Probing multivariate associations between structural neuroimaging phenotypes and genetic markers

Mert R. Sabuncu
Athinoula A. Martinos Center for Biomedical Imaging
Massachusetts General Hospital and Harvard Medical School

MICGen Workshop at MICCAI 2014
Sept 14, 2014
Structural Brain MRI

- sMRI is a ubiquitous imaging modality
- Shows neuroanatomical structures
- A typical scan is $256 \times 256 \times 256$ (1mm$^3$)
- Many neurological conditions are associated with sMRI-derived markers
- E.g., hippocampal volume is a sensitive marker of dementia
Neuroanatomy is largely heritable

Phenotype \( (P) = \text{Genotype} \ (G) + \text{Environment} \ (E) + G \times E \)

Heritability \( h^2 \): the proportion of the phenotypic variance in a trait attributable to the additive effects of genes

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Test-retest reliability</th>
<th>( h^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>intracranial volume (ICV)</td>
<td>0.995</td>
<td>0.849</td>
</tr>
<tr>
<td>total brain volume</td>
<td>0.997</td>
<td>0.981</td>
</tr>
<tr>
<td>left hemispheric cortical GM volume</td>
<td>0.992</td>
<td>0.521</td>
</tr>
<tr>
<td>right hemispheric cortical GM volume</td>
<td>0.991</td>
<td>0.492</td>
</tr>
<tr>
<td>total cortical GM volume</td>
<td>0.994</td>
<td>0.515</td>
</tr>
<tr>
<td>total subcortical GM volume</td>
<td>0.968</td>
<td>0.357</td>
</tr>
<tr>
<td>total GM volume</td>
<td>0.995</td>
<td>0.475</td>
</tr>
<tr>
<td>left hemispheric WM volume</td>
<td>0.996</td>
<td>0.416</td>
</tr>
<tr>
<td>right hemispheric WM volume</td>
<td>0.996</td>
<td>0.302</td>
</tr>
<tr>
<td>total WM volume</td>
<td>0.996</td>
<td>0.369</td>
</tr>
<tr>
<td>left hemispheric mean cortical thickness</td>
<td>0.899</td>
<td>0.688</td>
</tr>
<tr>
<td>right hemispheric mean cortical thickness</td>
<td>0.885</td>
<td>0.732</td>
</tr>
<tr>
<td>overall mean cortical thickness</td>
<td>0.935</td>
<td>0.734</td>
</tr>
<tr>
<td>left hemispheric total surface area</td>
<td>0.999</td>
<td>0.298</td>
</tr>
<tr>
<td>right hemispheric total surface area</td>
<td>0.997</td>
<td>0.288</td>
</tr>
<tr>
<td>total surface area</td>
<td>0.998</td>
<td>0.305</td>
</tr>
</tbody>
</table>
Neurological disorders are largely heritable

<table>
<thead>
<tr>
<th>Condition</th>
<th>Heritability</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td>&gt; 80%</td>
<td>Cannon et al. 1998</td>
</tr>
<tr>
<td>Alzheimer’s</td>
<td>&gt; 70%</td>
<td>Gatz et al. 1996</td>
</tr>
<tr>
<td>ALS</td>
<td>&gt; 60%</td>
<td>Al Chalabi et al. 2010</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>59%</td>
<td>Kendler et al. 1995</td>
</tr>
<tr>
<td>Panic Disorder</td>
<td>44%</td>
<td>Kendler et al. 1995</td>
</tr>
<tr>
<td>Major Depression</td>
<td>41%</td>
<td>Kendler et al. 1995</td>
</tr>
<tr>
<td>Autism</td>
<td>37%</td>
<td>Hallmayer et al. 2011</td>
</tr>
<tr>
<td>Phobia</td>
<td>35%</td>
<td>Kendler et al. 1995</td>
</tr>
<tr>
<td>Anxiety Disorder</td>
<td>32%</td>
<td>Kendler et al. 1995</td>
</tr>
<tr>
<td>Bulimia</td>
<td>30%</td>
<td>Kendler et al. 1995</td>
</tr>
</tbody>
</table>
Many neurological disorders are polygenic.

**LETTERS**

**Common polygenic variation contributes to risk of schizophrenia and bipolar disorder**

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[^1]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. Here we show, using two analytic approaches, the extent to which common genetic variation underlies the risk of schizophrenia. First, we implicate the major histocompatibility complex. Second, we provide molecular genetic evidence for a substantial polygenic component to the risk of schizophrenia involving thousands of common alleles of very small effect. We show that this component also contributes to the risk of bipolar disorder, but not to several non-psychiatric diseases.

Table 2, Supplementary Fig. 2 and section 5 and 6 in Supplementary Information.

The best imputed SNP, which reached genome-wide significance (rs3130297, \( P = 4.79 \times 10^{-8} \), T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb)), was also in the MHC, 7 kilobases (kb) from NOTCH4, a gene with previously reported associations with schizophrenia[^1]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at \( P < 10^{-5} \), found on the ancestral European haplotype[^2] (Table 1, Supplementary Table 3 and section 3 in Supplementary Information). However, it was not possible to ascribe the association to a specific HLA allele, haplotype or region (Supplementary Table 3 and

[^1]: Purcell et al. 2009
[^2]: rs3130297, \( P = 4.79 \times 10^{-8} \), T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb), was also in the MHC, 7 kilobases (kb) from NOTCH4, a gene with previously reported associations with schizophrenia.
Neurological disorders have distributed footprints

AD: Alzheimer’s Disease; bvFTD: behavioral variant frontotemporal dementia; SD: semantic dementia; PNFA: progressive nonfluent aphasia; CBS: corticobasal syndrome

Seeley et al. 2009
Cortical atrophy signature of Alzheimer’s disease

Dickerson et al. 2009
Neuroimaging genetics relationships are multivariate

- Multiple genes influence neuroanatomy and disease
- Epistasis, i.e., gene-gene interactions
- Gene-risk factor interactions
- Multivariate patterns of disease-associated anatomical alterations
MRI scans are 3D matrices of intensities

Image processing aims to extract biologically meaningful measurements

Several software tools available: FreeSurfer\(^1\), SPM\(^2\), CARET\(^3\), BrainSuite\(^4\)

---

\(^1\) freesurfer.net
\(^2\) www.fil.ion.ucl.ac.uk/spm
\(^3\) http://brainvis.wustl.edu/wiki/index.php/Caret:About
\(^4\) brainsuite.usc.edu
Cortical thickness and area values sampled at > 300,000 points on the surface
Segmentations and volume measurements of over 100 cortical and sub-cortical GM structures and 60 WM structures
Nonlinear across-subject volumetric registration
3 Approaches for Multivariate Neuroimaging Genetics

Build on prior work

- Build on prior case-control datasets:
  Supervised Machine Learning/Predictive Modeling
- Build on published results:
  Bayesian (Generative) Modeling
- Build on known biology:
  Kernel Machine Regression
Supervised Machine Learning/Predictive Modeling

- Classical stats deals with small \( p \), moderate/large \( n \)
- Typical ML, e.g., Support Vector Machines, deals with large \( p \), small \( n \)
- Supervised learning: construct model from training data
MRI-based Clinical Prediction

- Various classification/regression algorithms
  - Support Vector Machines, Random Forests, Relevance Vector/Voxel Machines
- Multivariate prediction is more accurate than univariate markers
- Detailed analysis of factors that influence prediction performance

Sabuncu and Konukoglu 2014
We employed Relevance Vector Machine to discriminate AD patients versus matched controls.
Figure 2. Thickness across AD-vulnerable cortex versus Alzheimer-associated non-APOE polygenic score in CN subjects with subthreshold levels of amyloid burden (N = 64, p = -0.29, P < 0.05).
Bayesian Modeling

Probabilistic Approach to Joint Modeling of Imaging and Genetics
Kayhan Batmanghelich, Adrian Dalca, Mert Sabuncu and Polina Golland
Kernel Machine Regression

- Model the aggregated effect of a collection of SNPs on the phenotype in a flexible framework
- Semiparametric model:
  - $y_i = X_i \beta + h(G_i) + \epsilon_i$
  - $G_i = (g_{i1}, \cdots, g_{iL})$ - a genomic region with $L$ genetic variants
  - $h(\cdot)$ - an arbitrary function located in a reproducing kernel Hilbert space $\mathcal{H}_K$ defined by a nonnegative-definite kernel function matrix $K$
  - $\epsilon_i \sim N(0, \sigma^2)$
  - Minimize the penalized square-error loss function w.r.t $\beta$, $\sigma^2$ and $h$:
    \[
    \mathcal{L} = \frac{1}{2} \sum_{i=1}^{N} \{y_i - X_i \beta - h(G_i)\}^2 + \frac{\lambda}{2} \|h\|_{\mathcal{H}_K}^2
    \]
Kernel Machine Regression

- Efficient hypothesis testing
- Choose kernel type for flexibility
- Can be extended to quantify overall heritability and $G \times E$ effects
  - Detecting Gene-Environment Interactions via a Kernel Machine Method
    Tian Ge, Thomas Nichols, Debashis Ghosh, Elizabeth Mormino, Jordan Smoller and Mert Sabuncu
  - Fast Heritability Analysis Using Genome-Wide Data via Kernel Machines
    Tian Ge, Thomas Nichols, Avram Holmes, Phil Lee, Joshua Roffman, Randy Buckner, Mert Sabuncu and Jordan Smoller
Summary: Multivariate Imaging Genetics

- **Supervised Machine Learning/Predictive Modeling**
  - Pros: Several algorithms, good prediction performance
  - Cons: Biological interpretability is limited

- **Bayesian (Generative) Modeling**
  - Pros: Flexible, well-established machinery
  - Cons: Computational cost, modeling assumptions

- **Kernel Machine Regression**
  - Pros: Flexible, efficient hypothesis testing
  - Cons: Non-Bayesian, so limited in quantifying and handling model uncertainty
Acknowledgements

- Collaborators: Randy Buckner, Jordan Smoller, Reisa Sperling, Bruce Fischl, Phil De Jager, Jonathan Rosand, Kayhan Batmanghelich, Tian Ge, Polina Golland, Adrian Dalca, ...

- Thanks to Tian Ge for the slides on kernel machine regression

- Funding from NIH NIBIB 1K25EB013649-01 and a BrightFocus grant (AHAF A2012333)