hautement significative. On remarquera aisément que ces gènes sont en nombre relativement élevé. Le but de notre projet est à ce point de trouver une méthode qui permette de sélectionner quel gène est réellement intéressant et important, et pour ce faire on introduit un second paramètre de sélection, qui sera les interactions physiques entre les produits des gènes. On utilisera pour cela des réseaux de gènes.
Ces réseaux permettent de calculer un nouveau score (p-value), le **PriorityScore**. Ce score est dépendant du réseau utilisé, car les interactions peuvent varier selon le mode de construction du réseau.



Le calcul de ce nouveau score est basé sur une moyenne ajustée des GeneScores des gènes en interaction directe avec le gène pour lequel on essaie de définir ce PriorityScore. Ce score est donc ajusté en fonction des éventuelles dépendances entre deux gènes, ce que l'on observe notamment lorsqu'il y a des gènes appartenant au même cluster. Ainsi, on restreint les gènes d'intérêt par cette deuxième valeur significative.

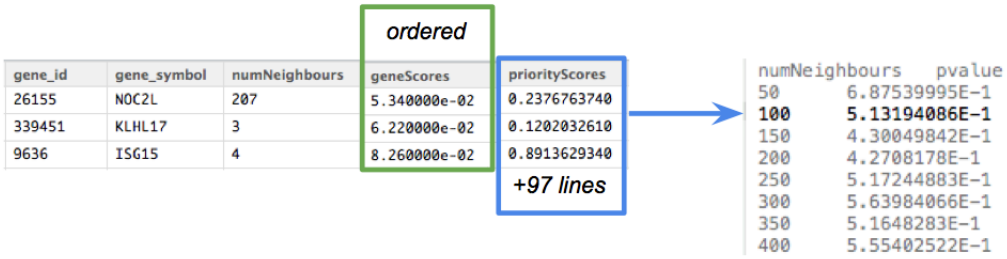
On peut ensuite intégrer ces deux scores dans un seul graphique: le Manhattan plot. Plus un gène a un genescore élevé, plus le point qui le représente sera haut dans le graphique. De plus, lorsque que le prioritysore dépasse un certain seuil significatif, le point est marqué en rouge.

Ceci nous permet donc de visualiser facilement quels gènes sont intéressants, autant du point de vue de leurs genescores que de leurs prioritysores.



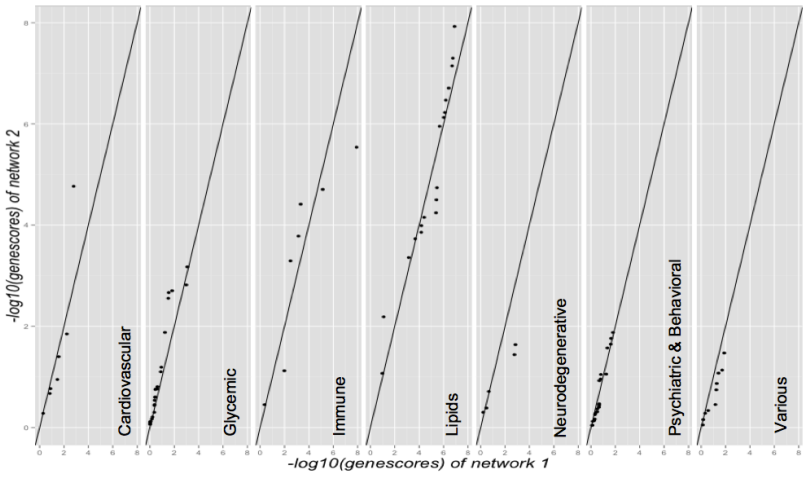
Dans nos premières données GWAS nous avons 130 différents phénotypes pour lesquels nous pouvons créer un Manhattan plot. Certains ayant de meilleurs résultats que d'autres, notre but serait dorénavant de comparer différentes méthodes de prioritisation. Comme il serait impossible de comparer à l'oeil les 130 Manhattan plots de correspondant à l'utilisation d'un 1er réseau, avec les 130 d'un deuxième réseau, et les 130 d'un troisième etc… Il nous faut trouver un moyen de comparer directement les réseaux entre eux.
On va donc faire des analyses d'enrichissement, c'est-à-dire qu'on va calculer des valeurs CUT pour chaque phénotype, pour que l'on puisse attribuer une seule valeur par Manhattan plot.

Cette valeur CUT, c'est simplement un p-value de nos Priority Scores. Pour la calculer, on ordonne l'ensemble de nos gènes en fonction de leurs genescores (du plus significatif au moins significatif). Nous prenons ensuite les X premiers priorityscores associés (dans notre cas les 100 premiers) et nous les comparons à une distribution aléatoire pour obtenir une pvalue. Cette valeur nous permet donc de résumer un Manhattan plot.



Une fois cette valeur calculée, il est relativement aisé de comparer deux réseaux entre eux. En effet, il s'agit simplement de comparer les pvalues obtenues avec différents priorityscores (donc différents réseaux) pour savoir lequel serait le plus efficace.

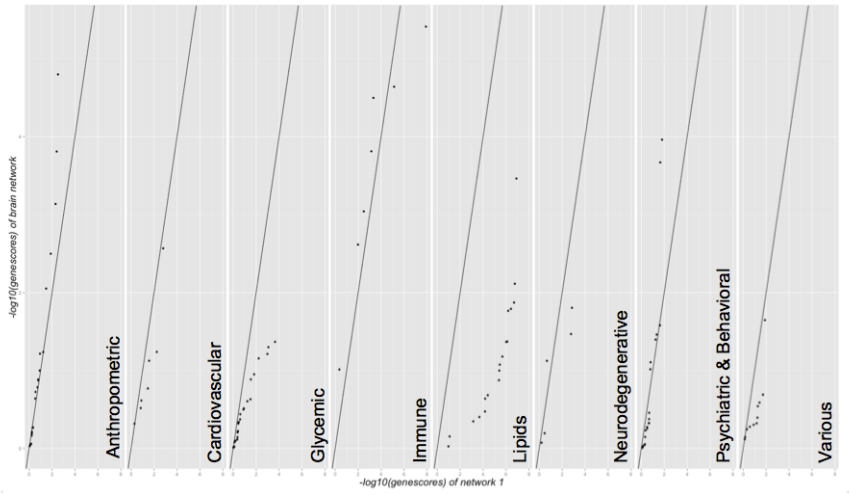
On obtient donc un graphique pour chaque groupe de traits phénotypiques où l'on plot sur l'axe des x les pvalues d'un premier réseau et sur l'axe des y celles d'un deuxième réseau. Un pattern de déviation de la ligne médiane vers l'un des axes représente donc une plus grande efficacité de celui-ci. Dans l'exemple ci-contre, on voit que le network 1 (protein-protein interaction) et le network 2 (diffusion tiré du network 1) sont relativement semblables, aucun pattern particulier ne se détache.



<u>Exemple d'application de notre méthode</u>

On s'est intéressés à un nouveau réseau tiré d'un article paru fin avril 2015 [1].

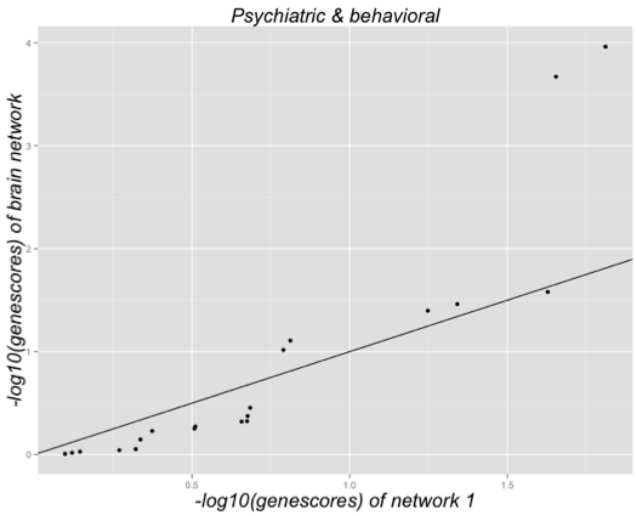
L'avantage d'un tissue-specific network, c'est qu'il permet de prédire des réponses spécifiques, en identifiant le rôle fonctionnel des gènes pour chaque type de tissus. On a donc utilisé un tissus specific network sur le cerveau dont ont été tirés des priority scores que nous avons employé suivant la méthode préalablement présentée.



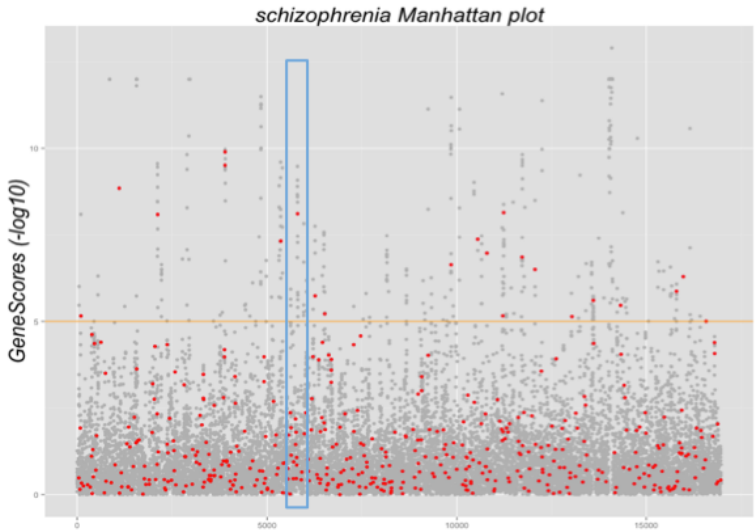
On peut donc comparer le network 1 (protein-protein interaction) sur l'axe des abscisses et le network brain sur l'axe des ordonnées.
On constate que le network 1 semble de manière générale meilleur, mais que les groupes « anthropometric » et « psychiatric & behavioral » ont de meilleurs résultats avec le network brain, ce qui semble cohérent.

On a décidé de s'intéresser plus particulièrement au groupe « psychiatric & behavioral » et de regarder en détail quels sont les 2 points qui semblent particulièrement bien s'en sortir dans le cas du network brain. Pour rappel, ces points présentent une distance intéressante avec la ligne médiane représentant l'endroit où les valeurs des pvalues sont les mêmes pour chacun des 2 réseaux.

Ces points sont représentatifs de 2 traits phénotypiques:
- le fait d'avoir fait des études universitaires
- la schizophrénie



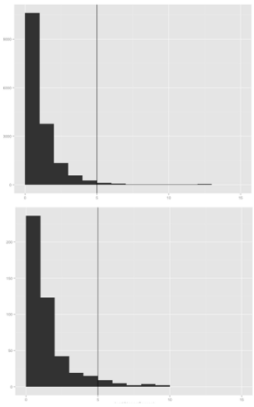
On a choisi de s'intéresser plus particulièrement à la schizophrénie qui, à notre avis, est le plus susceptible des deux d'avoir un background génétique.



On reprend donc les données particulières à la schizophrénie pour en tirer un Manhattan plot.

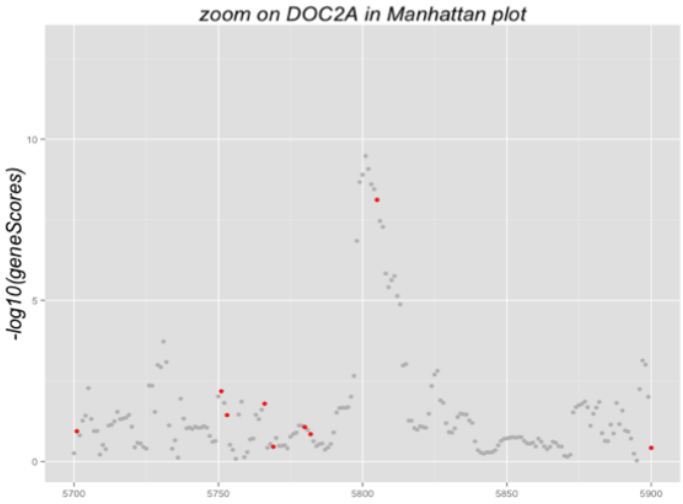
En regardant ce Manhattan plot, on voit qu'on a un nombre relativement important de points rouges en-dessous et en-dessus de la ligne représentant notre seuil significatif pour les genescores.

On souhaite vérifier rapidement à quel point ces données sont intéressantes. On aimerait savoir, parmi les points rouges que l'on observe, combien y en a-t-il de plus que ce qu'on attendrait dans le cas d'une distribution aléatoire. Pour avoir une idée de cet ordre de grandeur, on a fait deux histogrammes.



Celui du haut représente la distribution de l'ensemble des points du manhattan plot.
Celui du bas représente la distribution des points rouges.
La ligne verticale est au même niveau dans les 2 cas, qui est le seuil significatif représenté par la ligne horizontale du Manhattan plot.

Ceci nous a permis de calculer des ratios de distributions de part et d'autres de la droite pour chacun des deux histogrammes, que nous avons ensuite comparé entre eux pour obtenir un enrichissement d'un facteur de 2. C'est un facteur suffisamment élevé pour qu'on puisse faire confiance à ce que l'on voit et s'intéresser de plus près aux gènes qui ressortent (haut sur le Manhattan plot car genescore élevé et rouges car priorityscore élevé également).



Un bon exemple de l'intérêt de notre projet est de zoomer sur l'intervalle montré ci-contre.

On voit un ensemble de points gris placés haut sur le Manhattan plot, représentant donc des gènes avec un bon genescore.
Comment choisir lequel serait plus intéressant que les autres? Comment prioritiser?
On a pu le faire grâce à l'utilisation du network spécifique au cerveau par l'intermédiaire des priorityscore. On voit un gène qui se détache particulièrement: c'est le point rouge.

Après quelques recherches sur le gène correspondant au point rouge, DOC2A, on trouve qu'effectivement, il est exprimé dans le cerveau et joue un rôle dans la transmission des neurotransmetteurs au niveau synaptique. Il semble également jouer un rôle dans le cas de la schizophrénie et de l'autisme [2].

On a donc pu, grâce à une méthode statistique, mettre en évidence le rôle d'un gène dans un phénotype particulier, rôle corrélé par la recherche clinique dans notre cas.

Références

[1] *Olga G Troyanskaya et al., " Understanding multicellular function and disease with human tissue-specific networks. ", Nature Genetics, 47, 569–576 (2015)*

[2] *Glessner et al., "Strong synaptic transmission impact by copy number variations in schizophrenia", PNAS, (June 2010), vol. 107, no. 23*