Reliable Heritability Clustering – Finding Distinct Genetic Factors that Influence Cortical Thickness

Rachel Walton[1], Anderson Winkler[2], John Blangero[3], Peter Fox[4], David Glahn[2], Thomas Nichols[1]

1. University of Warwick Statistics, Coventry, United Kingdom. 2. Yale University School of Medicine, New Haven, CT. 3. Southwest Foundation for Biomedical Research, San Antonio, TX. 4. University Of Texas Health Science Center At San Antonio, San Antonio, TX.

Introduction

Brain imaging phenotypes exhibit extremely high heritability, or h^2, with up to 88% of variability in gray matter volume explained by additive genetic factors [1]. In this work we explore the multivariate patterns of heritability. Correlation between any pair of phenotypes is the combination of a genetic and environmental correlation. Hierarchical clustering can be performed using genetic correlations as a distance measure, which will identify brain regions that appear to be influenced by distinct genetic factors.

A weakness of hierarchical clustering, however, is the lack of information on the reliability of the clusters obtained. Other authors [2] have addressed this issue by trying various different exploratory methods. Here we use a parametric bootstrap on the estimated genetic correlations to judge the stability of clustering solutions. We evaluate five different transformations of the genetic correlations to find the most stable and interpretable measure.

Methods

Using a population study of Mexican-American families comprised of 632 subjects, we used Freesurfer to obtain surface thickness on each point in cortical mesh, and these thicknesses were averaged within 51 bilateral ROIs (Fig 1). For each pair of ROIs quantitative genetic analysis was performed with SOLAR. Bivariate analyses were performed to decompose phenotypic correlations between ROIs in terms of the genetic (ρ_g) and environmental (ρ_e) correlations, accounting for kinship & heritability. Maximum likelihood estimation provides point estimates, standard errors and P-values for all parameters. All genetic analyses were conducted with age, sex, age*sex, age2, and age2*sex included as covariates.

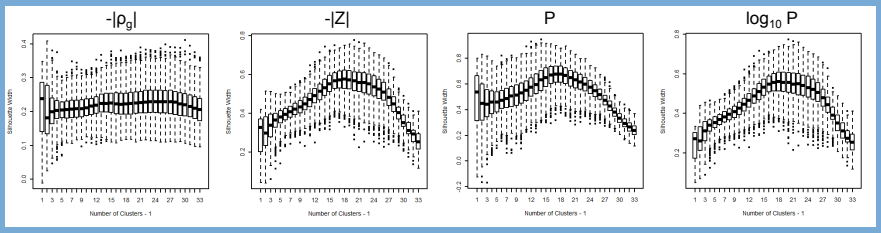
We apply hierarchical clustering (complete linkage) to 4 distance measures based on ρ_g:

- $-|\rho_g|$
- $-|Z|$, where Z is the test for Ho: $\rho_g = 0$
- P-value of Z
- $\log_{10}$ P-value of Z

The minus absolute value transformation is needed to ensure that 'small' distances correspond to strong genetic correlations; also, all measures are shifted to ensure the minimum distance is 0. Each distance measure represents a different compromise between reducing the weight of highly variable estimates and retaining the interpretability of the original measures.

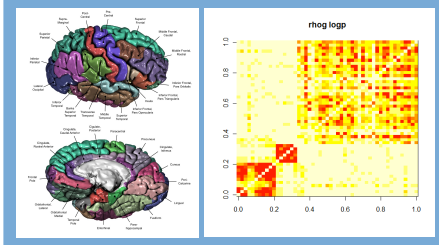
The fidelity of the clustering solution and the real data is measured by the cophenetic correlation, and compactness and separation of clusters is measured with the silhouette width [3]. To gauge stability we used a parametric bootstrap, consisting of 1000 Monte Carlo realizations of ρ_g where the population mean and standard deviation are set equal to the original point estimate and standard error, respectively. For each realization new clustering solution found and silhouette width found. Final fit is based on the original data, but the bootstrap is used to choose the most reliable distance measure.

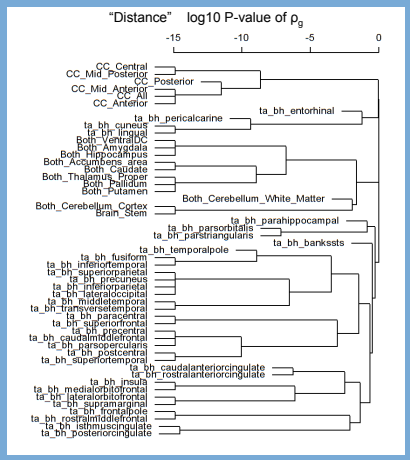
Figure 1. Reliability of genetic correlation of cortical thickness with different distance measures



Plots show cluster reliability (as measured with the Silhouette Width) versus number of clusters over 1000 bootstrap realizations. Clusters based on the untransformed ρ_g have poor reliability for any number of clusters, while the other measures improved reliability and suggest that the optimal number of clusters is around 19.

Figure 2. Clustering of cortical thickness based on $\log_{10}$ P of ρ_g estimate





Left shows the ROI regions considered, and middle shows the log10 P-value of the ρ_g estimate between each pair of ROIs. Clearly visible are the distinct genetic correlations within the 11 subcortical ROIs, the 6 corpus callosum ROIs, and the 34 cortical ROIs. Right shows the dendogram based on using log10 P as the measure of distance between each ROI. Clusters found on the right of the dendogram are those that we can be most confidence

Results

The cophenetic correlations between raw distance matrix and the hierarchical clustering solution are shown in Table 1. While original data results suggests that ρ_g has the best clustering, under Monte Carlo realizations, the cophenetic correlation values are much lower reflecting, and now the $\log_{10}$ P values provide the best stability, most closely matching the original ($\log_{10}$ P values distances)

	Cophenetic Correlations	
Distance Measure	Original Data	Bootstrap Samples
$-\|\rho_g\|$	0.7216	0.5320
$-\|Z\|$	0.7736	0.6403
P	0.5711	0.3774
$\log_{10}$ P	0.7224	0.6502

Table 1. Original and Bootstrap-resampled Cophenetic correlations.

Figure 1 shows the silhouette width over bootstrap samples for different number of clusters. All transformed distance measured out-performed $-|\rho_g|$ and offered more information on the optimal number of clusters. The 'shoulder' of the $-\log_{10}$ P curve suggests 19 clusters is a good number. Figure 2 shows the clustering solution using $-\log_{10}$ P as the distance measure; the dendigram shows that many pairs or triplets of ROI's have ultra-high confidence (log10 P having minimum possible value, -15), with other groupings having successively lower confidence.

Conclusions

We have demonstrated that the stability of clustering results using raw genetic correlation estimates can be low, and much greater stability of results can be obtained by basing distances on $\log_{10}$ P-values or other measures based on the statistical precision of the correlation estimates.

References

1. Glahn et al. (2007), "Neuroimaging Endophenotypes: Strategies for Finding Genes Influencing Brain Structure and Function," *Human Brain Mapping*, 28:488-501.
2. Schmitt et al. (2008), "Identification of Genetically Mediated Cortical Networks: A Multivariate Study of Pediatric Twins and Siblings." *Cerebral Cortex*, 18:1737-1747.
3. Knowles et al. (2005), "Computational cluster validation in post-genomic data analysis." *Bioinformatics*, 21:3201-3212.