Probing multivariate associations between structural neuroimaging phenotypes and genetic markers

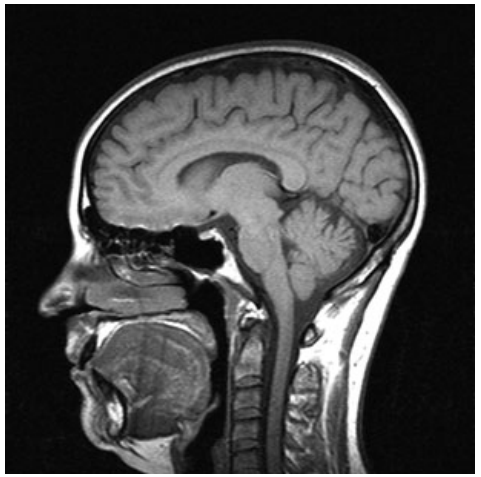
Mert R. Sabuncu
Athinoula A. Martinos Center for Biomedical Imaging

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) = Genotype (G) + Environment (E) + G×E
Heritability h^2: the proportion of the phenotypic variance in a trait attributable to the additive effects of genes

Measurement	Test-retest reliability	$\hat{h}^2$
intracranial volume (ICV)	0.995	0.849
total brain volume	0.997	0.981
left hemispheric cortical GM volume	0.992	0.521
right hemispheric cortical GM volume	0.991	0.492
total cortical GM volume	0.994	0.515
total subcortical GM volume	0.968	0.357
total GM volume	0.995	0.475
left hemispheric WM volume	0.996	0.416
right hemispheric WM volume	0.996	0.302
total WM volume	0.996	0.369
left hemispheric mean cortical thickness	0.899	0.688
right hemispheric mean cortical thickness	0.885	0.732
overall mean cortical thickness	0.935	0.734
left hemispheric total surface area	0.999	0.298
right hemispheric total surface area	0.997	0.288
total surface area	0.998	0.305

Neurological disorders are largely heritable

Condition	Heritability	Study
Schizophrenia	$> 80\%$	Cannon *et al.* 1998
Alzheimer's	$> 70\%$	Gatz *et al.* 1996
ALS	$> 60\%$	Al Chalabi *et al.* 2010
Alcoholism	59%	Kendler *et al.* 1995
Panic Disorder	44%	Kendler *et al.* 1995
Major Depression	41%	Kendler *et al.* 1995
Autism	37%	Hallmayer *et al.* 2011
Phobia	35%	Kendler *et al.* 1995
Anxiety Disorder	32%	Kendler *et al.* 1995
Bulimia	30%	Kendler *et al.* 1995

Many neurological disorders are polygenic

nature Vol 460|6 August 2009|doi:10.1038/nature08185

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

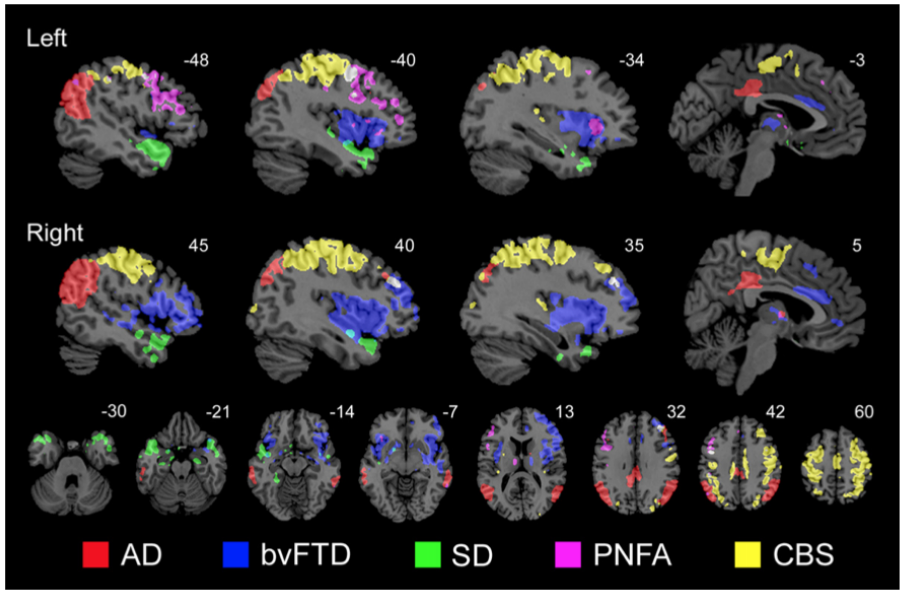
Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[1,2]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. 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[4]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at $P < 10^{-3}$, found on the ancestral European haplotype[5] (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

Purcell *et al.* 2009

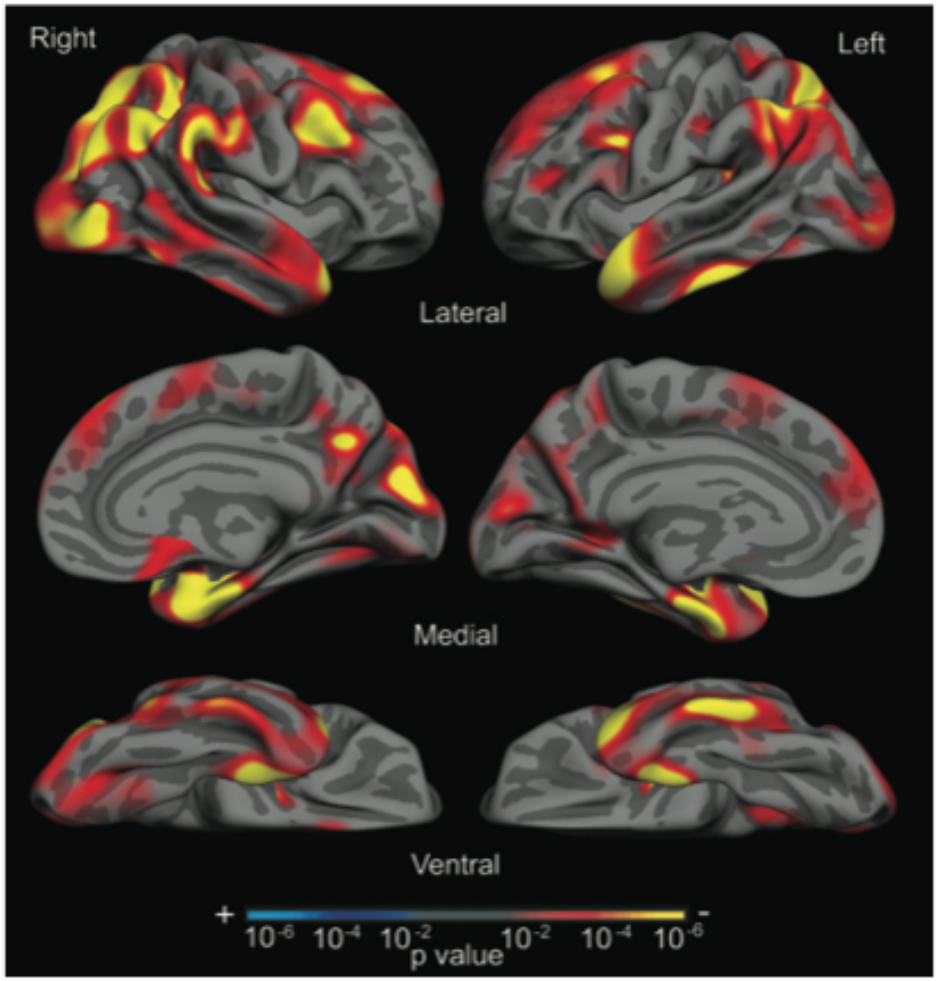
Neurological disorders have distributed footprints



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

Seeley *et al.* 2009

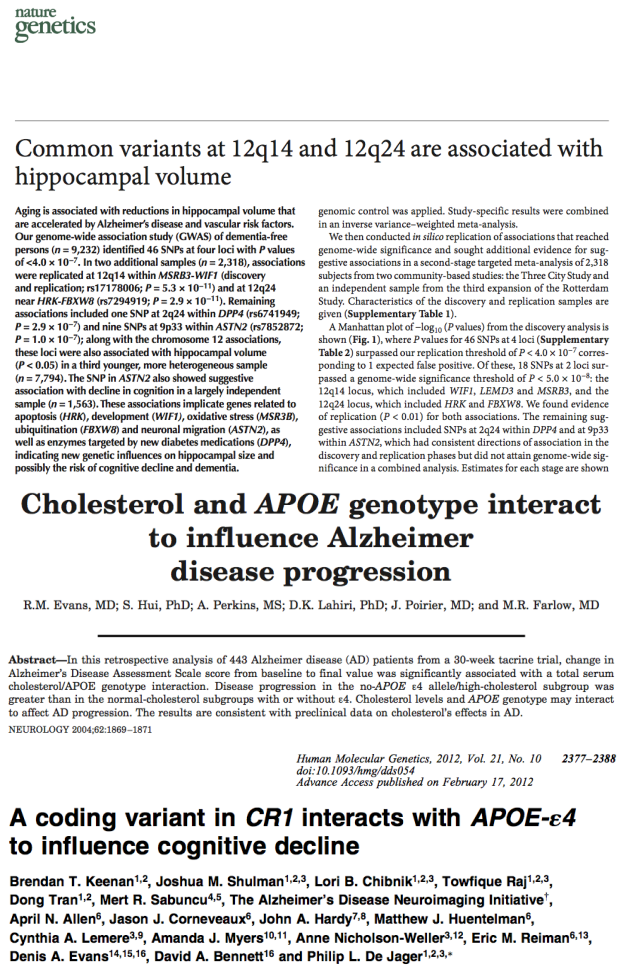
Cortical atrophy signature of Alzheimer's disease



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



nature genetics

Common variants at 12q14 and 12q24 are associated with hippocampal volume

Aging is associated with reductions in hippocampal volume that are accelerated by Alzheimer's disease and vascular risk factors. Our genome-wide association study (GWAS) of dementia-free persons ($n = 9,232$) identified 46 SNPs at four loci with P values of $<4.0 \times 10^{-7}$. In two additional samples ($n = 2,318$), associations were replicated at 12q14 within *MSRB3-WIF1* (discovery and replication; rs17178006; $P = 5.3 \times 10^{-11}$) and at 12q24 near *HRK-FBXW8* (rs7294919; $P = 2.9 \times 10^{-11}$). Remaining associations included one SNP at 2q24 within *DPP4* (rs6741949; $P = 2.9 \times 10^{-7}$) and nine SNPs at 9p33 within *ASTN2* (rs7852872; $P = 1.0 \times 10^{-7}$); along with the chromosome 12 associations, these loci were also associated with hippocampal volume ($P < 0.05$) in a third younger, more heterogeneous sample ($n = 7,794$). The SNP in *ASTN2* also showed suggestive association with decline in cognition in a largely independent sample ($n = 1,563$). These associations implicate genes related to apoptosis (*HRK*), development (*WIF1*), oxidative stress (*MSR3B*), ubiquitination (*FBXW8*) and neuronal migration (*ASTN2*), as well as enzymes targeted by new diabetes medications (*DPP4*), indicating new genetic influences on hippocampal size and possibly the risk of cognitive decline and dementia.

genomic control was applied. Study-specific results were combined in an inverse variance–weighted meta-analysis.

We then conducted *in silico* replication of associations that reached genome-wide significance and sought additional evidence for suggestive associations in a second-stage targeted meta-analysis of 2,318 subjects from two community-based studies: the Three City Study and an independent sample from the third expansion of the Rotterdam Study. Characteristics of the discovery and replication samples are given (**Supplementary Table 1**).

A Manhattan plot of $-\log_{10}$ (P values) from the discovery analysis is shown (**Fig. 1**), where P values for 46 SNPs at 4 loci (**Supplementary Table 2**) surpassed our replication threshold of $P < 4.0 \times 10^{-7}$ corresponding to 1 expected false positive. Of these, 18 SNPs at 2 loci surpassed a genome-wide significance threshold of $P < 5.0 \times 10^{-8}$: the 12q14 locus, which included *WIF1*, *LEMD3* and *MSRB3*, and the 12q24 locus, which included *HRK* and *FBXW8*. We found evidence of replication ($P < 0.01$) for both associations. The remaining suggestive associations included SNPs at 2q24 within *DPP4* and at 9p33 within *ASTN2*, which had consistent directions of association in the discovery and replication phases but did not attain genome-wide significance in a combined analysis. Estimates for each stage are shown

Cholesterol and *APOE* genotype interact to influence Alzheimer disease progression

R.M. Evans, MD; S. Hui, PhD; A. Perkins, MS; D.K. Lahiri, PhD; J. Poirier, MD; and M.R. Farlow, MD

Abstract—In this retrospective analysis of 443 Alzheimer disease (AD) patients from a 30-week tacrine trial, change in Alzheimer's Disease Assessment Scale score from baseline to final value was significantly associated with a total serum cholesterol/APOE genotype interaction. Disease progression in the no-*APOE* ε4 allele/high-cholesterol subgroup was greater than in the normal-cholesterol subgroups with or without ε4. Cholesterol levels and *APOE* genotype may interact to affect AD progression. The results are consistent with preclinical data on cholesterol's effects in AD.

NEUROLOGY 2004;62:1869–1871

Human Molecular Genetics, 2012, Vol. 21, No. 10 **2377–2388**
doi:10.1093/hmg/dds054
Advance Access published on February 17, 2012

A coding variant in *CR1* interacts with *APOE-ε4* to influence cognitive decline

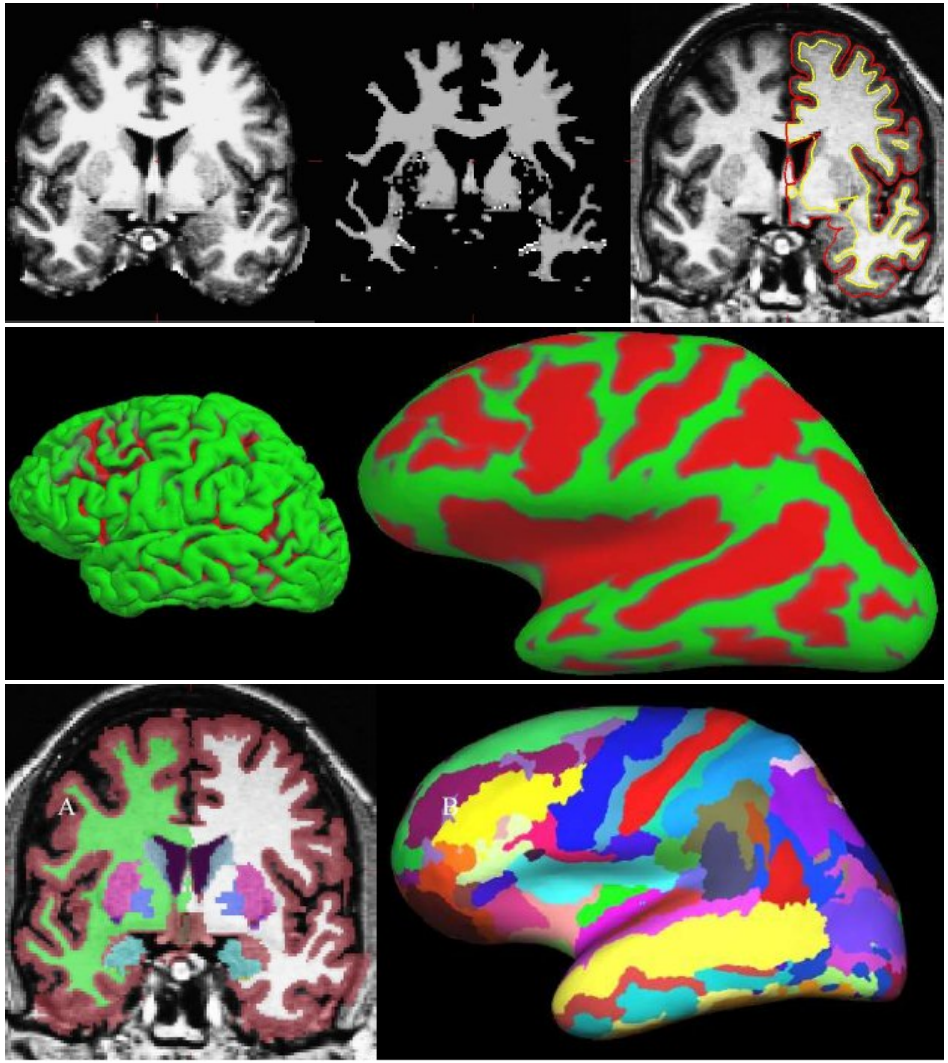
Brendan T. Keenan[1,2], Joshua M. Shulman[1,2,3], Lori B. Chibnik[1,2,3], Towfique Raj[1,2,3], Dong Tran[1,2], Mert R. Sabuncu[4,5], The Alzheimer's Disease Neuroimaging Initiative[†], April N. Allen[6], Jason J. Corneveaux[6], John A. Hardy[7,8], Matthew J. Huentelman[6], Cynthia A. Lemere[3,9], Amanda J. Myers[10,11], Anne Nicholson-Weller[3,12], Eric M. Reiman[6,13], Denis A. Evans[14,15,16], David A. Bennett[16] and Philip L. De Jager[1,2,3,*]

Processing sMRI scans

- 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

FreeSurfer: Overview



FreeSurfer: Morphometric Output

- 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

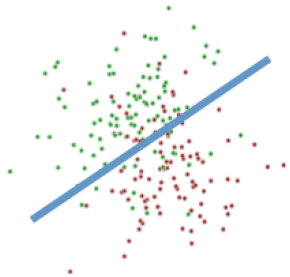
3 Approaches for Multivariate Neuroimaging Genetics

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

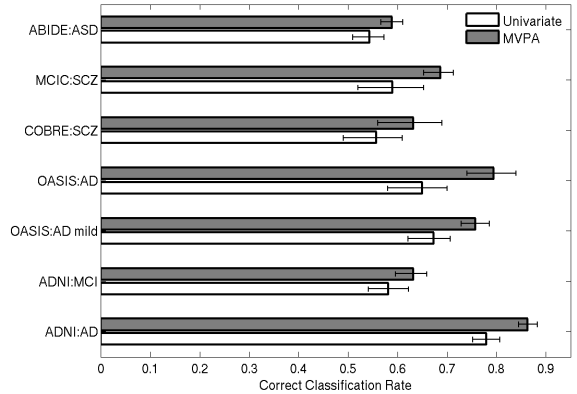
- Various classification/regression algorithms
  - Support Vector Machines, Random Forests, Relevance Vector/Voxel Machines
- Multivariate prediction is more accurate than univariate markers
- Detailed analysis of factors that influence prediction performance

Neuroinform
DOI 10.1007/s12021-014-9238-1

ORIGINAL ARTICLE

Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study

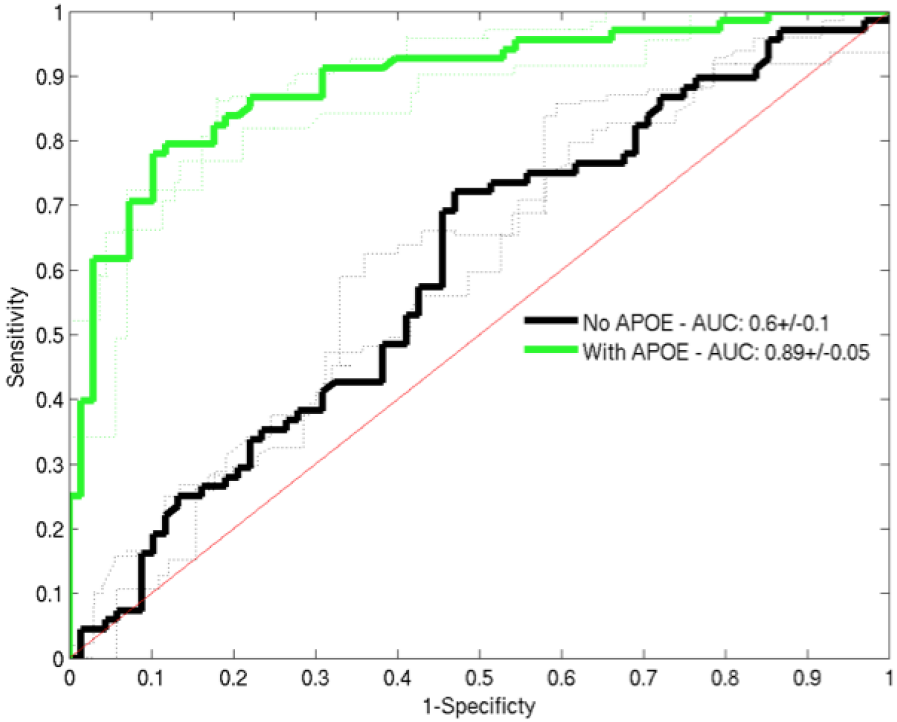
Mert R. Sabuncu · Ender Konukoglu ·



Sabuncu and Konukoglu 2014

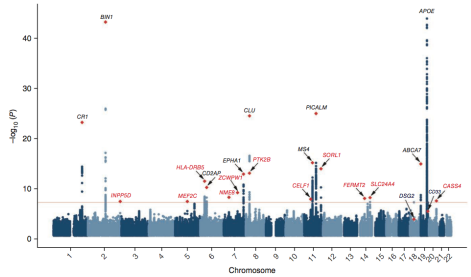
Polygenic Disease Risk

We employed Relevance Vector Machine to discriminate AD patients versus matched controls.



nature genetics

Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease



Association b/w Polygenic Risk and MRI Score in CN

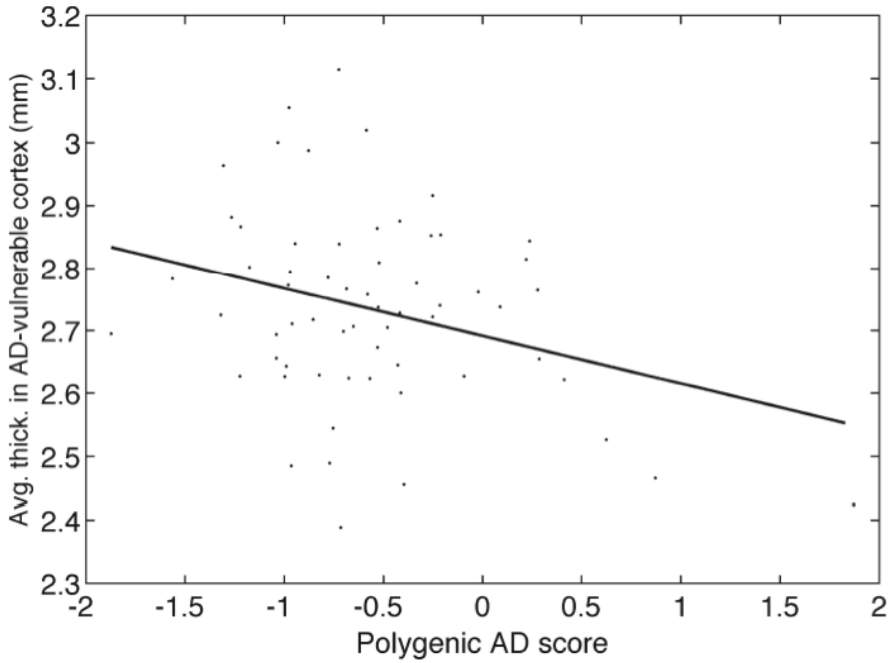


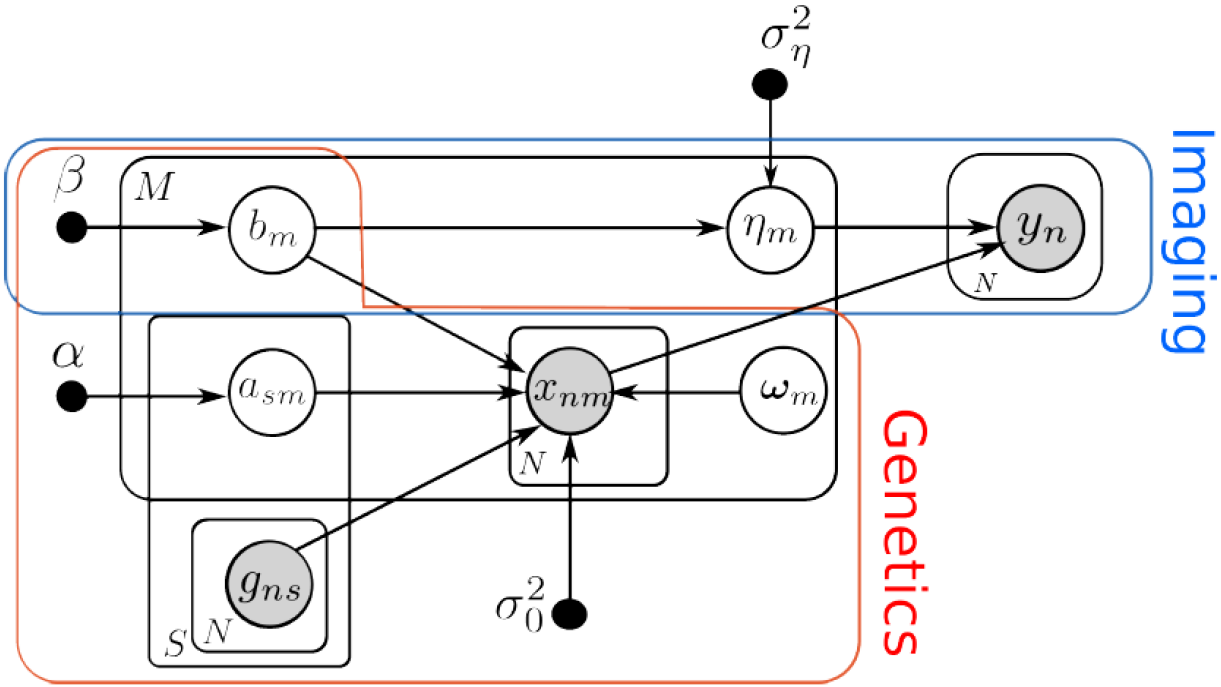
Figure 2. Thickness across AD-vulnerable cortex versus Alzheimer-associated non-APOE polygenic score in CN subjects with subthreshold levels of amyloid burden ($N = 64$, $\rho = -0.29$, $P < 0.05$).

Cerebral Cortex November 2012;22:2653-2661
doi:10.1093/cercor/bhr348
Advance Access publication December 13, 2011

The Association between a Polygenic Alzheimer Score and Cortical Thickness in Clinically Normal Subjects

Mert R. Sabuncu[1,2], Randy L. Buckner[1,3,4], Jordan W. Smoller[5], Phil Hyoun Lee[5], Bruce Fischl[1,2] and Reisa A. Sperling[1,6,7], for the Alzheimer's Disease Neuroimaging Initiative

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:
  - $y_i = \mathbf{X}_i\beta + h(\mathbf{G}_i) + \epsilon_i$
  - $\mathbf{G}_i = (g_{i1}, \cdots, g_{iL})$ - a genomic region with L genetic variants
  - $h(\cdot)$ - an arbitrary function located in a reproducing kernel Hilbert space $\mathcal{H}_\mathbf{K}$ defined by a nonnegative-definite kernel function matrix $\mathbf{K}$
  - $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$
  - Minimize the penalized square-error loss function w.r.t β, σ^2 and h:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N}\{y_i - \mathbf{X}_i\beta - h(\mathbf{G}_i)\}^2 + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_\mathbf{K}}$$

Liu et al., 2007

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)