HHS Public Access

Author manuscript

Proc IEEE Int Symp Biomed Imaging. Author manuscript; available in PMC 2018 June 19.

Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2017 ; 2017: 1226–1230. doi:10.1109/ISBI.2017.7950738.

APPROXIMATING PRINCIPAL GENETIC COMPONENTS OF SUBCORTICAL SHAPE

Boris A. Gutman[1], **Fabrizio Pizzagalli**[1], **Neda Jahanshad**[1], **Margaret J. Wright**[2,3], **Katie L. McMahon**[2], **Greig de Zubicaray**[4], and **Paul M. Thompson**[1]

[1]Imaging Genetics Center, University of Southern California, Los Angeles, CA, USA

[2]Queensland Brain Institute, University of Queensland, Brisbane, QLD, Australia

[3]Centre for Advanced Imaging, University of Queensland, Brisbane, QLD, Australia

[4]Queensland University of Technology, Brisbane, QLD, Australia

Abstract

Optimal representations of the genetic structure underlying complex neuroimaging phenotypes lie at the heart of our quest to discover the genetic code of the brain. Here, we suggest a strategy for achieving such a representation by decomposing the genetic covariance matrix of complex phenotypes into maximally heritable and genetically independent components. We show that such a representation can be approximated well with eigenvectors of the genetic covariance based on a large family study. Using 520 twin pairs from the QTIM dataset, we estimate 500 principal genetic components of 54,000 vertex-wise shape features representing fourteen subcortical regions. We show that our features maintain their desired properties in practice. Further, the genetic components are found to be significantly associated with the CLU and PICALM genes in an unrelated Alzheimer's Disease (AD) dataset. The same genes are not significantly associated with other volume and shape measures in this dataset.

Index Terms

imaging genetics; subcortical shape; brain imaging; genome-wide association study; Alzheimer's disease

1. INTRODUCTION

The field of Imaging Genetics aims to connect the human genetic code with quantitative imaging-based phenotypes. Two complimentary approaches are typically used: on the one hand, we search for common genetic variants (GV) which explain some variance in well-established phenotypes. On the other hand, we would like to discover signatures of complex imaging phenotypes associated with specific genes of interest. Both of these approaches are hampered by the very large dimensionality of both the genetic (dim = 10^6 – 10^8) and imaging (dim = 10^3 – 10^6) data, particularly in brain MR imaging. For example, simply identifying several new variants associated with subcortical volume required a Genome-

Wide Association Study (GWAS) of more than 20,000 subjects [1]. Alternative strategies for identifying GV-phenotype interactions exist. The most common approach in the genetics community connects a single (univariate) phenotype of interest with a weighted set of GVs known as a polygenic risk score [2]. The reverse strategy connects many phenotypes to individual GV's. In the meta-analytic context, this approach, known as TATES [3], combines many mass-univariate studies into a weighted phenotype score for each variant. Because this approach requires a correlation analysis between every phenotype and every GV, it becomes implausible for dense imaging phenotypes. Traditional fully multivariate linear models, such as Canonical Correlation Analysis (CCA) [4] or the closely related Partial Least Squares (PLS), have also been applied to imaging genetics. Though potentially more complex, these models remain computationally tractable even for very high-dimensional data, as they can be readily kernelized [5]. Further, the PLS approach can be extended into a multi-site meta-analysis study, making its use quite practical [6]. However, interpreting CCA or PLS component pairs – a set of weights for both the phenotypes and the genotypes – becomes challenging. Expensive post-hoc procedures may be needed to identify specific GVs implicated by the models [7].

In this paper, we take the intermediate approach more akin to the TATES model, but without the computational requirements. Our aim is to find a set of phenotypic components that best represent the genetically correlated aspects of the phenotype. We would like a set of features that maximally capture the phenotype heritability using a minimal number of components. In other words, by analogy to independent components analysis, we want components with (1) maximum heritability and (2) minimum genetic correlation between them. Should we find such a set of components, we should expect each of them to be maximally associated with some set of GVs, while at the same time having a different set of associated GVs from the other components. In this way, we hope to recover a universal set of features from a complex imaging phenotype that can be used to both empower a GWAS study and to improve the chances of finding associations with candidate polymorphisms. Simplified versions of the proposed pipeline have been used in brain imaging, particularly ones that focus on cortical parcellation [8]. However, the simplification of the complex genetic structure of the brain to discrete binary clusters is likely to miss important effects, likely reducing the power of such a representation. This effect has been shown in the context of disease biomarkers [9], for example.

For computational tractability, we use the narrow-sense heritability model in a family study of monozygotic and dizygotic twins to compute a genetic covariance matrix. While a full genetic components analysis would require an iterative log-likelihood optimization [10], we approximate the components by the eigenvectors of the genetic covariance matrix. Our imaging features are comprised of roughly 54,000 vertex-wise features mapped to boundary surfaces of fourteen subcortical regions. Our components have the desired properties of high heritability and genetic independence. At the same time, when applied to an unseen Alzheimer's dataset with different demographics and clinical characteristics, the resulting feature set is able to identify significant association with AD risk genes not associated with other subcortical shape-based features. This suggests that our approach may indeed lead to better-powered imaging genetics studies.

2. GENETIC COVARIANCE DECOMPOSITION

We begin by describing the univariate analysis of heritability. Family-based heritability estimates use maximum likelihood variance decomposition methods to break down the population variance and identify narrow sense heritability based on the expected kinship of the individuals. The covariance matrix Ω for a pedigree of individuals is given by:

$$\Omega = 2 \cdot \Phi \cdot \sigma_g^2 + I \cdot \sigma_e^2, \quad (1)$$

where $\boldsymbol{\sigma}_g^2$ is the genetic variance due to the additive genetic factors, $\boldsymbol{\Phi}$ is the kinship matrix representing the pair-wise kinship coefficients among all individuals, $\boldsymbol{\sigma}_e^2$ is the variance due to individual-specific environmental effects, and I is an identity matrix. Narrow sense heritability is defined as the fraction of phenotypic variance $\boldsymbol{\sigma}_p^2$ attributable to additive genetic factors,

$$h^2 = \sigma_g^2 / \sigma_P^2. \quad (2)$$

By analogy to general (phenotype) covariance, we can extend (1) to genetic covariance between phenotype X and Y. This can be estimated by fitting the covariance decomposition model

$$\Omega_{XY} = 2 \cdot \Phi \otimes \sum_g + I \otimes \sum_e, \quad (3)$$

where ⊗ is the Kronecker product operator, and Σ_g, Σ_e are the genetic and environmental covariance matrices. The intuition behind the closely related genetic correlation, $\rho_g = \sigma_{XY_g}^2 / \left(\sigma_{X_g}^2 \sigma_{Y_g}^2\right)$, may be described as the extent to which relatedness of two individuals predicts phenotype Y in one individual, knowing phenotype X in the other. Of more interest to us is another interpretation: genetic correlation may imply shared genetic associations, also known as pleiotropy (Fig. 1).

With the above interpretation of genetic correlation in mind, it should become clear why a principal genetic decomposition can be expected to increase power in association studies. Now, suppose we wish to find a set of genetic and environmental principal components (PCs) $\boldsymbol{\Psi}_{\boldsymbol{G}}$ and $\boldsymbol{\Psi}_{\boldsymbol{E}}$ to fit our data:

$$\boldsymbol{y} = \boldsymbol{F}\beta + \boldsymbol{\Psi}_{\boldsymbol{G}}\alpha + \boldsymbol{\Psi}_{\boldsymbol{E}}\gamma + \varepsilon, \quad (4)$$

with $\boldsymbol{F}$ representing fixed effects, e.g. sex and age. An optimization problem to recover the genetic components can then be written as −2 log L =

$$const + m_G \log|2\Phi| + \log|\boldsymbol{R}| + \log|\boldsymbol{C}| + \boldsymbol{y}^T \boldsymbol{P} \boldsymbol{y}. \quad (5)$$

In (5), the last two terms are functions of both the PCs and the fixed effect term in (4). Kirkpatrick [10] solves (5) approximately with iterative PC pair rotation. However, with even a modest-sized problem, such an optimization becomes numerically challenging. Here, we opt instead for directly decomposing the genetic covariance matrix into its spectral components:

$$\sum_g = \tilde{\boldsymbol{\Psi}}_G^T \Lambda \tilde{\boldsymbol{\Psi}}_G \quad (6)$$

Computing Σ_g requires solving a series of simpler bivariate log likelihood problems, which nevertheless have some computational cost as well. We use standard PCA on two unrelated sets of individuals of equal size in our family study for initial dimensionality reduction. Phenotypic PCs are computed separately on each set and combined, with Σ_g estimated on PC coordinates. Because our initial feature set is twice the number of twin pairs, the genetic covariance matrix cannot be full-rank. For this reason, we estimate the genetic covariance structure of the PC coordinates, and order the Σ_g eigenvectors by their heritability. In practice, this approach works quite well, eliminating redundant components. The final genetic component loadings can then be trivially mapped via a linear combination to the original shape vertex coordinates.

3. SUBCORTICAL SHAPE FEATURES

Our subcortical shape measures are computed using a previously described pipeline [11, 12], available via the ENIGMA Shape package[1]. Briefly, structural MR images are parcellated into cortical and subcortical regions (FreeSurfer 5.3). The binary region images are then surfaced with triangle meshes and parametrically (spherically) registered to a common region-specific surface template [13]. This leads to a one-to-one surface correspondence across the dataset at roughly 27,000 vertices describing the left and right thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and nucleus accumbens. Each vertex $\boldsymbol{p}$ is endowed with two shape descriptors:

1. Medial Thickness, $D(\boldsymbol{p}) = \|\boldsymbol{c_p} - \boldsymbol{p}\|$, where $\boldsymbol{c_p}$ is the point on the medial curve $\boldsymbol{c}$ closest to $\boldsymbol{p}$
2. Log of the Jacobian determinant J arising from the template mapping φ, J: $T_{\varphi(\boldsymbol{p})}\mathcal{M}_t \rightarrow T_{\boldsymbol{p}}\mathcal{M}$

The resulting shape descriptors can be readily mapped to the subcortical models, and have been shown to be highly heritable both in family and general population studies [11,12].

[1] http://enigma.usc.edu/protocols/imaging-protocols/

4. EXPERIMENTS

Our principal genetic components were estimated using 1040 twins scanned in the Queensland Twin Imaging Study (QTIM), 148 monozygotic pairs, mean age 23.1 +/− 3.1. We used the Sequential Oligogenic Linkage Analysis Routines (SOLAR) software package[2] [14] for heritability and genetic correlation analysis. SOLAR uses maximum likelihood variance decomposition methods. We visualize the principal components of the phenotypic (full) variance and genetic variance in Figures 2 and 3. To assess the parsimony of the genetic structure representation, we first visualize the genetic covariance matrices of the first 20 PCs (Fig. 4). Model fit quality using up to the first five components is compared in Table 1, using the Akaike Information Criterion $AIC = 2k - 2\log(L)$

To test our hypothesis that genetic PCs are better able to capture effect of known variants, we selected 16 genes known to increase risk of acquiring Alzheimer's Disease. We used 686 subjects from the ADNI 1 cohort, 144 AD patients, 337 patients with Mild Cognitive Impairment (MCI), and 205 healthy controls, mean age = 76.0 +/−5.1. Using the first 100 genetic component coordinates as phenotypes, we performed candidate variant association analysis on the selected AD risk genes. We performed the same analysis using (1) phenotypic PC coordinates, (2) vertex-wise features, using False Discovery Rate correction, and (3) region volumes.

The effect of APOE was amply detected using several of the phenotypic components as well as hippocampal vertex-wise measures, consistent with the literature. No other gene was significantly associated with any of the traditional subcortical volume and shape measures tested. Encouragingly, the PICALM and CLU genes were significantly associated with two distinct genetic components of variance after Bonferroni correction. The overall comparison of standard PCA and principal genetic components is displayed in Figure 5.

5. CONCLUSION

We have presented a straightforward approach to identify maximally heritable components of complex imaging phenotypes. We use the family-based study design to enable robust estimation of genetic components via eigenvalue decomposition of the genetic covariance matrix. The deeply structured nature of brain imaging, at odds with the largely unstructured human genome, presents both a challenge and an opportunity in imaging genetics. Here, we have exploited the structure of the brain to generate a universal set of imaging-based keys which we hope can be used to unlock the brain's genetic code with more ease. Our preliminary results using a modest-sized Alzheimer's Disease dataset suggest that our genetic components do indeed generalize to unseen data as we had hoped. Future work will continue to use our genetic shape representation in both GWAS and candidate GV studies, as well as explore using spatial regularization in genetic component recovery.

[2] http://www.nitrc.org/projects/se_linux

Acknowledgments

This work was supported in part by *NIH Big Data to Knowledge* (BD2K) Center of Excellence grant U54 EB020403, funded by a cross-NIH consortium.

References

1. Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivieres S, Jahanshad N, et al. Common genetic variants influence human subcortical brain structures. Nature. 2015; 520:224–229. [PubMed: 25607358]
2. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. Mar.2013 9:e1003348. [PubMed: 23555274]
3. van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013; 9:e1003235. [PubMed: 23359524]
4. Liu J, Calhoun VD. A review of multivariate analyses in imaging genetics. Frontiers in Neuroinformatics. 2014; 8:29. [PubMed: 24723883]
5. Lorenzi M, Gutman B, Hibar DP, Altmann A, Jahanshad N, Thompson PM, et al. Partial least squares modelling for imaging-genetics in Alzheimer's disease: Plausibility and generalization. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016:838–841.
6. Lorenzi M, Gutman B, Thompson P, Alexander D, Altmann A, Ourselin S, et al. Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study. 12th International Symposium on Medical Information Processing and Analysis (SIPAIM 2016). 2016 submitted.
7. Lorenzi M, Gutman B, Altmann A, Hibar D, Jahanshad Neda, Alexander D, et al. Linking Gene Pathways and Brain Atrophy in Alzheimer's Disease. Alzheimer's Association International Conference. 2016
8. Chen C-H, Fiecas M, Gutiérrez ED, Panizzon MS, Eyler LT, Vuoksimaa E, et al. Genetic topography of brain morphology. Proceedings of the National Academy of Sciences. Oct 15.2013 110:17089–17094. 2013.
9. Gutman BA, Hua X, Rajagopalan P, Chou Y-Y, Wang Y, Yanovsky I, et al. Maximizing power to track Alzheimer's disease and MCI progression by LDA-based weighting of longitudinal ventricular surface features. Neuroimage. 2013; 70:386–401. [PubMed: 23296188]
10. Kirkpatrick M, Meyer K. Direct Estimation of Genetic Principal Components: Simplified Analysis of Complex Phenotypes. Genetics. 2004; 168:2295–2306. [PubMed: 15611193]
11. Gutman BA, Jahanshad N, Ching CR, Wang Y, Kochunov PV, Nichols TE, et al. Medial demons registration localizes the degree of genetic influence over subcortical shape variability: An N= 1480 meta-analysis. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. 2015:1402–1406.
12. Roshchupkin* GV, Gutman* BA, Vernooij MW, Jahanshad N, Martin NG, Hofman A, et al. Heritability of the shape of subcortical brain structures in the general population. Nature Communications. 2016 vol Accepted.
13. Gutman, BA., Madsen, SK., Toga, AW., Thompson, PM. A Family of Fast Spherical Registration Algorithms for Cortical Shapes. In: Shen, L.Liu, T.Yap, P-T.Huang, H.Shen, D., Westin, C-F., editors. Multimodal Brain Image Analysis. Vol. 8159. Springer International Publishing; 2013. p. 246-257.
14. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. May.1998 62:1198–211. [PubMed: 9545414]

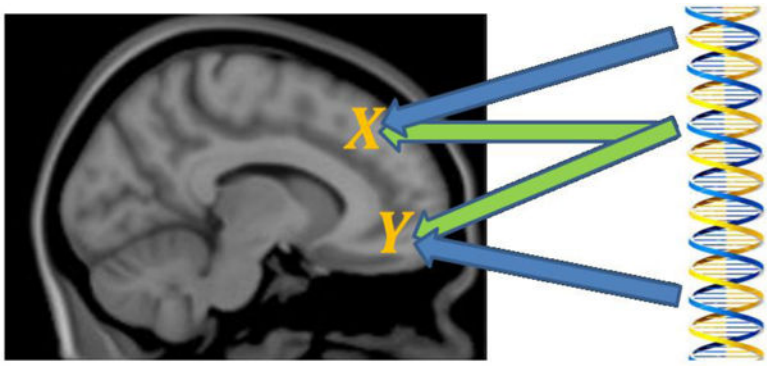

Figure 1. Interpreting Genetic Correlation

Trait pleiotropy, or shared genetic associations, is one of the explanations.

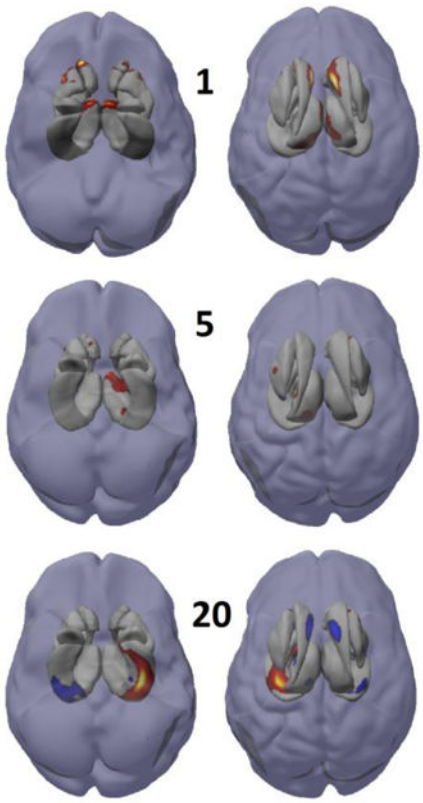

Figure 2. Principal phenotypic components of subcortical shape

The correlation structure captures the expected local spatial correlation, with the main components mainly dominated by a specific region.

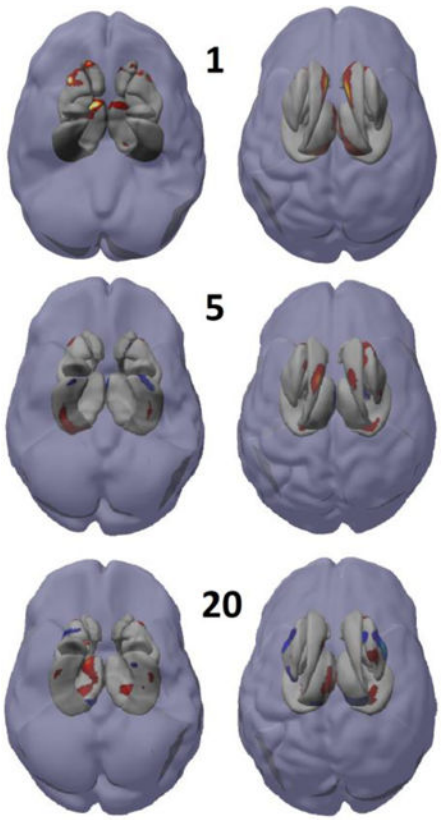

Figure 3. Principal genetic components of subcortical shape

The genetic correlation structure is more complex, with more inter-regional cross-talk.

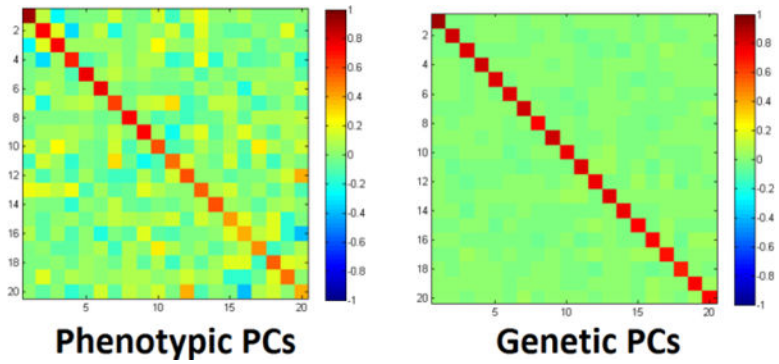

Figure 4. Genetic correlation of principal components

Heritability is displayed on the diagonal. Principal genetic components capture more shape heritability with minimal redundancy.

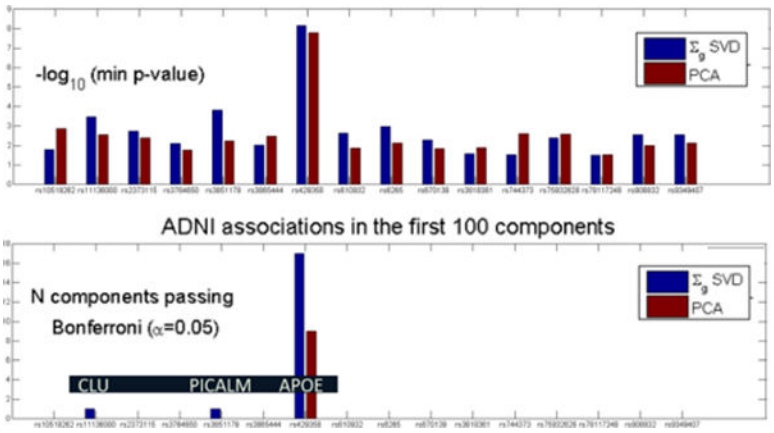

Figure 5. Associations of component coordinates with AD genes PCA vs. Principal Genetic Components

Top: log of the smallest p-value for each AD risk gene. **Bottom:** number of components passing Bonferroni correction. The genetic components are able to identify two additional genes.

Table 1

AIC of heritability models

As in figure 4, the table shows that the genetic components better explain shape heritability than standard principal components (lower AIC is better).

# of components	2	3	4	5
Genetic PCs	5772	4544	3156	1548
Phenotypic PCs	6146	5330	3398	1724