Clustering of subjects on subsets of SNPs

JJ Meulman, EJC de Geus, G Willemsen, JJ Hottenga,
PF Sullivan, JH Smit, BWJH Penninx, DI Boomsma

Background

As part of the Genetic Association Information Network we conducted a genome wide association study of 1,738 cases with major depressive disorder (MDD) and 1,802 controls selected to be at low liability for MDD from The Netherlands (Boomsma et al. 2008; Sullivan et al. in press) .

Genotyping & Quality Control

- Perlegen 600k SNP chip.
- Software PLINK.
- Genome build 36.
- Duplicate and Mendelian errors < 2 per SNP.
- MAF > 0.01.
- Missing genotypes < 0.05 in SNPs and individuals.
- HWE > 0.00001.
- Genomic control inflation factor = 1.0.
- 435,291 SNPs

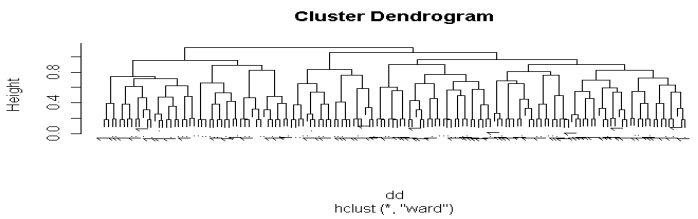
Results

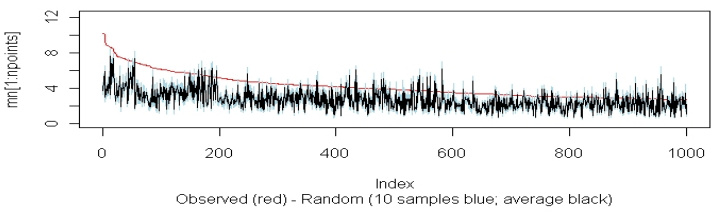
Traditional GWA analysis showed that 11 of the top 200 signals localized to a 167 kb region overlapping the gene *PCLO*, with p-values of 7.7x10-7 for rs2715148 and 1.2x10-6 for rs2522833.

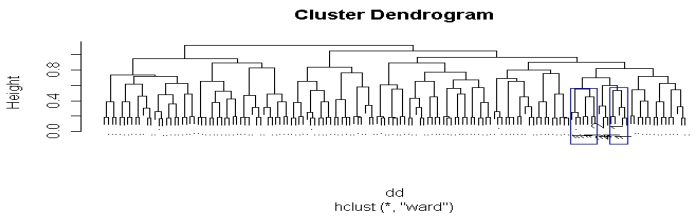
Next aim

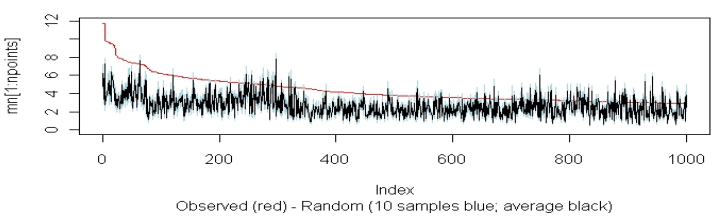
In a next step, we started to explore the co-clustering of subjects on subsets of SNPs. We apply a very general class of techniques, developed by Friedman and Meulman (2004), for the clustering of objects on subsets of attributes, called "COSA" (**C**lustering **O**bjects on **S**ubsets of **A**ttributes).

The COSA algorithm is aimed to detect subgroups of subjects that preferentially cluster on *subsets* of the attribute variables (SNPs). Common data analysis approaches in systems biology are to cluster the attributes first, and only after having reduced the original many-attribute data set to a much smaller one, one tries to cluster the objects. The problem here is that we would like to select those SNPs that discriminate most among subjects (so we do this while regarding all attributes multivariately). Therefore, two tasks have to be carried out simultaneously: cluster the subjects into homogeneous groups, while selecting different subsets of SNPs (one for each cluster). The SNP subset for any discovered cluster may be completely, partially or non-overlapping with those for other clusters.
In a preliminary COSA analysis (with a random sample of 150 subjects on 13,122 SNPs within genes on the whole genome (one SNP per gene)), two small clusters of MDD cases were detected, displayed at the two panels at the left. The COSA weights for the SNPs associated with these two clusters are displayed in the two panels at the right. In the high regions, the weights for the 2 clusters are significantly higher than the weights for 10 random samples of size N/2.









Acknowledgements: The GAIN initiative, The Netherlands Organization for Scientific Research, The Center for Cognitive and Neurological Research, GenomEUTwin, European Research Council, NARSAD, Geestkracht (ZonMW)