HIERARCHICAL CLUSTERING OF THE GENETIC CONNECTIVITY MATRIX REVEALS THE NETWORK TOPOLOGY OF GENE ACTION ON BRAIN MICROSTRUCTURE: AN N=531 TWIN STUDY

Ming-Chang Chiang[1], Marina Barysheva[1], Katie L. McMahon[2], Greig I. de Zubicaray[3], Kori Johnson[2], Nicholas G. Martin[4], Arthur W. Toga[1], Margaret J. Wright[4], Paul M. Thompson[1]

[1]Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Los Angeles, CA, USA
[2]University of Queensland, Centre for Advanced Imaging, Brisbane, Australia
[3]University of Queensland, School of Psychology, Brisbane, Australia
[4]Queensland Institute of Medical Research, Brisbane, Australia

ABSTRACT

Genetic correlation (r_g) analysis determines how much of the correlation between two measures is due to common genetic influences. In an analysis of 4 Tesla diffusion tensor images (DTI) from 531 healthy young adult twins and their siblings, we generalized the concept of genetic correlation to determine common genetic influences on white matter integrity, measured by fractional anisotropy (FA), at all points of the brain, yielding an *N*x*N* genetic correlation matrix r_g(x,y) between FA values at all pairs of voxels in the brain. With hierarchical clustering, we identified brain regions with relatively homogeneous genetic determinants, to boost the power to identify causal single nucleotide polymorphisms (SNP). We applied genome-wide association (GWA) to assess associations between 529,497 SNPs and FA in clusters defined by hubs of the clustered genetic correlation matrix. We identified a network of genes, with a scale-free topology, that influences white matter integrity over multiple brain regions.

Index Terms— diffusion tensor imaging, twins, hierarchical clustering, scale-free topology, genome-wide association

1. INTRODUCTION

A major goal of imaging genetics is to identify specific single nucleotide polymorphisms (SNP) that affect brain structure. Several international initiatives are collecting brain imaging data from thousands of subjects, along with genome-wide association scans (GWAS). Recently, voxel-based searches have been performed, for genetic markers that correlate with image-derived measures at each voxel in the brain (e.g., vGWAS [1]). An unbiased search of around half a million SNPs across the brain does not require prior hypotheses about which genetic variants matter, or which brain regions are relevant. However, its overall power is low due to the need to correct for millions of statistical tests across the genome and across the image.

Here we take a different approach to genetic discovery in imaging databases. We examine networks of genes that influence the integrity of the brain using diffusion tensor imaging (DTI). DTI is a variant of standard MRI, in which the 3D directionality of water diffusivity at each location in the brain is used to quantify fiber tract coherence and integrity [2]. We boost the power to discover genes that affect brain integrity, by merging powerful concepts from classical twin studies, hierarchical clustering, and network topology. In genetic analyses, the *genetic* correlation, or r_g, is defined as the proportion of the correlation between two traits that can be explained by common underlying genetic factors [3]. Here we generalize this concept to pairs of points in a 3D brain image. We form the genetic correlation matrix r_g(x,y), to define a 4876-by-4876 genetic correlation matrix between all pairs of selected voxels in the brain. Using network theory, we infer the dominant *hubs* of this *N*x*N* connectivity network by using hierarchical clustering to aggregate voxels with common genetic determination. By construction, these voxel sets, or clusters, are more homogeneous in their genetic determinants than random pairs of points in the brain. We then performed genome-wide association (GWA) to identify SNPs associated with fiber integrity in those clusters. SNPs are common variants in the genome, and over half a million are commonly genotyped in a GWA study. We then computed Spearman's rank correlation coefficient ρ for every pair of SNPs, to compare their partial regression coefficients with respect to FA, across all the clusters. A second co-occurrence matrix was derived to represent SNP pairs that tended to jointly affect brain regions. By applying network topology analysis to this SNP association matrix, a SNP network influencing white matter integrity was identified. The network was found to exhibit a scale-free topology, with several dominant hubs and structured sub-networks.

2. METHODS

2.1. Participants and SNP genotyping

531 healthy adult twins (103 monozygotic pairs and 106 dizygotic pairs) and their non-twin siblings (mean age: 23.7±2.1 SD years; 217 M/314 F) were recruited from 271 different families. Subjects' demographic information and the twin/sibling composition of the families are detailed in [4]. Genotype data were collected from 484 subjects. Genomic DNA samples were isolated from blood cells using standard protocols, and were analyzed on the Human610-Quad BeadChip (Illumina, Inc., San Diego, CA). 529,497 SNPs remained after quality control procedures [5]. We further excluded 12 subjects who were identified to be ancestry outliers [5], so 472 subjects remained in the subsequent GWA.

2.2. Image processing and registration

All MR images were collected using a 4 Tesla Bruker Medspec MRI scanner (Bruker Medical, Ettingen, Germany), with a

transverse electromagnetic (TEM) headcoil, at the Centre for Advanced Imaging (University of Queensland, Australia). Diffusion-weighted images were acquired using single-shot echo planar imaging with a twice-refocused spin echo sequence, to reduce eddy-current induced distortions. Imaging parameters were: 21 axial slices (5 mm thick), FOV = 23 cm, TR/TE 6090/91.7 ms, 0.5 mm gap, with a 128×100 acquisition matrix. 30 images were acquired: 3 with no diffusion sensitization (i.e., T_2-weighted images) and 27 diffusion-weighted images (b = 1145.7 s/mm^2) with gradient directions evenly distributed on an imaginary hemisphere. The reconstruction matrix was 128×128, yielding a 1.8×1.8 mm^2 in-plane resolution. Total scan time was 3.05 minutes. For each subject, a fractional anisotropy (FA) image was derived from the DTI, which had been resampled to isotropic voxel resolution (with dimensions: 128×128×93 voxels, resolution: 1.7×1.7×1.7 mm^3). The FA maps were then fluidly registered to a mean FA template, after appropriate preprocessing, as detailed in [4]. We averaged the fluidly-registered FA images voxelwise across all subjects and restricted subsequent data analysis to regions with average FA > 0.2, to focus the analysis on major white matter fiber structures. Each participant's FA map was smoothed using an isotropic Gaussian filter with full width at half maximum (FWHM) = 6 mm.

2.3. Defining regions of interest (ROI) for GWAS

We defined ROIs for GWAS from brain regions where genetic influences on white matter integrity were high, using a standard structural equation model (SEM). The SEM partitions the observed variance in FA at every voxel, across the subject sample, into components due to additive genetic factors (A), shared environment (C) and unique environment (E), giving the significance map for the genetic factors, p(A) [6, 7]. p(A) was further assessed using the standard false discovery rate method (FDR; [8]) to correct for multiple comparisons. We selected voxels where the genetic component, A, contributed at least 60% to the total variation of FA, and its significance passed the FDR ≤ 5% threshold. This led to 4876 candidate voxels for the ROIs. By doing this, we ensured that the main effect (i.e., genetic influences on FA) for the genetic correlation analysis below was significant. As FDR was applied, no more than 5% of the voxels declared as genetically influenced are likely to be false positive findings.

We estimated the genetic correlation coefficient r_g between any two voxels, using a classical "cross-trait cross-twin" analysis [7]. This yielded a 4876-by-4876 *genetic* correlation matrix. r_g = 1 indicates that correlation between FA of the two voxels is 100% attributable to the same underlying genetic factors. To generate ROIs for further analysis, we then identified clusters of voxels that were in high genetic correlation with each other using the hierarchical clustering method in the MATLAB Statistics Toolbox (Mathworks, Inc., Natick, MA). Hierarchical clustering treated every one of the 4876 voxels as a *node* in a network. Clustering began by treating each node as a single cluster. We then repeatedly merged the two nodes with the highest similarity, or the shortest distance, to create larger clusters. Similarity between any two nodes was defined by the extent of their interconnectivity, quantified by an index of topological overlap (TO) [9]:

$$w_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}. \quad (1)$$

Here a_{ij} = 0 or 1 is the adjacency function between nodes i and j. a_{ij} = 1 if the genetic correlation r_g between them is not less than 0.95; this indicates that the two nodes are connected. $l_{ij} = \sum_{u \neq i,j} a_{iu} a_{uj}$ is the number of nodes to which both nodes i and j are connected, and $k_i = \sum_{u \neq i} a_{iu}$ is the connectivity of node i. w_{ij} is a weight lying between 0 and 1; higher w_{ij} indicates that nodes i and j are linked to more common neighbors, and thus are more likely part of highly connected *modules* of the network [9]. We defined the distance between any two nodes i and j as $d_{ij} = 1 - w_{ij}$. The distance between two clusters that contain more than one node was defined as the average distance across all pairs of nodes in the two clusters. Finally, we generated a hierarchical tree of clusters at different distance levels (**Fig. 1**). Here a cluster at level d means that the distance between every pair of all the sub-clusters it contains is not greater than d. By setting a threshold of d = 0.5, we obtained 233 clusters. We selected 18 of them as our ROIs – the ones that were no less than 50 voxels (246 mm^3) in size.

2.4. GWAS for white matter integrity

The mean FA for each ROI was associated respectively with each genotyped SNP across the whole genome (n = 529,497), using the Quantitative Transmission Disequilibrium Test (QTDT) program [10]. FA was regressed against an intercept, subjects' age and sex included as nuisance covariates, and the SNP genotypic value. For a SNP with two alleles A and a, its genotypic value = 1 for AA, 0 for Aa, and -1 for aa. A total of 529,497 SNPs × 18 ROIs = 9,530,946 associations were performed. We used the traditional Bonferroni method to correct for multiple comparisons across the genome: a SNP with a P-value less than 0.05/529,497 = 9.4 × 10^{-8} was declared to be significantly associated with FA. For each GWAS-significant SNP, we further corrected for multiple comparisons across the ROIs using the bootstrap method. The P-value of the SNP was converted to a Z-value using the inverse normal cumulative distribution function. Then, the P-value adjusted for multiple comparisons, or the q-value, was derived by comparing the sum of Z-values across all the 18 ROIs for that SNP with the bootstrap distribution of 100,000 sums of Z-values randomly selected from the 529,497 SNPs. A SNP with $q \leq 0.05$ was considered to reach overall significance.

2.5. Gene interconnection network

We extended our analysis from detecting the effects of individual genetic loci at single locations, to constructing a gene network that influences white matter integrity over multiple brain regions. In the network scheme, the connectedness between SNPs was derived from the correlations between their effects on the phenotype (here, FA). We selected SNPs whose association with FA in at least one ROI had a significance P-value < 10^{-5}. This criterion was somewhat relaxed (compared to the very strict GWAS-significant threshold with P < 9.4 × 10^{-8}) to include more SNPs in the network analysis. This might therefore allow higher sensitivity to detect statistical dependencies between gene effects. We computed the Spearman's rank correlation coefficient, ρ, for every pair of SNPs, based on their partial regression coefficients with respect to FA across all the 18 ROIs. The partial regression coefficient β for a SNP represents change in FA due to the unit additive effect of the dominant allele (e.g., the effect of allele A that results in difference in FA between Aa and aa genotypes). We then applied a hard threshold on the correlation matrix to visualize the network: SNPs i and j were considered to be connected (i.e., a_{ij} = 1) if $\rho_{ij} \geq 0.9$.

3. RESULTS

Figure 1 displays the 18 ROIs (≥ 50 voxels) identified by hierarchical clustering based on the 4876-by-4876 topological overlap (TO) matrix. We assumed that there should be more power to detect associations between FA in these ROIs and SNPs identified from the whole genome, as the FA of voxels in an individual ROI was strongly determined by a common set of genes. This is corroborated by the 24 SNPs found in GWAS, which showed strong "genome-wide significant" associations with white matter integrity (FA), with an uncorrected P-value $< 9.4 \times 10^{-8}$, and an overall significance (across all the 18 ROIs) q-value ≤ 0.05.

Figure 2 shows the connectivity network of the SNPs that affect white matter integrity. Two SNPs are defined as *connected*, if their effects on FA (their partial regression coefficients with respect to FA) are strongly correlated ($\rho \geq 0.9$). Most SNPs have no or only scarce connections with other SNPs. However, a small proportion of SNPs is highly connected with each other, and forms a connected subset, or *module*, of the network. In fact, this SNP network is in a *scale-free topology* [11], as shown in **Figure 3**, where the number of connections approximately obeys a power-law distribution. If n is the number of connections of a SNP to others, the probability of n, namely $f(n)$, followed an approximate power law, $f(n) \propto n^{-1.25}$ (Pearson's correlation coefficient $r = -0.86$, $P = 4.0 \times 10^{-6}$).

4. DISCUSSION

In this study we applied network theory methods to identify networks of genetic variants that influence brain fiber integrity in 531 healthy young adults. Using the genetic correlation as a means to find voxels with common genetic influences, we defined ROIs, whose component voxels had FA values that were determined by a highly overlapping set of genes (quantified by the genetic correlation matrix, $r_g(x,y)$, that compares all pairs of voxels). We detected 24 SNPs from the whole genome with significant effects on white matter integrity, after conservative Bonferroni-type correction for multiple comparisons across the genome and across all the ROIs. The associations between these SNPs and FA require replication, when data from larger independent cohorts becomes available, for example via the Enigma project (http://enigma.loni.ucla.edu). Two advantages of this genetic network analysis are apparent: (1) genetic correlation can tap into the natural latent structure of gene action in a brain image; (2) voxel clustering by genetic affinity leads to high power to find SNPs with correlated effects in genome-wide scans.

We also recovered a network of SNPs that influence brain white matter integrity. This network exhibited a scale-free topology, with connection probabilities approximately following a power law. The scale-free property has been discovered in various biological networks, e.g., for gene or protein interactions, or even the World Wide Web and other social networks [9]. Scale-free networks have high degrees of error tolerance, and may remain stable even their nodes are randomly damaged [11]. As white matter integrity is critical for normal brain function, we believe that the underlying genetic network that influences it needs to be sufficiently robust, to resist environmental effects and stressors.

Our method will likely empower work in imaging genomics and connectomics (e.g., the Human Connectome Project). As SNP aggregation leads to substantial dimension reduction, the resulting networks become strong candidates for verification and replication, without the computational and statistical burden of a full image-wide genome-wide search.

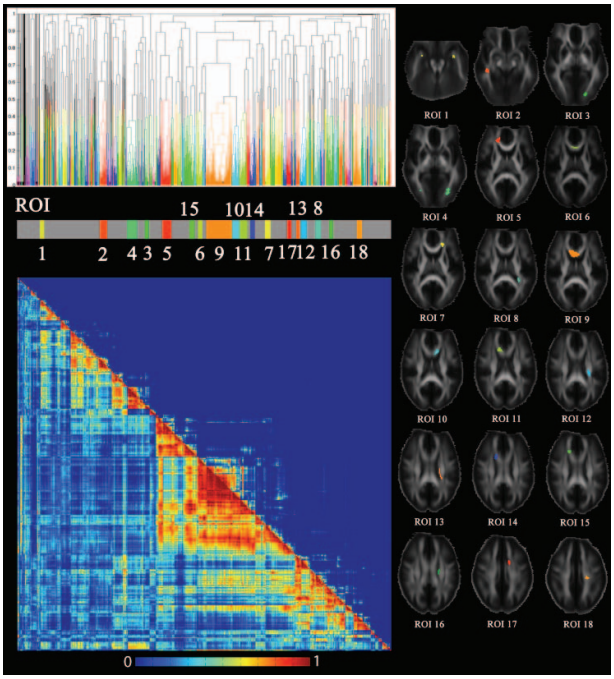

Fig. 1. The dendrogram (*upper left*) shows how the hierarchical clustering method aggregates the 4876 candidate voxels into clusters, based on the topological index between each pair of voxels. Each branch of the dendrogram coincides the corresponding column in the color-coded matrix (*lower left*), where the TO index matrix is shown in the upper triangle, and the genetic correlation matrix in the lower triangle. 18 clusters composed of no less than 50 voxels were selected as ROIs. They are shown (*right*) using a specific color for each ROI. The ROIs are sequentially numbered from the inferior to superior level of the brain.

REFERENCES

1. Stein, J.L., et al., *Voxelwise genome-wide association study (vGWAS).* Neuroimage, 2010. **53**(3): p. 1160-74.
2. Basser, P.J., J. Mattiello, and D. LeBihan, *MR diffusion tensor spectroscopy and imaging.* Biophys J, 1994. **66**(1): p. 259-67.
3. Neale, M.C., L.R. Cardon, and the NATO Scientific Affairs Division, *Methodology for genetic studies of twins and families.* 1992, Dordrecht; Boston: Kluwer Academic Publishers. xxv, 496 p.
4. Chiang, M.C., et al., *Genetics of white matter development: A DTI study of 705 twins and their siblings aged 12 to 29.* Neuroimage, 2011. **54**(3): p. 2308-2317.
5. Medland, S.E., et al., *Common variants in the trichohyalin gene are associated with straight hair in Europeans.* Am J Hum Genet, 2009. **85**(5): p. 750-5.
6. Posthuma, D., et al., *Multivariate genetic analysis of brain structure in an extended twin design.* Behav Genet, 2000. **30**(4): p. 311-9.

7. Chiang, M.C., et al., *Genetics of brain fiber architecture and intellectual performance.* J Neurosci, 2009. **29**(7): p. 2212-24.
8. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J R Statist Soc B, 1995. **57**(1): p. 289-300.
9. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.
10. Abecasis, G.R., L.R. Cardon, and W.O. Cookson, *A general test of association for quantitative traits in nuclear families.* Am J Hum Genet, 2000. **66**(1): p. 279-92.
11. Albert, R., H. Jeong, and A.L. Barabasi, *Error and attack tolerance of complex networks.* Nature, 2000. **406**(6794): p. 378-82.

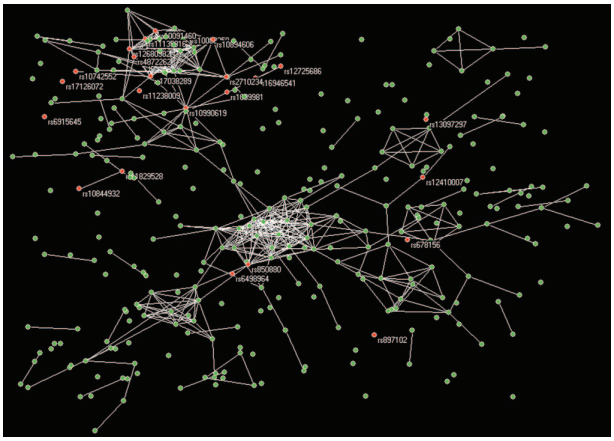

Fig. 2. Gene interconnection network. The nodes of the network, or SNPs, are displayed as green circles. SNPs whose associations with FA reach overall significance (as listed in **Table 1**) are colored in red, with their names labeled as well. Two SNPs linked by a white line have strongly correlated effects on FA, and thus these two SNPs are defined as "connected".

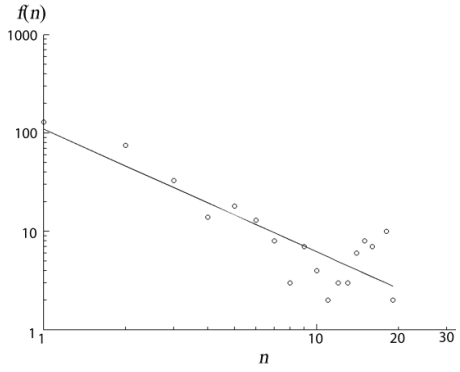

Fig. 3. Scale-free topology of the gene network. The log-log plot for the frequency of n connected SNPs, or $f(n)$, with respect to n shows that the gene network displayed in Fig. 2 is in a scale-free topology. $f(n) = 109.55 \cdot n^{-1.25}$ (Pearson's correlation coefficient $r = -0.86$, $P = 4.0 \times 10^{-6}$). $n = 1$ means that the SNPs are isolated and not connected to any other SNPs.

Table 1. SNPs whose associations with FA reach overall significance, at levels that correct for the genome-wide search, and correct for the number of ROIs assessed.

SNP	ROIs (P-value)	Chromosome and name of gene containing the SNP
rs12410007	6 (3.0×10^{-8})	Chr1, *KIAA1026*
	9 (1.0×10^{-8})	
rs12725686	12 (2.0×10^{-11})	Chr1, *FHAD1*
	13 (9.0×10^{-10})	
rs850880	11 (9.0×10^{-8})	Chr2
	15 (6.0×10^{-8})	
rs13097297	12 (8.0×10^{-8}),	Chr3
	13 (2.0×10^{-8})	
rs17038289	13 (1.0×10^{-8})	Chr4, *HADH*
rs6915645	8 (6.0×10^{-10})	Chr6
rs11829528	12 (2.0×10^{-9})	Chr7, *LHFPL3*
	13 (2.0×10^{-8})	
rs2710234	13 (2.0×10^{-8})	Chr7
rs10088359	12 (8.0×10^{-10})	Chr8
	13 (1.0×10^{-11})	
rs10091460	12 (1.0×10^{-9})	Chr8
	13 (2.0×10^{-11})	
rs11135816	12 (8.0×10^{-10})	Chr8
	13 (1.0×10^{-11})	
rs12680982	13 (3.0×10^{-8})	Chr8
rs4872262	13 (3.0×10^{-8})	Chr8
rs10990619	12 (1.0×10^{-8})	Chr9
	13 (6.0×10^{-11})	
rs10894606	13 (2.0×10^{-8})	Chr11, *OPCML*
rs10742552	12 (3.0×10^{-8})	Chr11
	13 (1.0×10^{-8})	
rs11238009	8 (4.0×10^{-8})	Chr11
rs897102	1 (8.0×10^{-9})	Chr11
rs10844932	12 (2.0×10^{-8})	Chr12
rs1009981	13 (2.0×10^{-8})	Chr12
rs16946541	13 (8.0×10^{-8})	Chr12, *MED13L*
rs17126072	8 (2.0×10^{-8})	Chr14, *DDHD1*
rs6498964	12 (7.0×10^{-8})	Chr16
	13 (9.0×10^{-9})	
rs678156	13 (6.0×10^{-8})	Chr16, *LPIN2*

Chr: chromosome. *DDHD1*: DDHD domain containing 1; *KIAA1026*: kazrin; *HADH*: hydroxyacyl-CoA dehydrogenase; *OPCML*: opioid binding protein/cell adhesion molecule-like; *LHFPL3*: lipoma HMGIC fusion partner-like 3; *FHAD1*: forkhead-associated (FHA) phosphopeptide binding domain 1; *MED13L*: mediator complex subunit 13-like; *LPIN2*: lipin 2.

Acknowledgments. This work was funded by NIH R01 grants HD050735, R01 EB008281, NHMRC Grant 496682, and P41 RR013642.