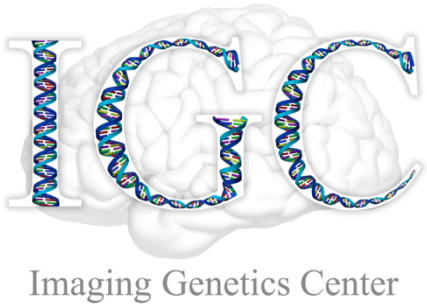
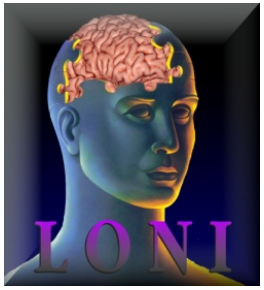
Voxelwise Gene-wide Association Methods and Multiple Testing

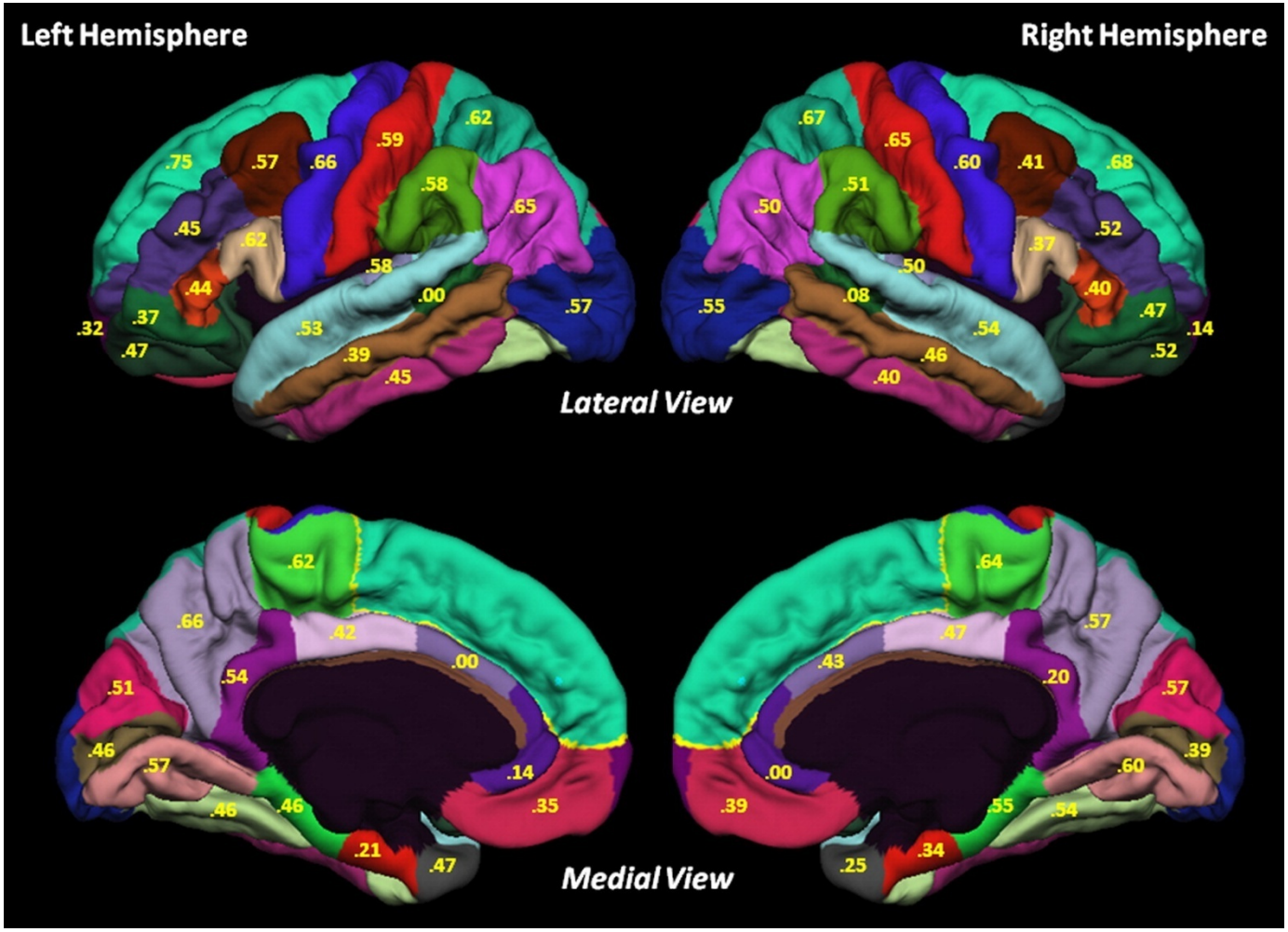
Derrek P. Hibar
derrek.hibar@loni.ucla.edu



UCLA



Brain structure is highly heritable



Must be specific genetic variants explaining the high heritability (most of which are unknown)

(Kremen et al., 2010)

The Endophenotype Approach

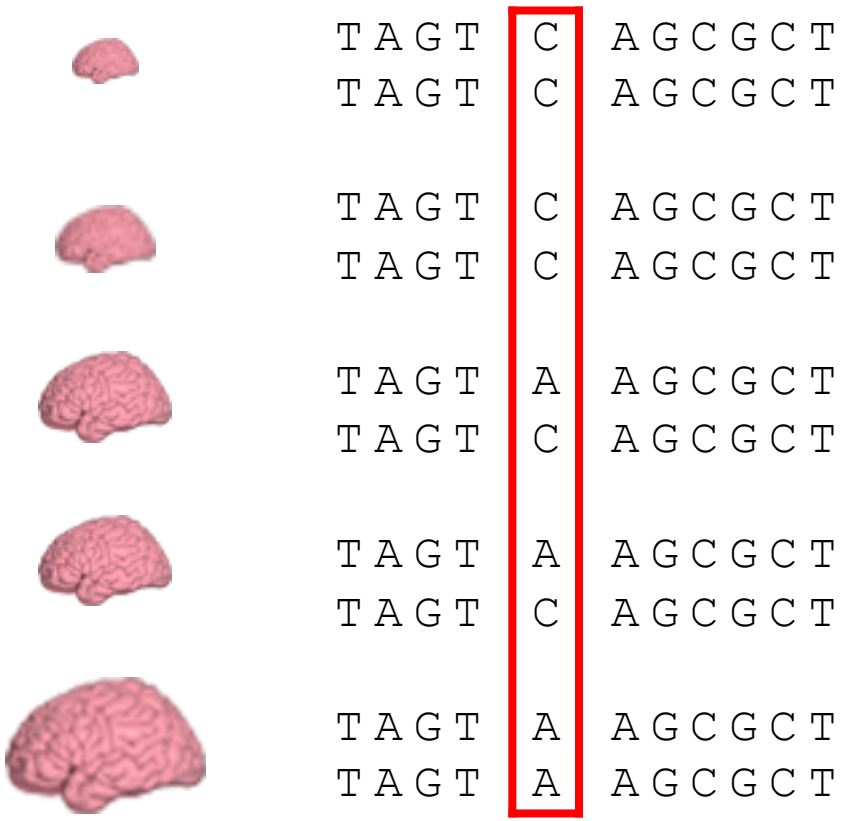
```
TAGT A AGCGCT
TAGT A AGCGCT
TAGT A AGCGCT
TAGT C AGCGCT
```

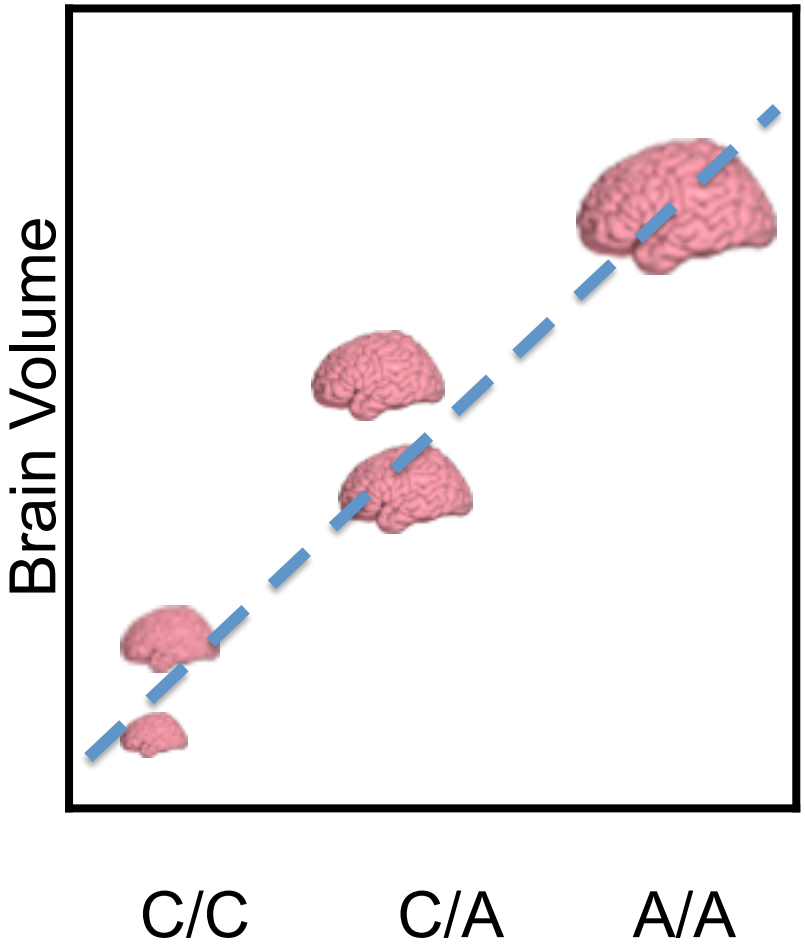
Schizophrenia

Ashley Egan 2012

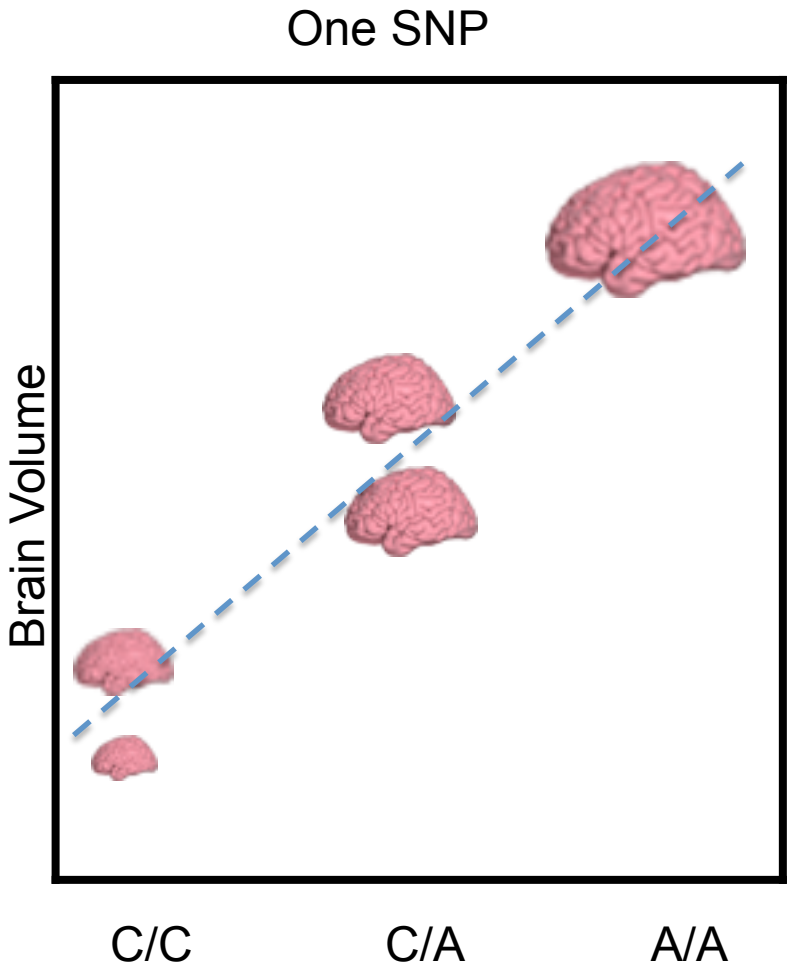
Genetic Variation (SNPs) → Endophenotype (Brain Structure) → Disease Status

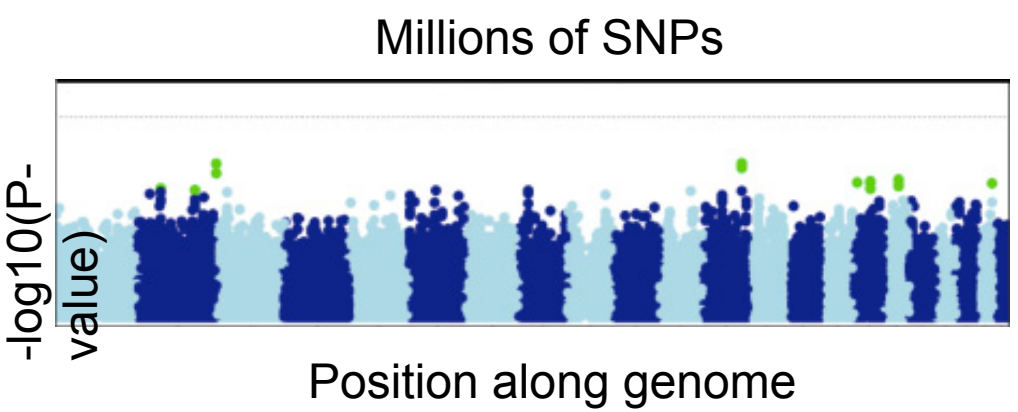
Finding Genetic Variants Influencing Brain Structure





Genome-wide association study

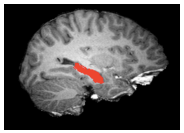
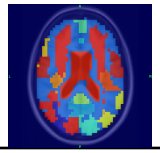
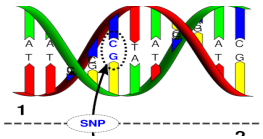
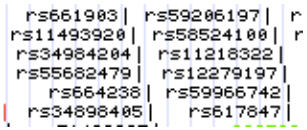




An unbiased search to find where in the genome a common variant is associated with a trait.

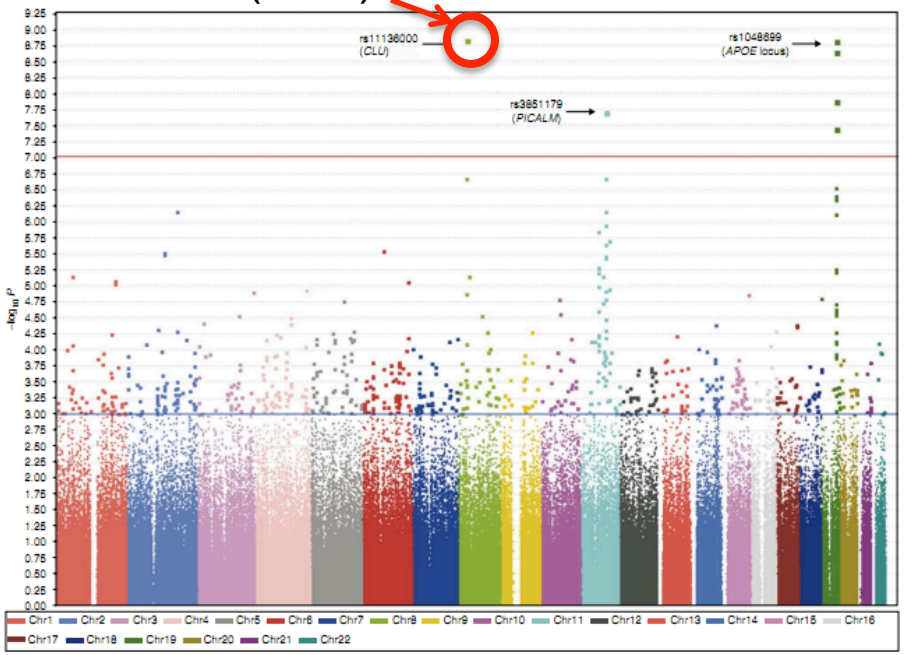
(adapted from Andy Saykin)

Imaging Genetics Menu

Genetics \ Imaging	Candidate ROI 	Many ROI 	Voxelwise
Candidate SNP 	Imager	Imager	Imager
Candidate Gene	Geneticist		
Genome-wide SNP 	Geneticist		
Genome-wide Gene	Geneticist		

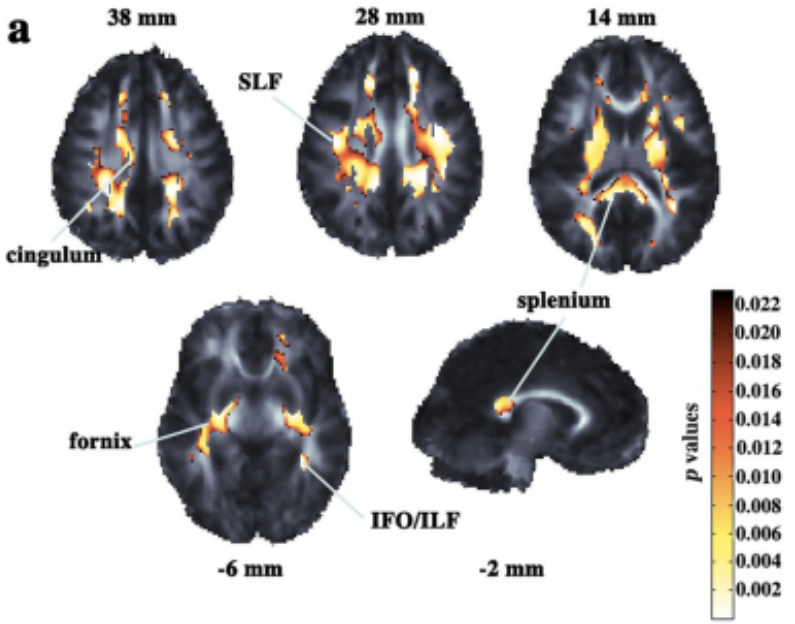
Characterizing the Effect of a Known Variant

rs11136000 (*CLU*)



Genome-wide association identifies variant within the CLU gene in ~4000 Alzheimer's patients and ~8000 controls – but what does it do?

(Harold et al., 2009)



The Alzheimer's associated variant broadly affects white matter integrity in a young cohort – may create early predisposition for disease

(Braskie et al., 2009)

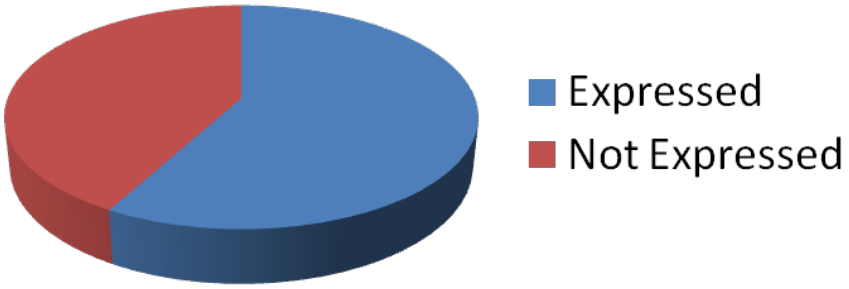
Advantages/Disadvantages

Advantages	Disadvantages
Candidate SNPs allow you to test a specific biological hypothesis	It is highly likely that we don't know the genetic underpinnings of a trait like brain structure so in general don't know the right SNP to pick
Strong hypothesis drives clearly interpretable results	In order to be widely accepted, the variant needs to have strong prior evidence (genome-wide significant in a meta-analysis or have clear function)
Multiple comparisons burden is reduced (one SNP – many voxels)	Unable to search the genome, only characterize the effect of a known variant
Quick way to provide functional relevance to unbiased genome-wide search results	

Low prior probability of any candidate to have effects on brain structure

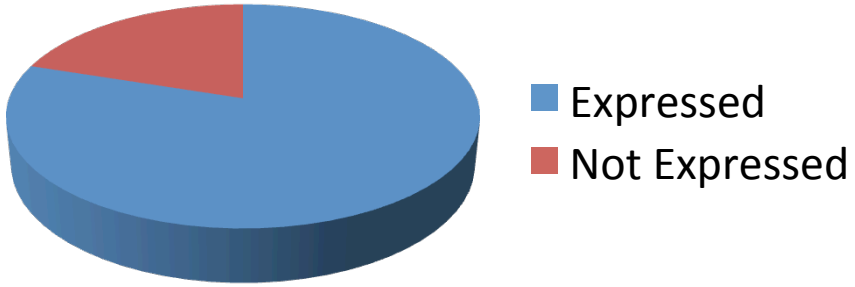
"Choosing candidate genes is generally on the basis of limited information and therefore excludes the vast majority of genes expressed in the central nervous system" (Glatt & Freimer, 2002)

Percentage of Genes Expressed in Human Cortex (Gene Chip)



(Myers et al., 2007)

Percentage of Genes Expressed in Mouse Brain (ISH)



(Lein et al., 2007)

We generally don't know the theoretical genetic underpinnings of a phenotype

(Freimer & Sabatti, 2004)

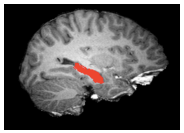
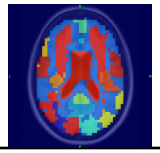
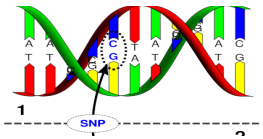
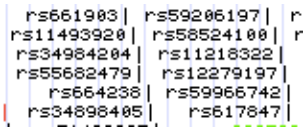
Candidate Genes -- Lack of Replication in ENIGMA

Gene	SNP (proxy)	A1	A2	Freq	Beta	SE	P-value
Full Discovery sample – including patients							
BDNF	rs6265	T	C	0.191	-0.280	7.193	0.969
TOMM40	rs2075650	A	G	0.832	17.657	8.348	**0.034**
CLU	rs11136000	T	C	0.399	6.173	5.791	0.287
PICALM	rs3851179	T	C	0.361	10.192	5.803	0.079
ZNF804A	rs1344706	A	C	0.584	5.555	5.639	0.325
COMT	rs4680	A	G	0.532	-7.045	5.637	0.211
DISC1	rs821616 (rs1754606 r^2=1.00)	T	C	0.718	0.465	6.204	0.940
NRG1	rs35753505 (rs12681411 r^2=0.835)	C	G	0.638	-2.870	6.065	0.636
DTNBP1	rs1011313	T	C	0.093	-7.900	9.703	0.416
DTNBP1	rs1018381 (rs875463 r^2=1.00)	A	C	0.080	-1.491	10.054	0.882

Stein et al., 2012; Nature Genetics

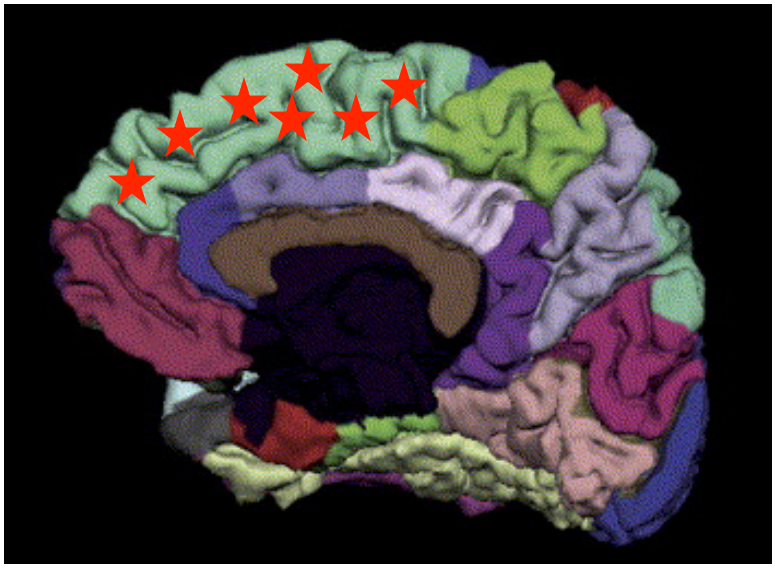
(adapted from Andy Saykin)

Imaging Genetics Menu

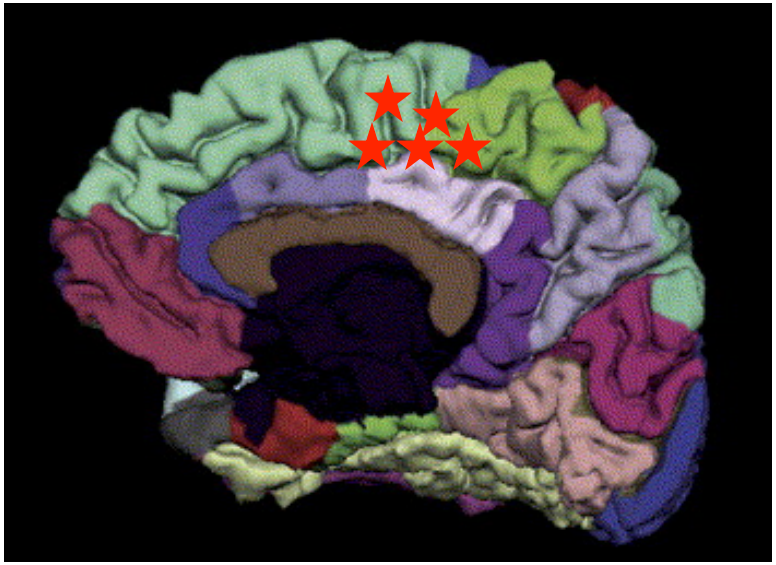
Genetics \ Imaging	Candidate ROI 	Many ROI 	Voxelwise
Candidate SNP 	Imager	Imager	Imager
Candidate Gene	Geneticist		
Genome-wide SNP 	Geneticist		
Genome-wide Gene	Geneticist		

Voxelwise vs. ROI approach

Dependent on geometry of the signal



★ = signal



Signal overlaps with ROI definition
ROI more powerful

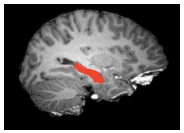
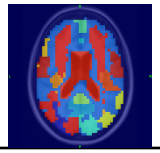
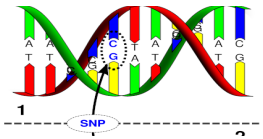
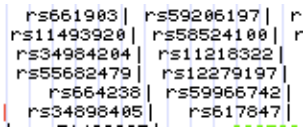
Signal does not overlap with ROI definition
Voxelwise more powerful

In search for genetic effects on brain structure –
we generally are not clear where they are

(Desikan et al., 2006)

(adapted from Andy Saykin)

Imaging Genetics Menu

Genetics \ Imaging	Candidate ROI 	Many ROI 	Voxelwise
Candidate SNP 	Imager	Imager	Imager
Candidate Gene	Geneticist		
Genome-wide SNP 	Geneticist		
Genome-wide Gene	Geneticist		

vGWAS

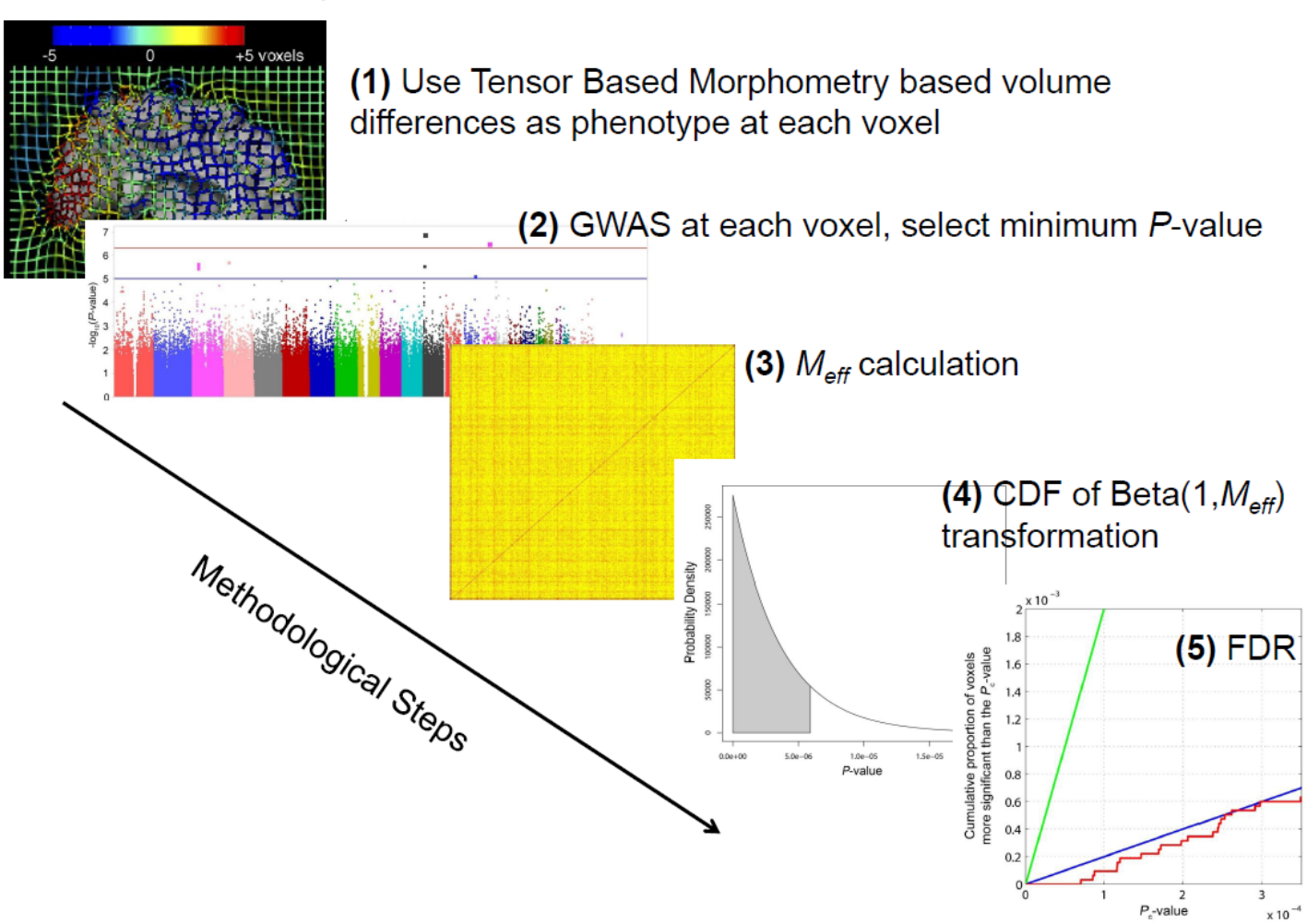
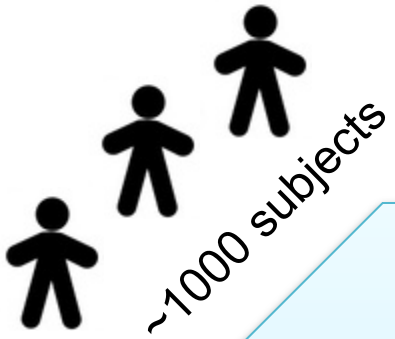


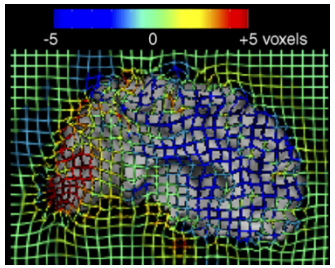
Figure 1: Summary of vGWAS method for conducting association and assessing statistical significance.

(Stein et al., 2010)

Multiple Testing Problem



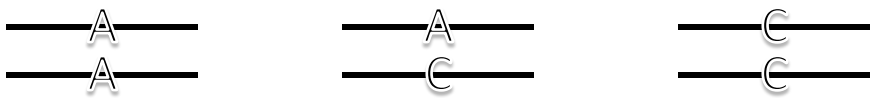
~1000 subjects



~30,000 voxels in the brain

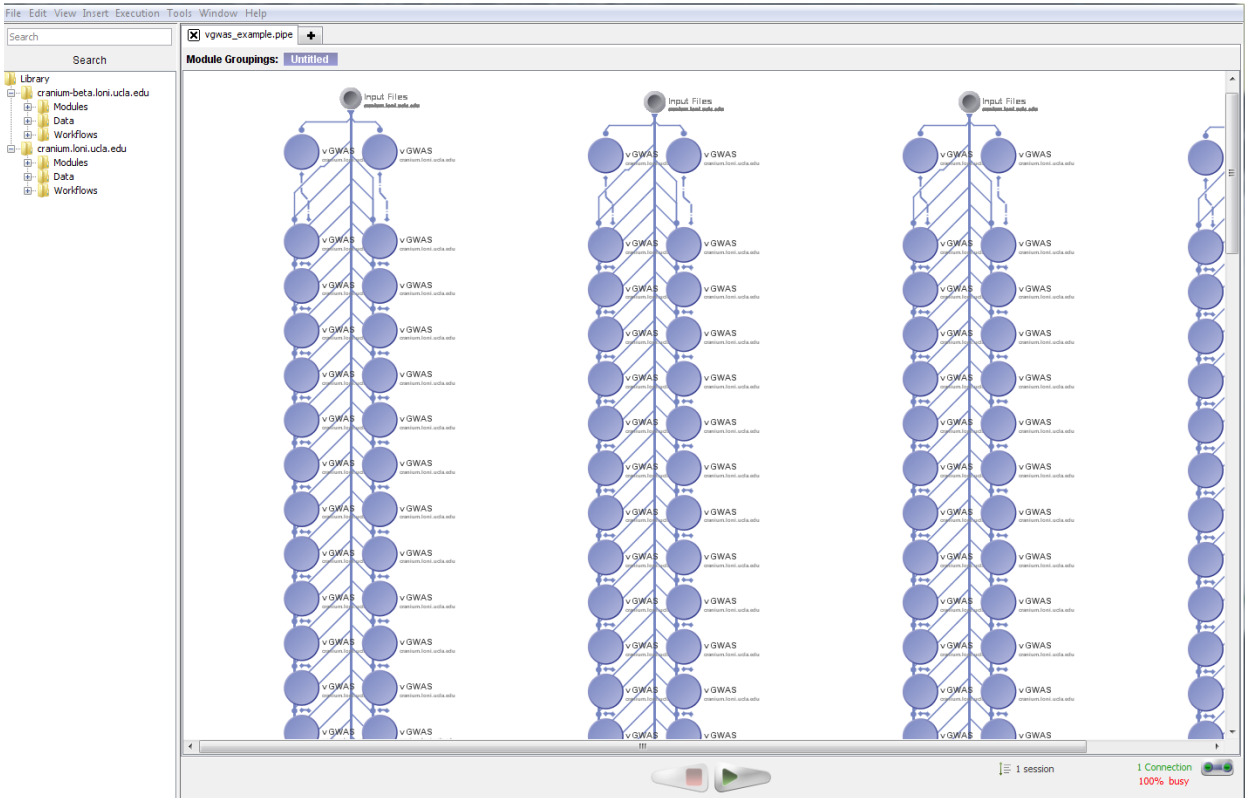
1.8 x 10^{10} tests!

~600,000 genetic markers (SNPs)



Computationally Intensive: GWAS on each voxel

Genome-wide association on each phenotype takes ~9 minutes / phenotype.
31,622 voxels * 9 minutes = 198 days of computation!
Across 300 nodes total computation time is 27h.

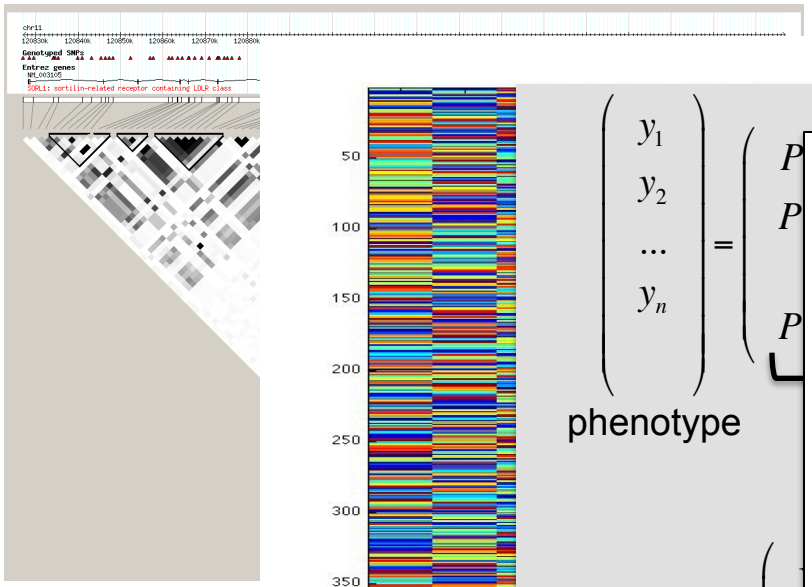


http://pipeline.loni.ucla.edu/

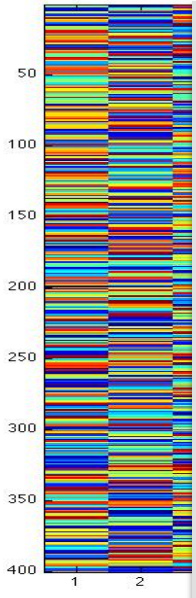
Gene-based association tests

- VEGAS (Liu et al., 2010)
- SIMES-GATES (built-in to KGG)
- Lasso regression
- Ridge regression
- Elastic net (Kohannim et al., 2012; ISBI)
- Principal components regression (PCReg)
- Many more…

Gene-based association test



Find all the
th



Conduct P
compone
v

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} P \\ P \\ \\ P \end{pmatrix}$$

phenotype

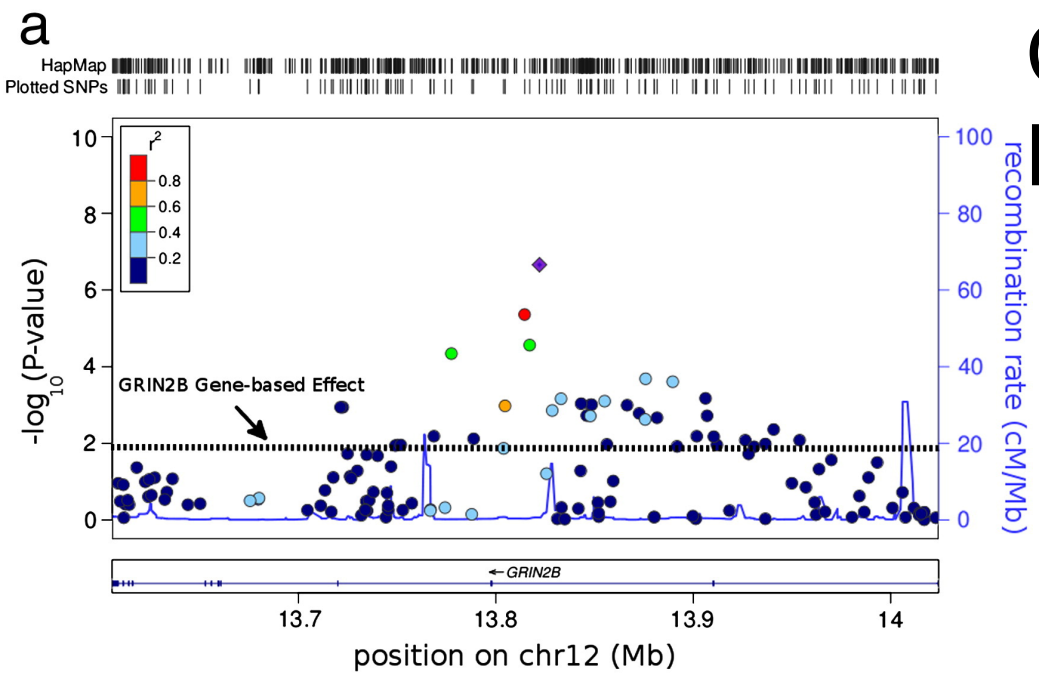
$$F_{k,df(\text{full})} = \frac{RSS(\text{reduced}) - RSS(\text{full})}{df(\text{reduced}) - df(\text{full})} \Big/ \frac{RSS(\text{full})}{df(\text{full})}$$

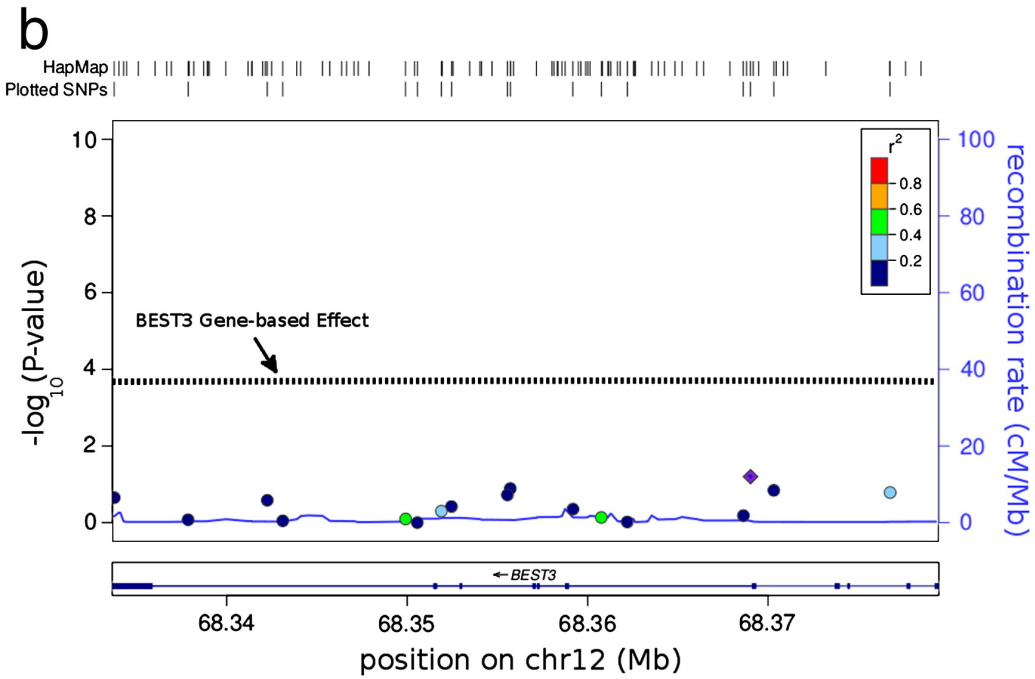
A partial F-test is used to test the joint effect of the SNP PCs statistically controlling for the effects in the reduced model.

Get one P-value per gene

Principal Component Regression

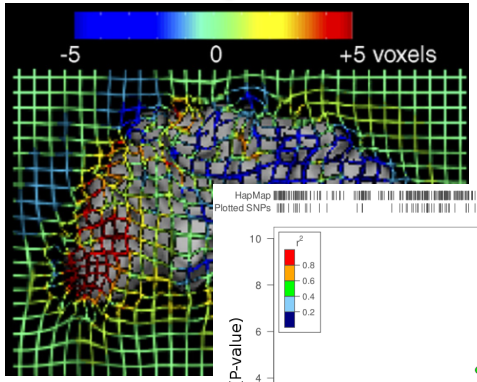
Comparison of PCReg and MLR



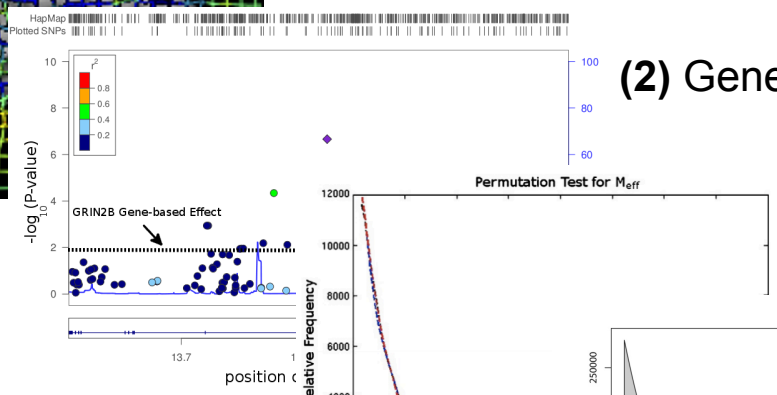


(Hibar et al., 2011)

vGeneWAS

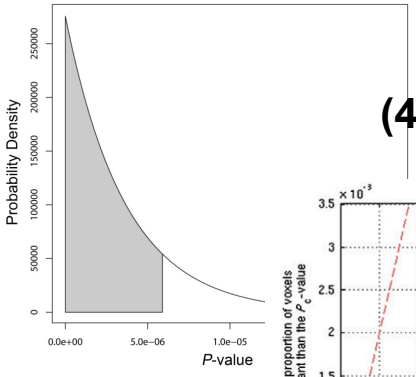


(1) Use Tensor Based Morphometry based volume differences as phenotype at each voxel

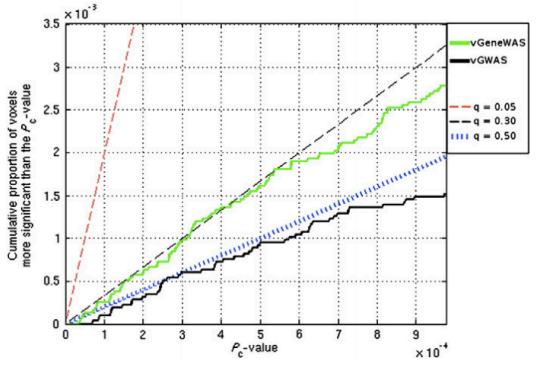


(2) GeneWAS at each voxel, select minimum *P*-value

(3) M_{eff} calculation through permutation and then estimating beta parameter of the Beta distribution



(4) CDF of Beta(1,M_{eff}) transformation



(5) FDR

Methodological Steps

(Hibar et al., 2011)

ADNI Dataset

Subjects

731 Caucasian subjects to avoid population stratification

Diagnosis:

- 172 Alzheimer's disease pat's
- 356 Mild Cognitive Impairment
- 203 healthy elderly

Demographics:

- 75.56 +/- 6.82 years
- 430 males

Genetics

Illumina 610-Quad BeadChip

Exclusions:

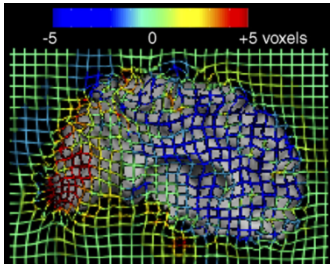
- genotype call rate < 95%,
- deviation from Hardy-Weinberg equilibrium $P<5.7x10^{-7}$
- minor allele frequency < 0.10

448,293 SNPs

- Manual SNP annotation into gene groups using the PLINK web interface.

18,044 genes in analysis

Imaging Phenotype

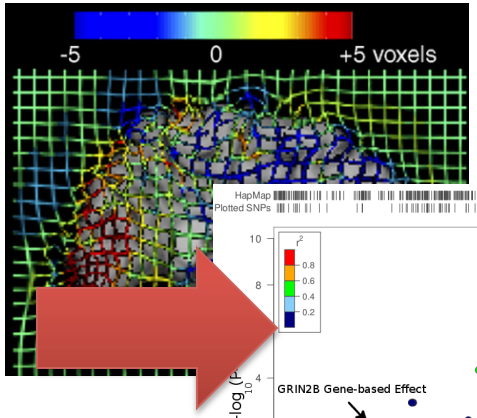


Tensor Based Morphometry

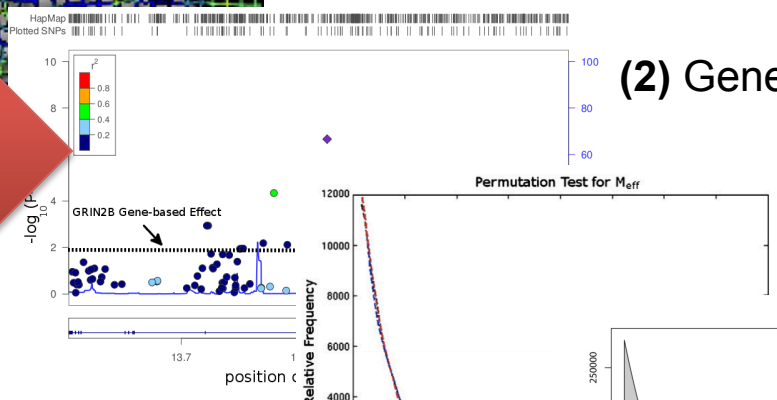
Each voxel encodes volume change relative to a study-specific template

31,622 voxels in the brain when downsampled to 4x4x4 mm^3 voxels

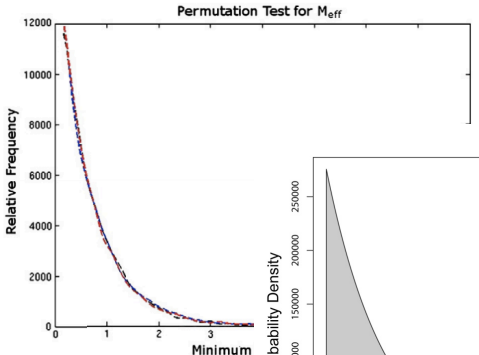
vGeneWAS



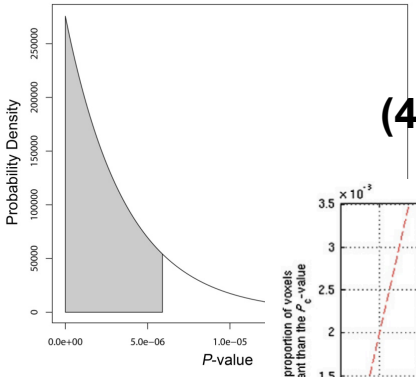
(1) Use Tensor Based Morphometry based volume differences as phenotype at each voxel



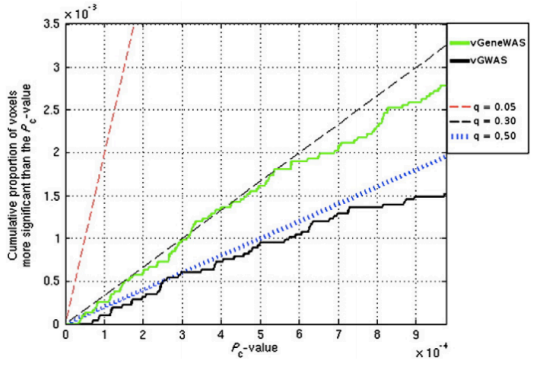
(2) GeneWAS at each voxel, select minimum *P*-value



(3) M_{eff} calculation through permutation and then estimating beta parameter of the Beta distribution



(4) CDF of Beta(1,M_{eff}) transformation

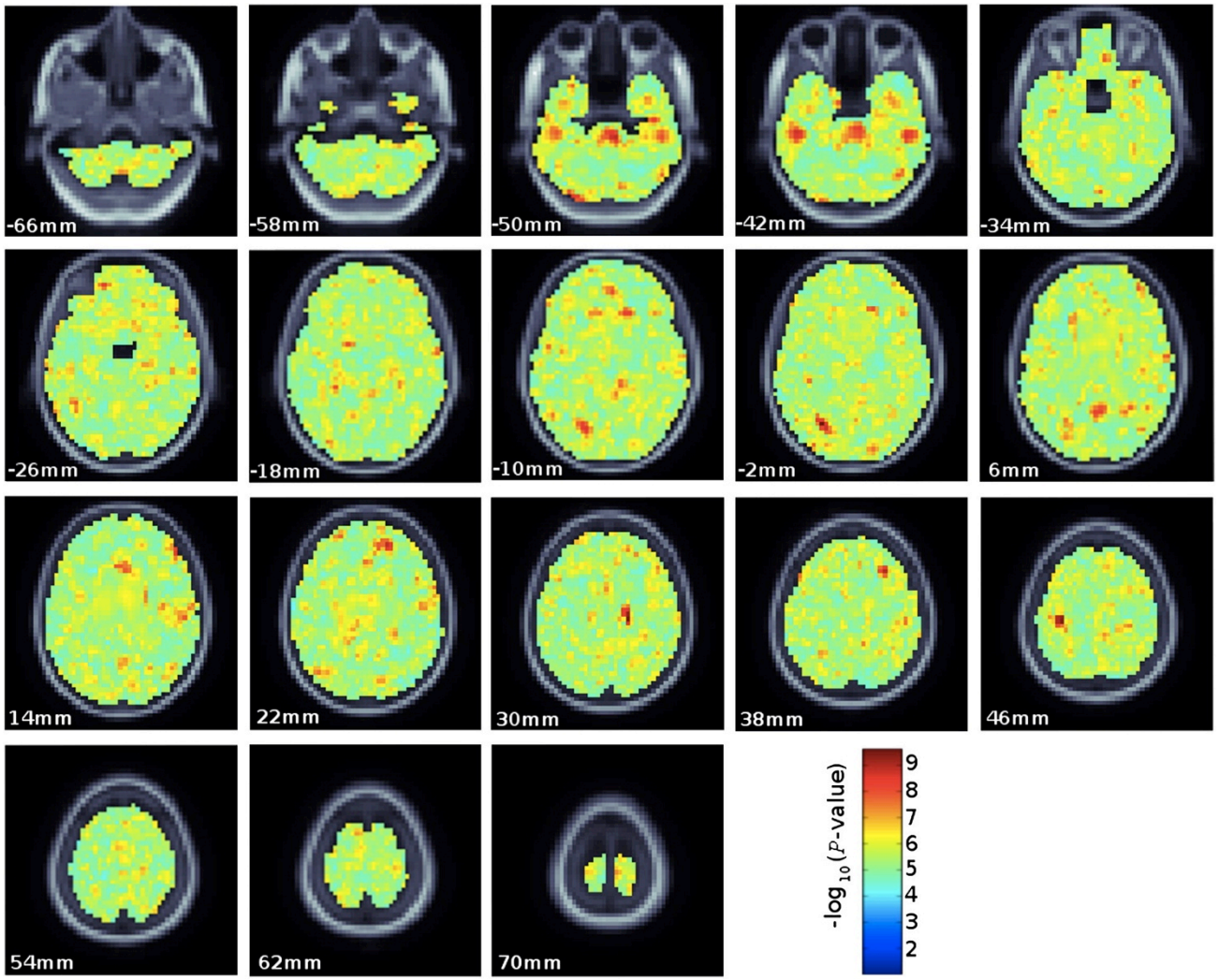


(5) FDR

Methodological Steps

(Hibar et al., 2011)

Raw minimum *P*-value at each voxel



(Hibar et al., 2011)

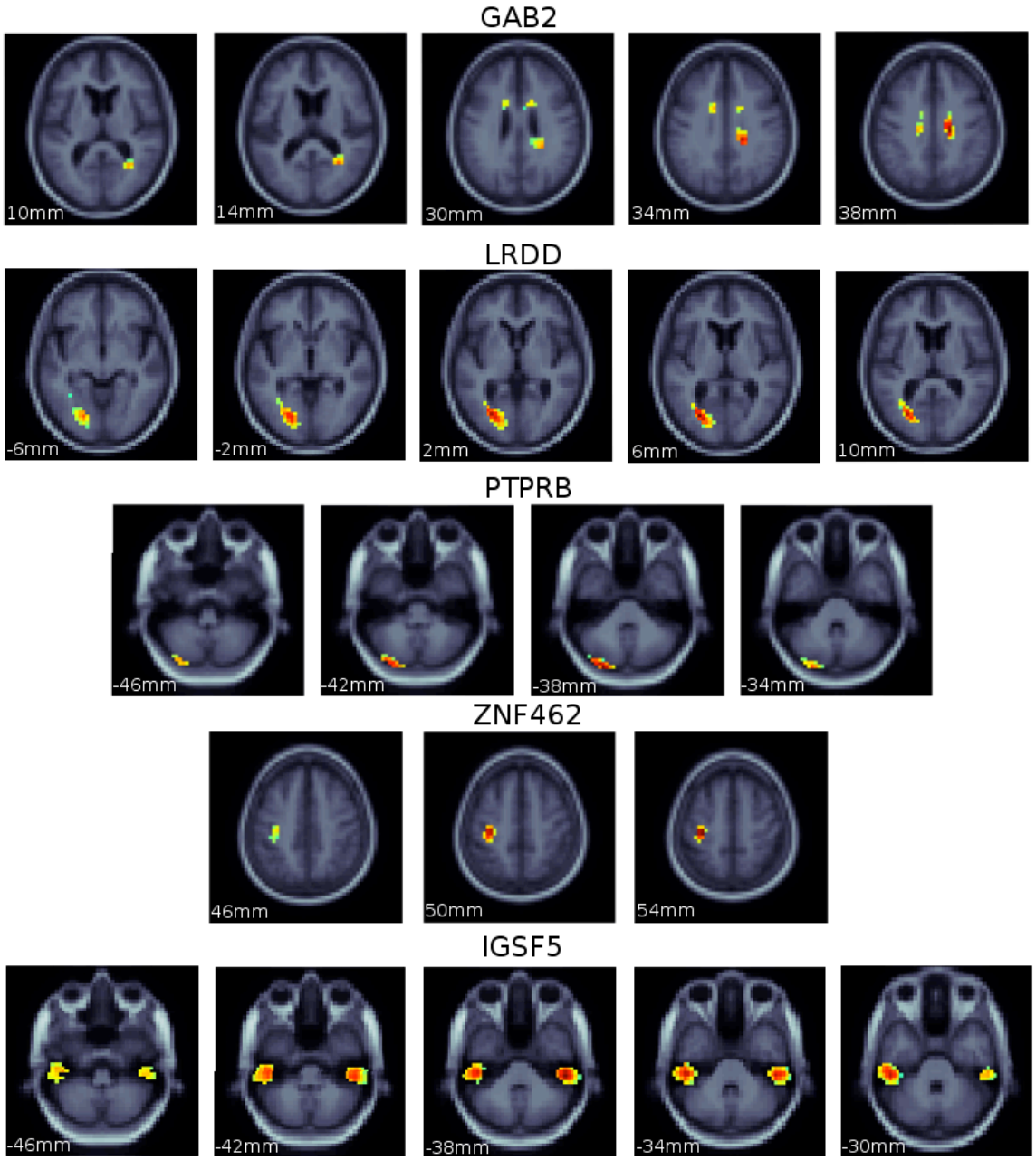
Most associated genes

Chr	Gene	# of SNPs in gene	# of eigenSNPs	Minimum P-value	Mean P-value	Volume (mm3)	Proportion of brain volume	Clustermax (mm3)	# of clusters
11	GAB2	20	10	2.36 × 10− 9	1.50 × 10− 5	6336	0.0049	2688	9
11	LRDD	2	2	2.60 × 10− 9	1.32 × 10− 5	8128	0.0063	7872	4
12	PTPRB	17	13	2.84 × 10− 9	1.81 × 10− 5	3200	0.0024	3008	5
9	ZNF462	9	6	3.29 × 10− 9	1.84 × 10− 5	2688	0.0021	2688	1
21	IGSF5	27	14	5.32 × 10− 9	1.62 × 10− 5	16,384	0.013	9344	3
2	SLC25A12	10	5	9.48 × 10− 9	2.66 × 10− 5	1856	0.0014	1792	2
11	MRE11A	9	6	9.86 × 10− 9	8.80 × 10− 6	9344	0.0072	9216	3
19	SLC8A2	11	7	1.06 × 10− 8	3.18 × 10− 5	5632	0.0043	5376	3
15	CHRM5	3			1.77 × 10− 5	1280	0.00099	1216	2
18	SPIRE1				× 10− 5	6016	0.0046	3072	12
3	C3orf64				× 10− 5	4352	0.0034	2112	4
21	S100B				× 10− 5	9344	0.0072	6656	7
1	CRCT1				× 10− 5	4096	0.0032	3456	4
19	ZNF626				2.12 × 10− 5	2560	0.002	2112	3
1	ELK4	1		.05 × 10− 8	3.27 × 10− 5	4032	0.0031	2688	4
11	RSF1	8	6	9.30 × 10− 8	1.31 × 10− 5	768	0.00059	768	1
20	WFDC11	2	2	1.06 × 10− 7	2.49 × 10− 5	1280	0.00099	512	5
6	SCML4	27	18	1.07 × 10− 7	1.67 × 10− 5	3328	0.0026	3328	1
12	ERP27	8	14	1.08 × 10− 7	2.61 × 10− 5	2624	0.002	2176	2

AD Risk Gene (Reiman et al., 2007)

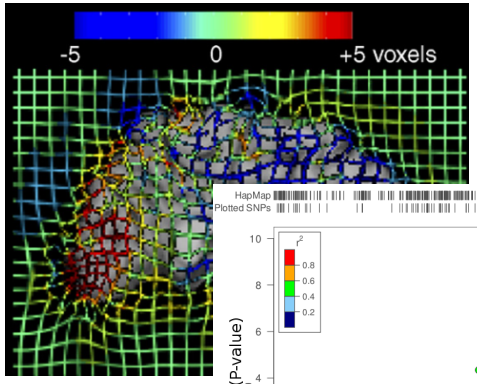
(Hibar et al., 2011)

Most associated voxels for most associated genes

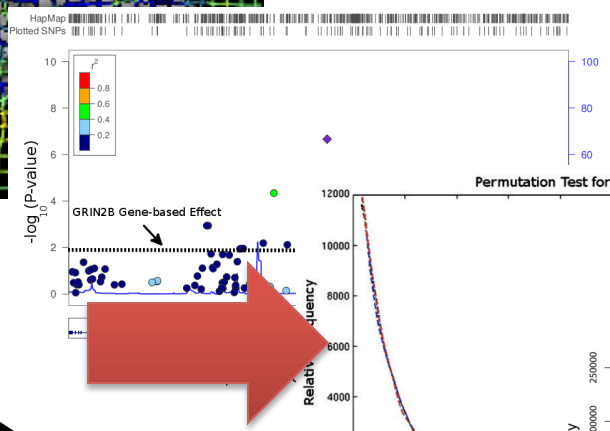


(Hibar et al., 2011)

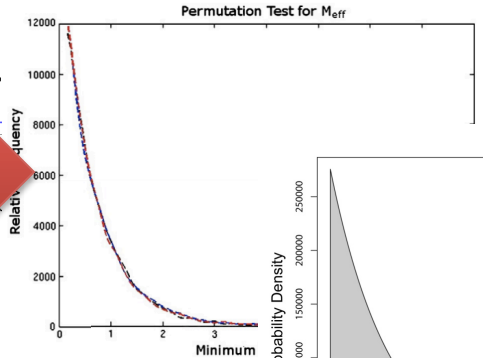
vGeneWAS



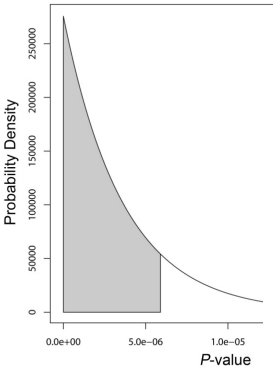
(1) Use Tensor Based Morphometry based volume differences as phenotype at each voxel



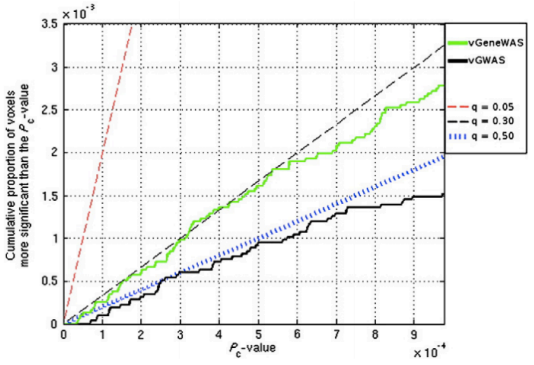
(2) GeneWAS at each voxel, select minimum *P*-value



(3) M_{eff} calculation through permutation and then estimating beta parameter of the Beta distribution



(4) CDF of Beta(1,M_{eff}) transformation

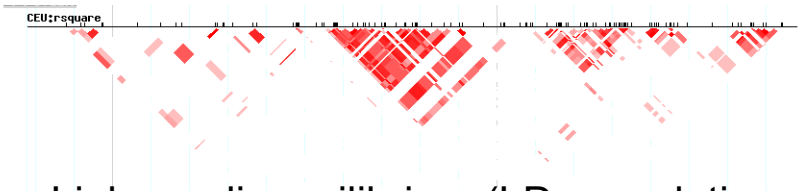


(5) FDR

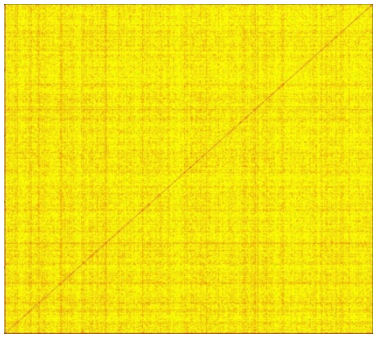
Methodological Steps

(Hibar et al., 2011)

Multiple Comparisons: Correlation of Genetic Markers



Linkage disequilibrium (LD; correlation between genetic markers) means that all tests are **not** independent.



simpleM: a method to determine the effective number of tests conducted M_{eff} where $M_{eff} \leq M$

1. Create correlation matrix
2. Calculate eigenvalues through PCA
3. Number of principal components which jointly explain 99.5% of variance = M_{eff}

TABLE I. Derived significance thresholds for the Illumina 1M data

		Permutation		SimpleM	
Chromosome	Number of SNPs	M_{eff}	α_p	M_{eff}	α_p
1	63,560	30,881	1.66e−06	31,898	1.61e−06
2	64,367	31,760	1.62e−06	33,386	1.54e−06
3	53,622	26,328	1.95e−06	27,675	1.85e−06
4	47,096	24,292	2.11e−06	24,474	2.10e−06
5	48,758	24,187	2.12e−06	25,229	2.03e−06
6	52,870	25,667	2.00e−06	26,165	1.96e−06
7	43,084	23,040	2.23e−06	23,123	2.22e−06
8	41,932	23,254	2.21e−06	22,314	2.30e−06
9	35,591	20,078	2.55e−06	20,061	2.56e−06
10	40,195	21,186	2.42e−06	21,277	2.41e−06
11	39,420	21,101	2.43e−06	20,359	2.52e−06
12	38,787	20,875	2.46e−06	21,318	2.41e−06
13	28,656	14,936	3.43e−06	15,087	3.40e−06
14	26,026	14,414	3.56e−06	13,829	3.71e−06
15	23,896	12,277	4.18e−06	13,312	3.85e−06
16	24,517	14,346	3.58e−06	14,626	3.51e−06
17	23,054	11,558	4.44e−06	12,578	4.08e−06
18	22,885	11,629	4.41e−06	12,583	4.08e−06
19	16,918	9,152	5.60e−06	9,646	5.32e−06
20	19,995	10,177	5.04e−06	11,169	4.59e−06
21	11,012	6,150	8.34e−06	6,247	8.21e−06
22	12,388	6,708	7.65e−06	7,004	7.32e−06
Genome-wide	778,629	378,368	1.36e−07	413,360	1.24e−07

Similar to permutation derived values

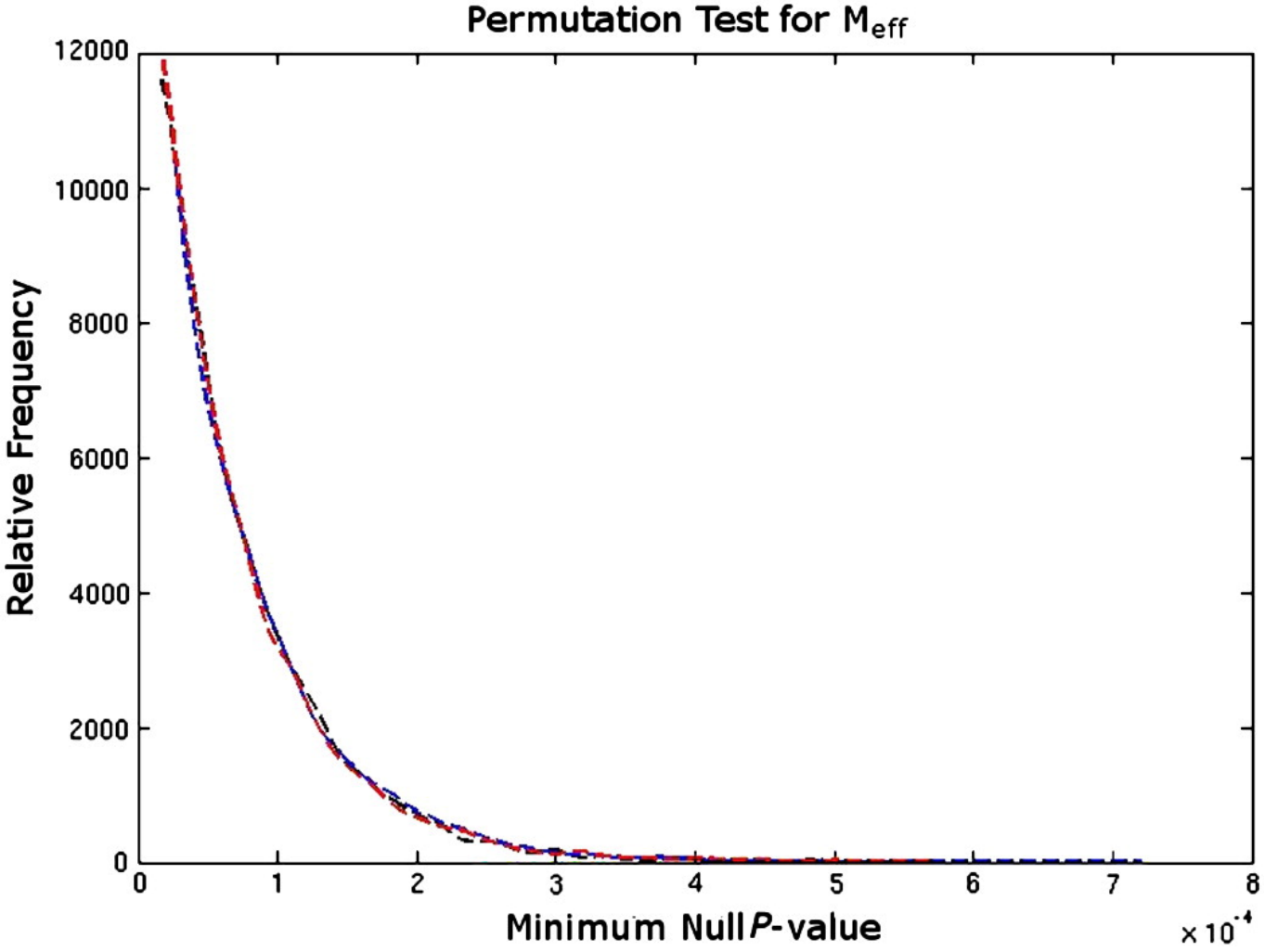
Correct *P*-values through a Beta(1, M_{eff}) distribution

(Gao et al., 2008; Gao et al., 2010)

Permutation Procedure

- Select a set of uncorrelated voxels
- Collect residuals of Pheno ~ Age + Sex
- Permute the residuals and then perform a gene-wide scan using PCReg.
- Store the p-value of the most associated gene
- Repeat permutation + gene-was (x5000)
- Null distribution of p-values

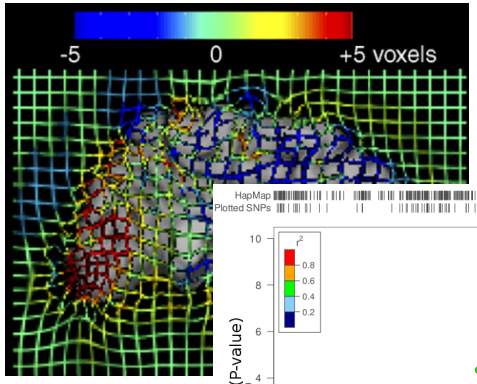
Permutation Results



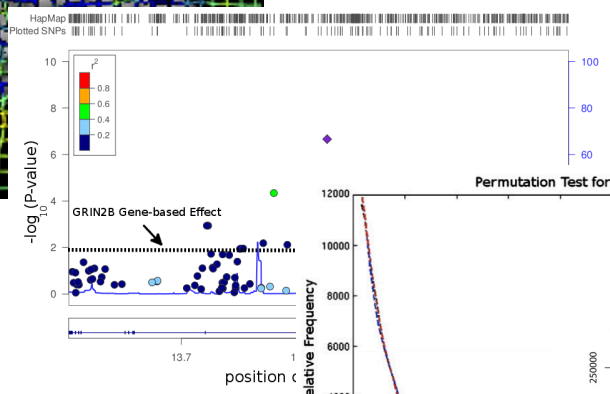
Effective number of tests M_{eff}

- The number of independent tests in this case follows (Ewens and Grant, 2001):
  - $f_{\min}(x) = n(1-x)^{n-1}$
- This is a Beta(a, b) distribution where $a = 1$ and $b = n$; where n is the number of independent genes tested
- Estimate b using a modified version of *betafit* that fixes $a = 1$ before estimating b

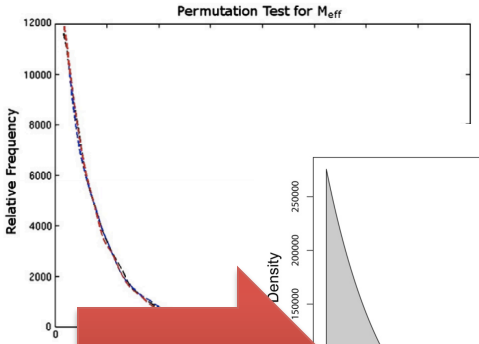
vGeneWAS



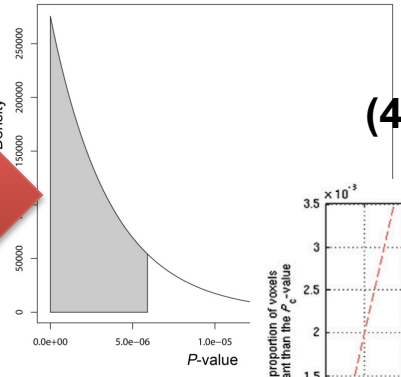
(1) Use Tensor Based Morphometry based volume differences as phenotype at each voxel



(2) GeneWAS at each voxel, select minimum *P*-value

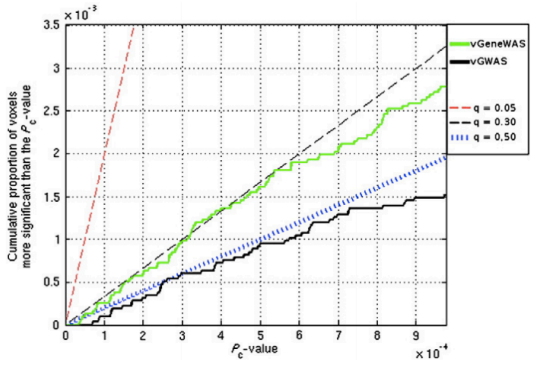


(3) M_{eff} calculation through permutation and then estimating beta parameter of the Beta distribution



(4) CDF of Beta(1,M_{eff}) transformation

Methodological Steps



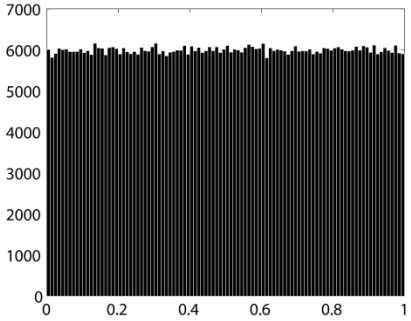
(5) FDR

(Hibar et al., 2011)

Multiple Comparisons Example

Error: Not accounting for multiple comparisons

Null *P*-values: Uniform Distribution

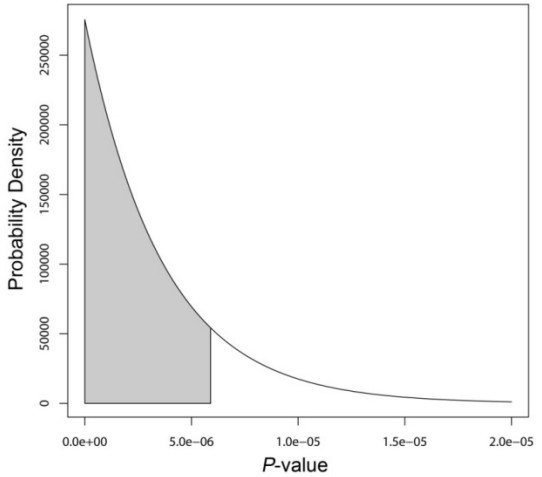


600,000 draws from a uniform distribution (Like GWAS on one phenotype). Minimum P-value gives very low P-values (1.7057×10^{-6} , 1.1026×10^{-6})

Wow, I have such low P-values! But all of this is randomness (simulated from null distributions)

Accounting for multiple comparisons assuming independence: Beta Distribution

Beta(1,600000) distribution



Models the multiple comparisons by picking the minimum *P*-value after 600,000 draws from a uniform distribution.

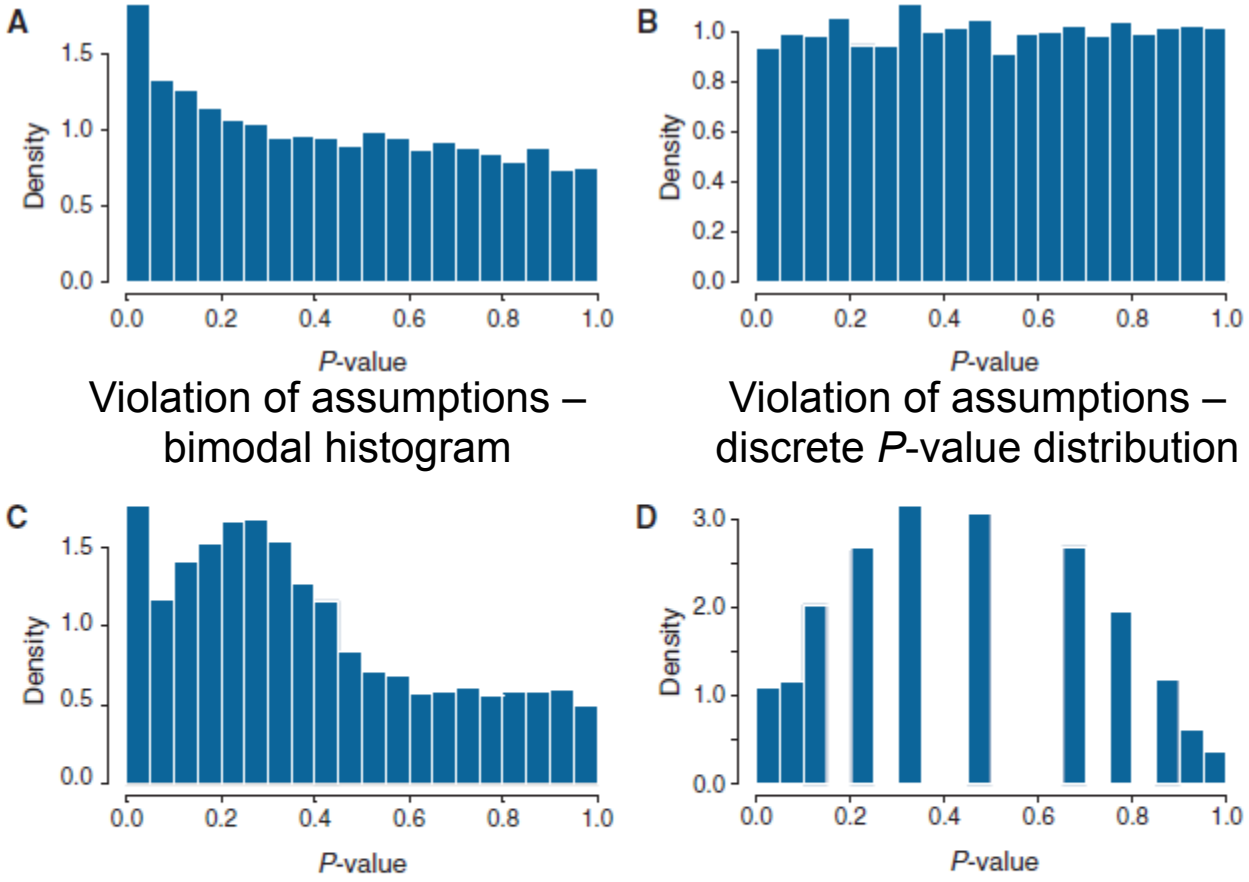
Adjustment through CDF of *Beta*(1,600000) gives corrected P-values

Raw P-value	**Corrected P-value**
1.7057×10^{-6}	0.646
1.1026×10^{-6}	0.484

Histogram visualization

FDR significant – over-representation of low *P*-values

Null *P*-values – no violations of assumptions



Violation of assumptions – bimodal histogram

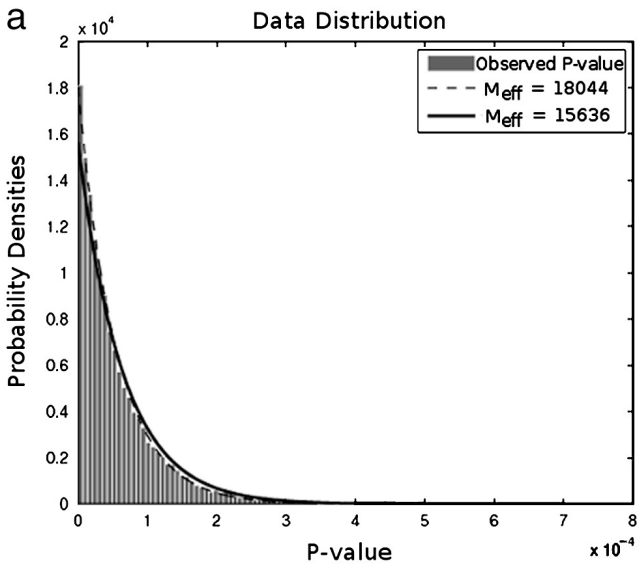
Violation of assumptions – discrete *P*-value distribution

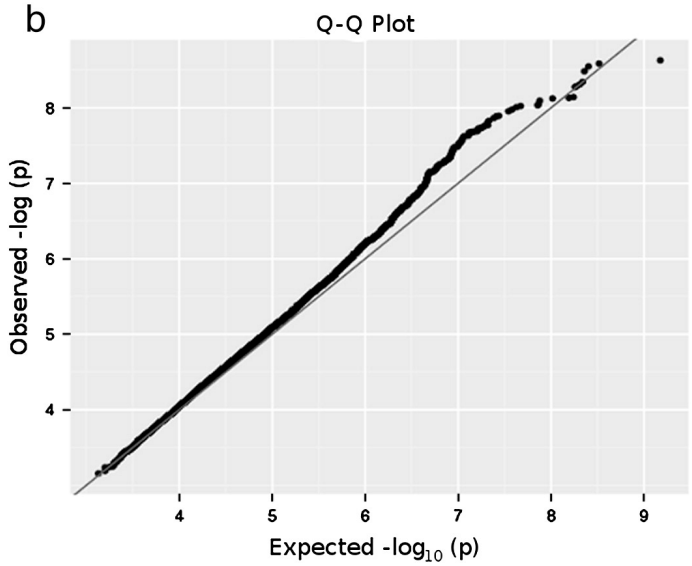
Figure 1: Examples of *P*-value histograms.

(Pounds, 2006; Dabney & Storey, 2006)

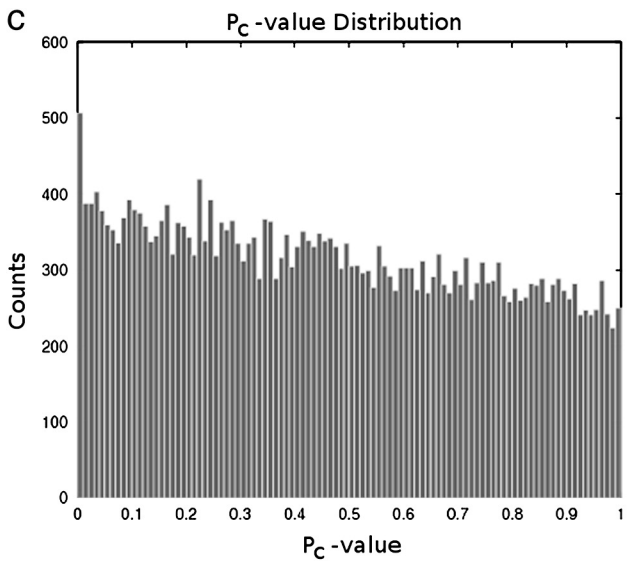
How well do results fit distributions?

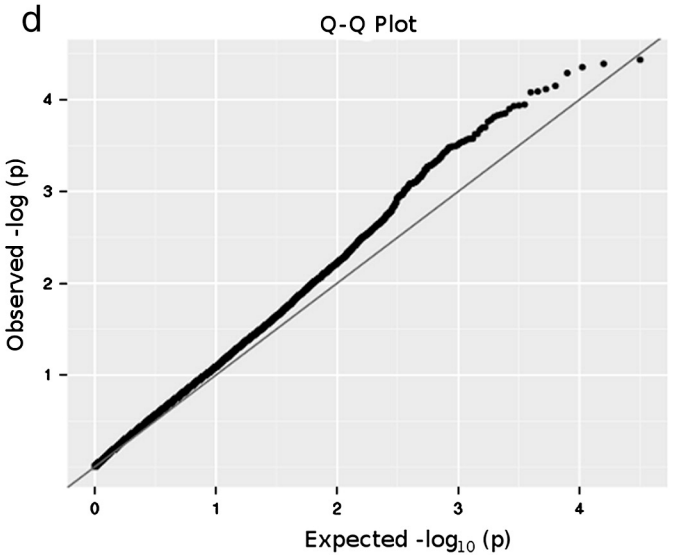
Raw *P*-value distribution



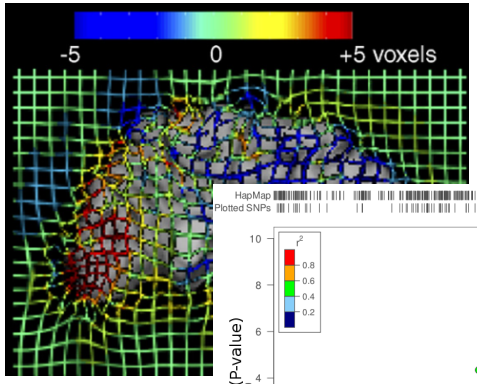


Corrected *P*-value distribution

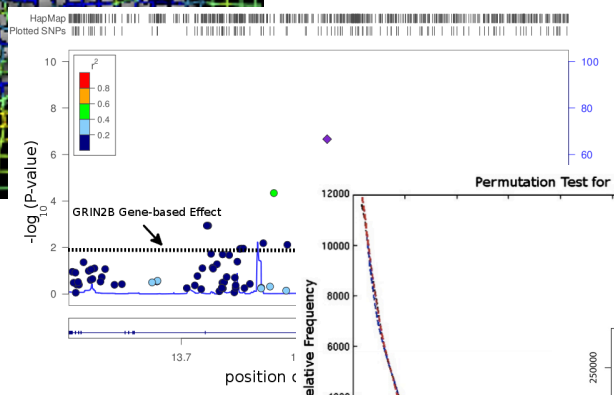




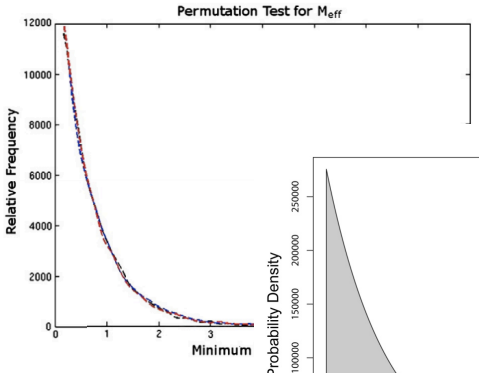
vGeneWAS



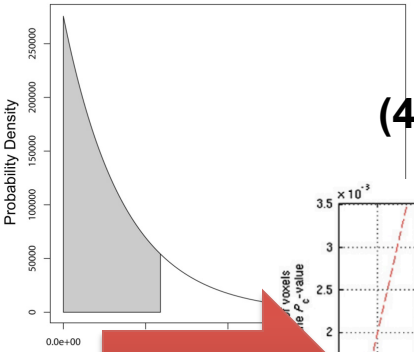
(1) Use Tensor Based Morphometry based volume differences as phenotype at each voxel



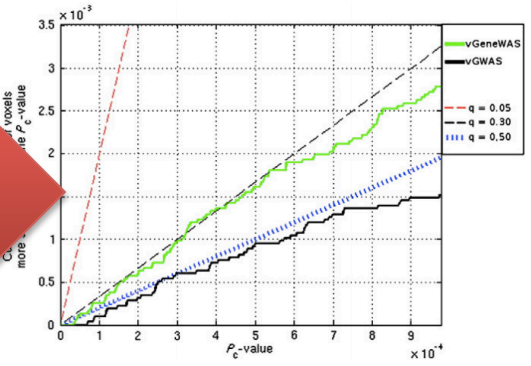
(2) GeneWAS at each voxel, select minimum *P*-value



(3) M_{eff} calculation through permutation and then estimating beta parameter of the Beta distribution



(4) CDF of Beta(1,M_{eff}) transformation



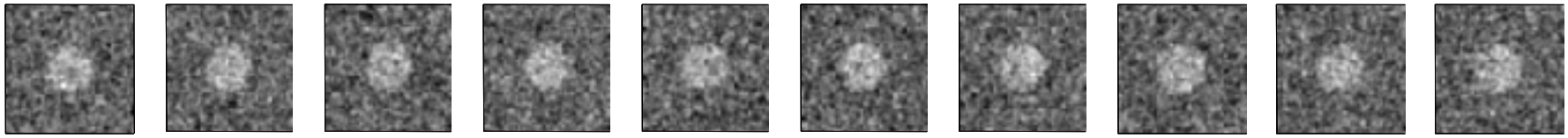
(5) FDR

Methodological Steps

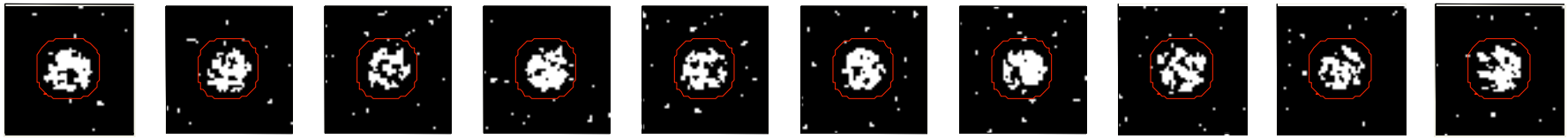
(Hibar et al., 2011)

Multiple Comparisons: Correction Across Voxels Through False Discovery Rate (FDR)

Signal + Noise



Control of False Discovery Rate at 10%

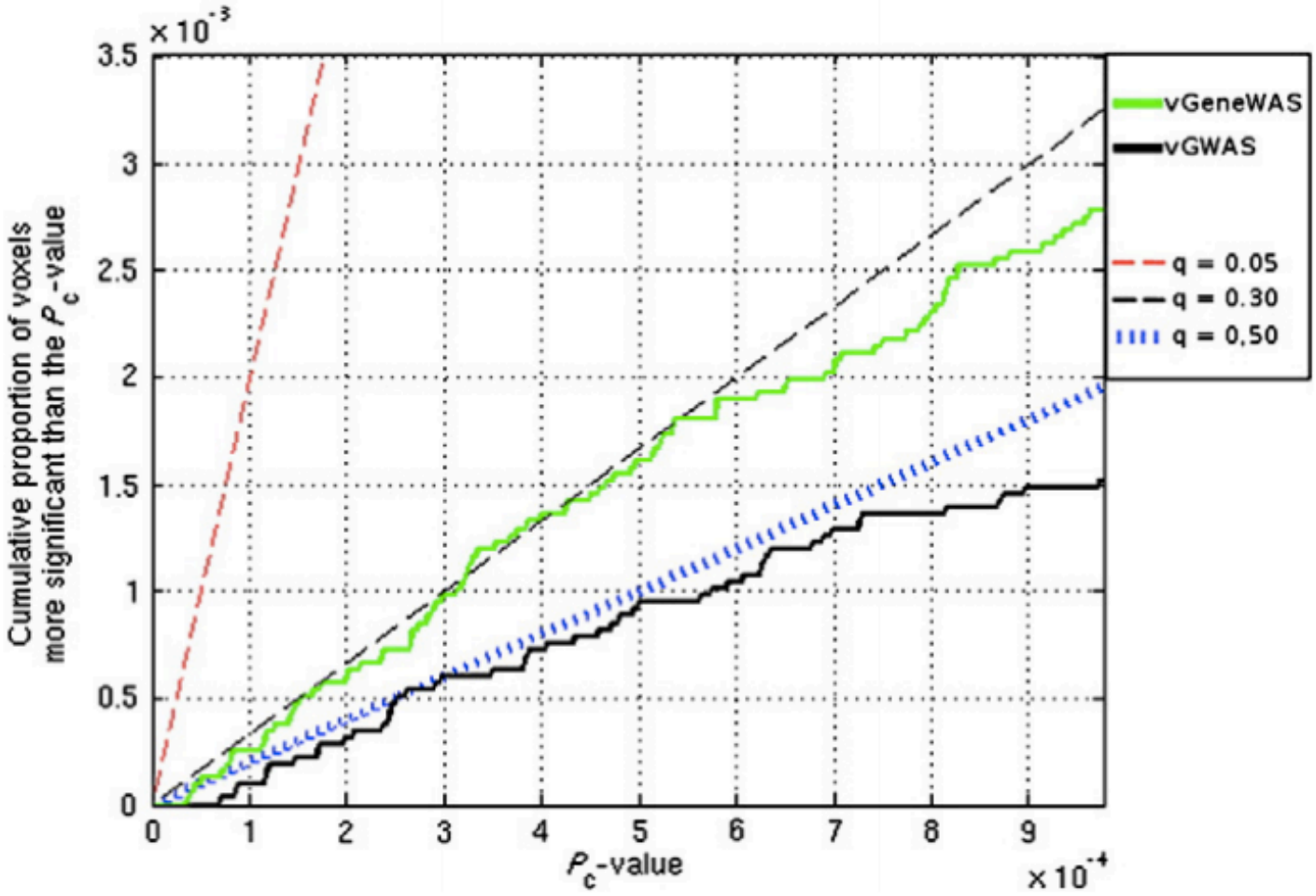


6.7% 10.4% 14.9% 9.3% 16.2% 13.8% 14.0% 10.5% 12.2% 8.7%

Percentage of Activated Pixels that are False Positives

(Tom Nichols website: http://www.sph.umich.edu/~nichols/FDR/; Genovese et al., 2002)

Power of vGeneWAS



vGeneWAS is more powerful than vGWAS in certain circumstances

(Hibar et al., 2011)

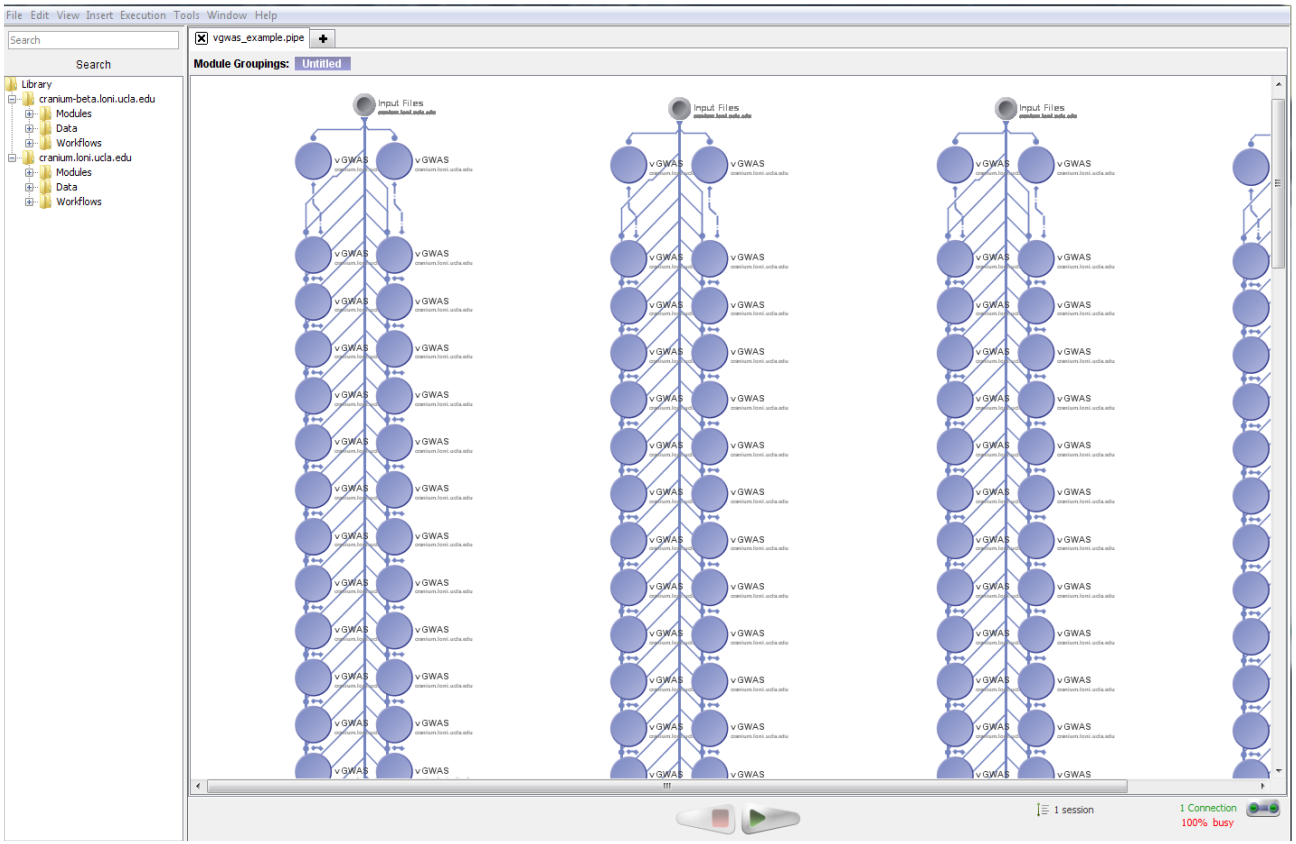
Computationally Intensive: Full GeneWAS at each voxel

Gene-wide association at each voxel takes ~6 minutes/phenotype.

31,622 voxels * 6 minutes = 132 days of computation!

Across 10 nodes using multi-core threading the

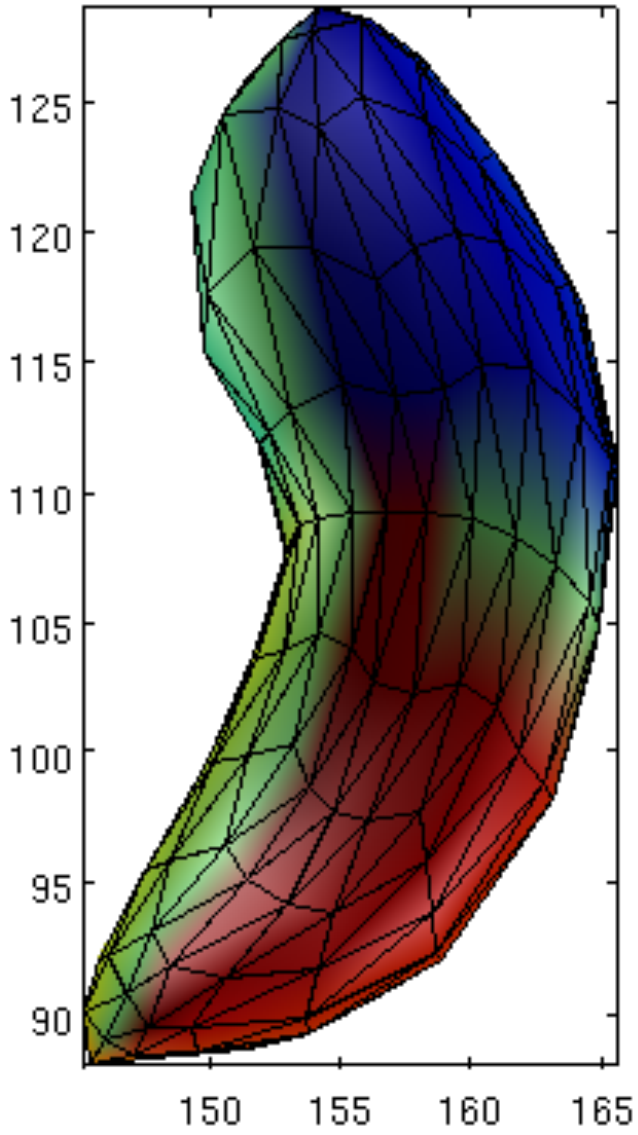
total computation time was 2 weeks!



Advantages/Disadvantages

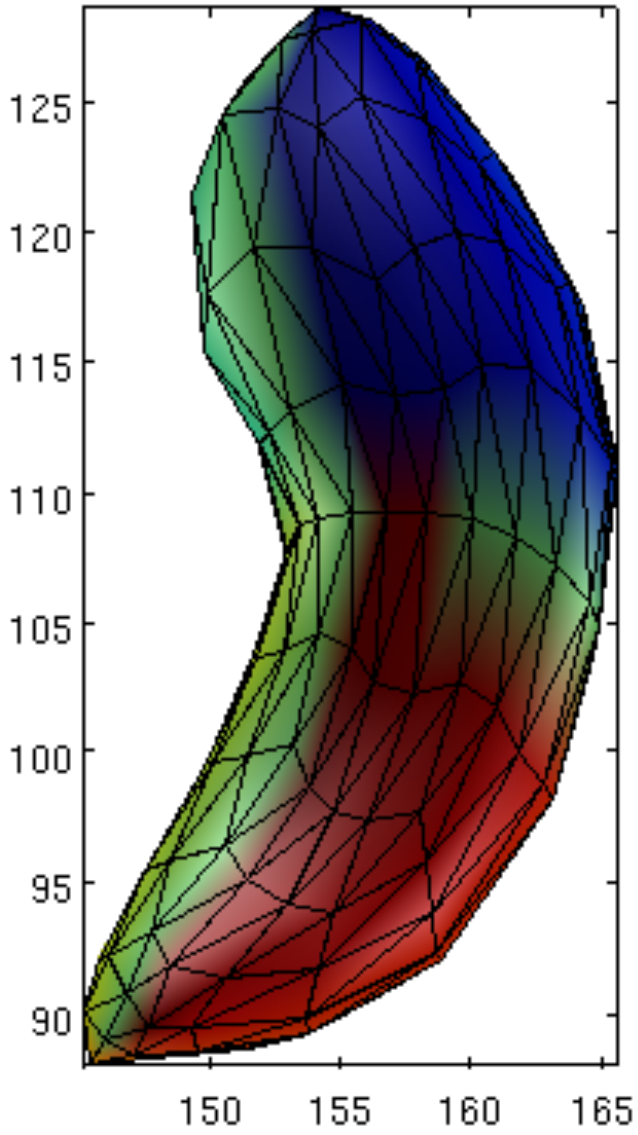
Advantages	Disadvantages
Able to jointly search the genome and imaging space to answer the question “where in the genome and where in the brain”	A strong association of one voxel to one gene is hard to interpret, we’re more interested in how a gene affects many parts of the brain
An unbiased approach to discovery, grouping by functional unit	Computationally intensive process (several days of processing) with a huge number of statistical tests
Has a small amount of data reduction because we group by gene, and is more powerful than vGWAS, depending on the effect	Selecting only the minimum P-value means that we lose a lot of information about other genes

Reducing the burden of multiple comparisons correction?


125
120
115
110
105
100
95
90
150
155
160
165

Surface-Based Morphometry

- Styner (2006) developed spherical harmonics (SPHARM) framework for describing 3D mesh surfaces
- Surfaces have been used as traits to differentiate clinical populations: schizophrenia (Styner 2006), BPD (Ong 2012), MDD (Tae 2011), AD (Looi 2010).
- Seem to be promising traits, but there are only a few studies out there applying it to clinical populations.
- We can use existing software like SPHARM-MAT to perform surface-based morphometric analysis (http://imaging.indyrad.iupui.edu/projects/SPHARM/)

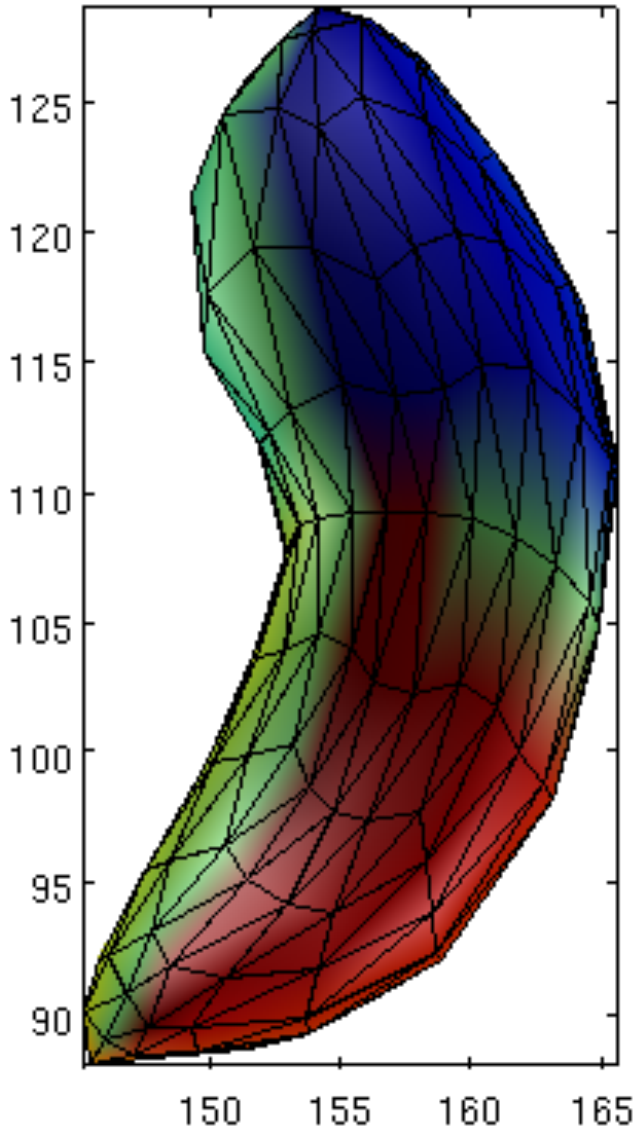


GWAS at each vertex? Significance criterion because unwieldy $5 \times 10^{-8}/365 = 1.4 \times 10^{-10}$

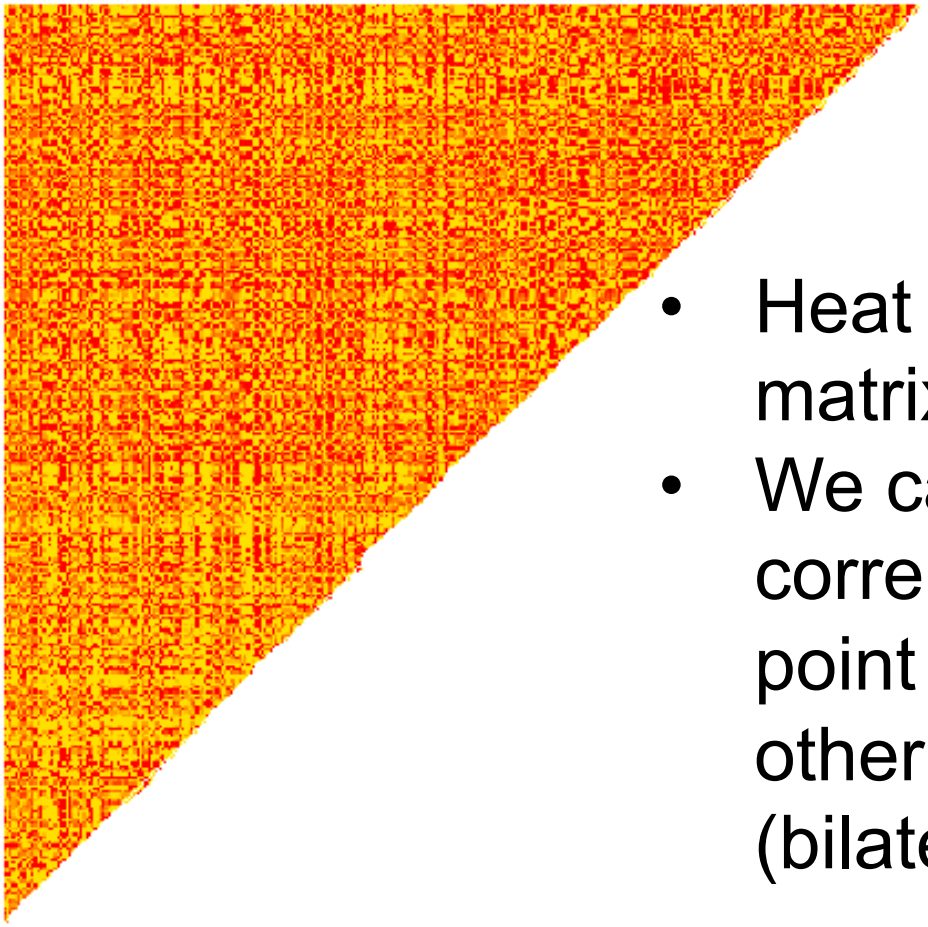
Derrek P. Hibar, Sarah E. Medland, Jason L. Stein, Sungeun Kim, Li Shen, Andrew J. Saykin, Greig I. de Zubicaray, Katie L. McMahon, Grant W. Montgomery, Nicholas G. Martin, Margaret J. Wright, Srdjan Djurovic, Ingrid Agartz, Ole A. Andreassen, Paul M. Thompson (2013). **Genetic Clustering on the Hippocampal Surface for Genome-wide Association Studies**, MICCAI 2013

How can we sensibly reduce the total number of tests?

- Grouping regions by common genetic determinants using a genetic correlation (r_g) as in Chiang 2012 and Chen 2012
- Use structural equation modeling (SEM) and bivariate trait analysis (Chiang 2009) in pairs of dizygotic/monozygotic twins to determine the extent to which two traits share common genetic determinants (r_g).

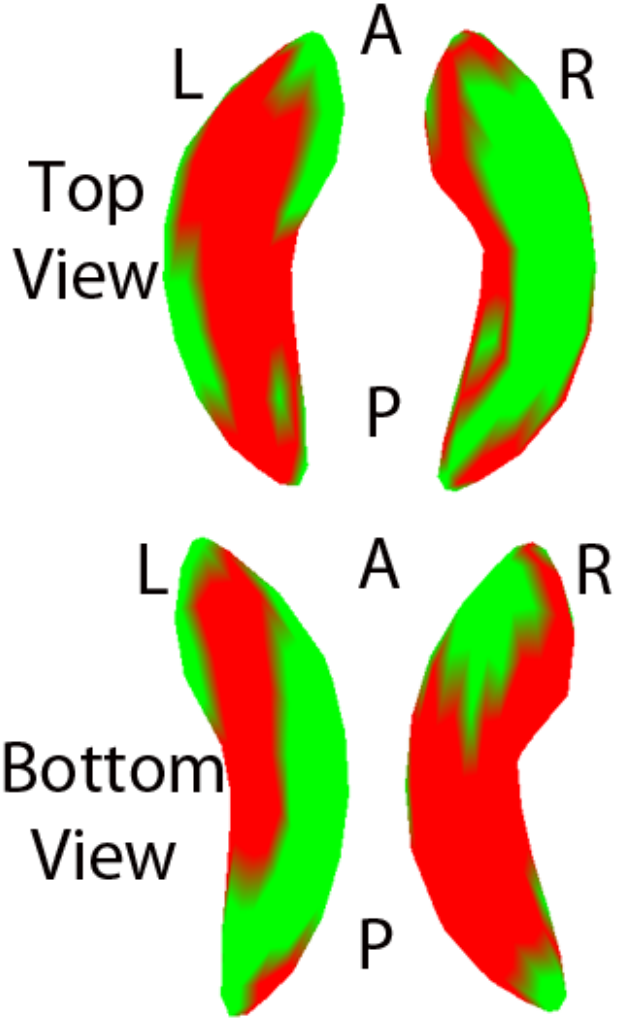

125
120
115
110
105
100
95
90
150
155
160
165

Genetic Clustering

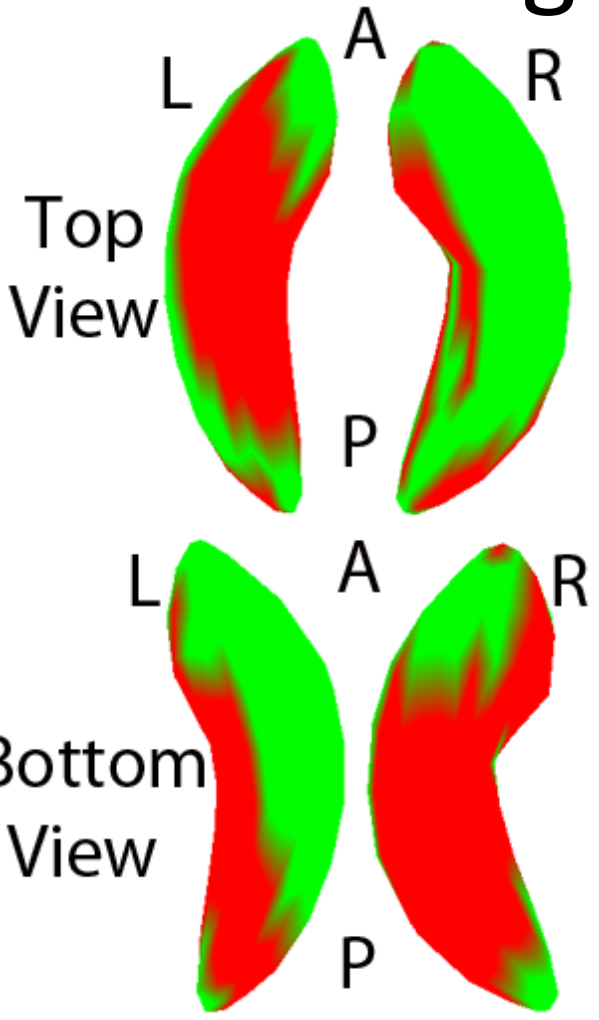


- Heat map of the r_g correlation matrix
- We calculated the genetic correlation between a given point on the surface and all other points on the surface (bilaterally)

Genotypic Clustering



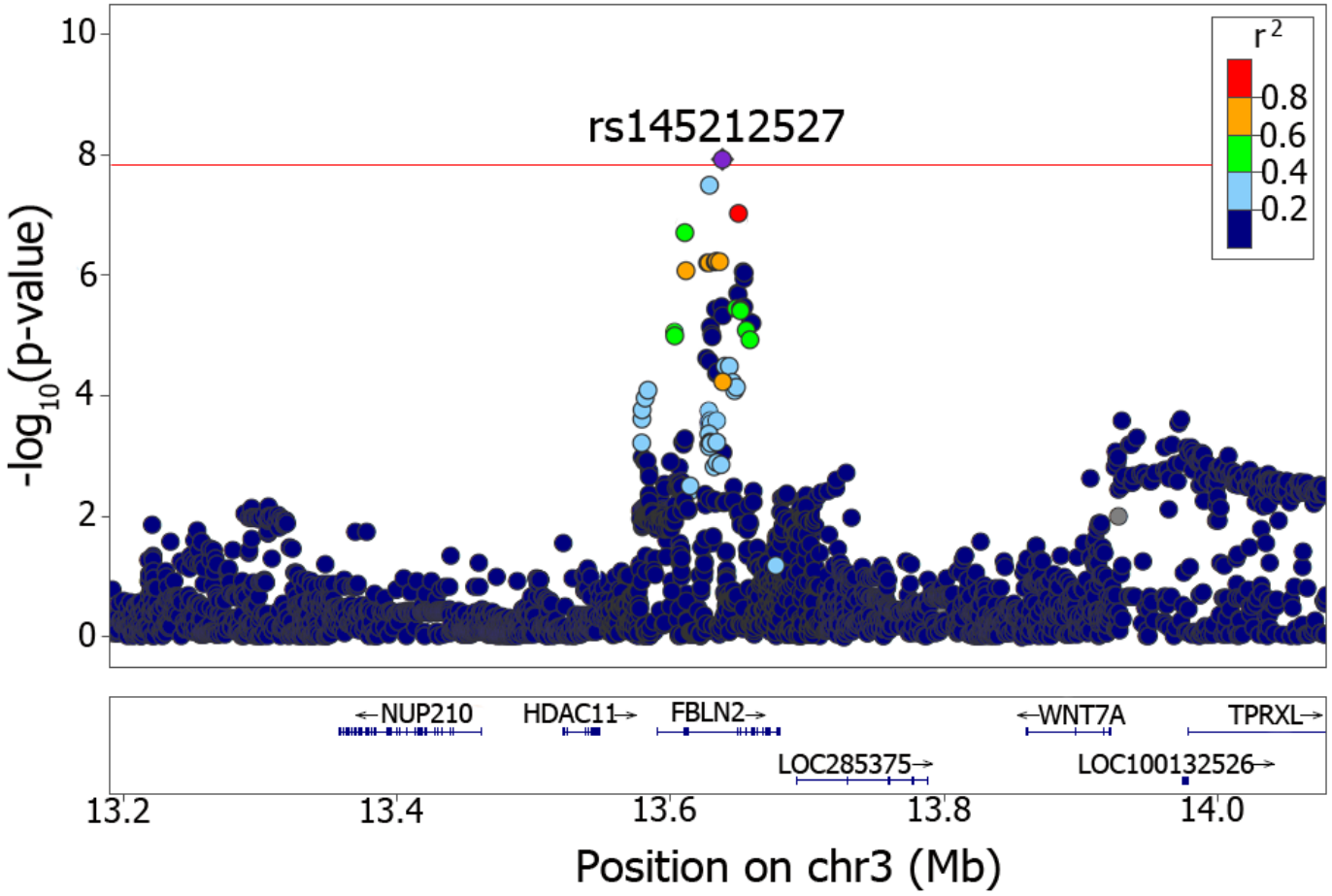
Phenotypic Clustering



GWAS

- We averaged the surface differences within each grouping based on the genetic correlation and the phenotypic correlation separately. And performed a GWAS.
- We performed a GWAS of these clustered regions in three separate datasets: ADNI (n=511), QTIM (n=571), and TOP (n=172) and then combined meta-analytically using an inverse variance-weighted method.

FBLN2


rs145212527
r 2
0.8
0.6
0.4
0.2
-log10(p-value)
10
8
6
4
2
0
NUP210
HDAC11
FBLN2
WNT7A
TPRXL
LOC285375
LOC100132526
13.2
13.4
13.6
13.8
14.0
Position on chr3 (Mb)

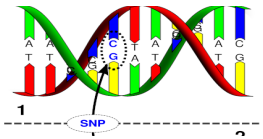
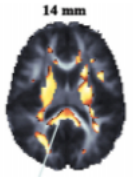
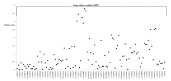
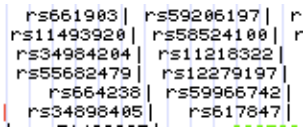
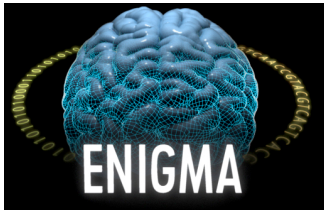
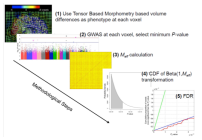
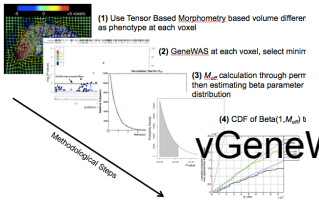
Novel Approaches

- Vounou et al., sparse Reduced Rank Regression -> leverage the sparsity of images and the genome to select features simultaneously.
- Ge et al., RFT + LSKM to increase power to detect associations
- Rosenblatt et al., vGWAS Revisited. Oral session O-T3 Tuesday 11:45p and poster 1288.

Novel Approaches

- Wan et al., Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. 2011. Sparse regression models for finding associations between selected regions on the hippocampal surface and candidate SNPS.

Imaging Genetics Menu

Genetics \ Imaging	Candidate ROI	Many ROI	Voxelwise
Candidate SNP 			 CLU & DTI ZNF804A & functional connectivity
Candidate Gene	 SORL1 & hippocampal volume		
Genome-wide SNP 			 vGWAS
Genome-wide Gene			 vGeneWAS

Scripts

- R code for performing association tests of a single SNP, a set of SNPs individually, or a set of SNPs as a group at each point within a user-provided mask
- TBM, VBM, DTI, fMRI, and others
- The set of scripts and examples can be found here:

https://github.com/dhibar/VoxelwiseRegression

Useful web resources

UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgGateway
Genome visualization magic.

Hapmap: http://hapmap.ncbi.nlm.nih.gov/
Allele frequencies in multiple populations.

Allen Brain Atlas: http://www.brain-map.org/
See where a gene is expressed.

Entrez Gene: http://www.ncbi.nlm.nih.gov/gene/
See the gene ontology (what it does).

dbSNP: http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp
The database of every documented genetic variation.

Plink: http://pngu.mgh.harvard.edu/~purcell/plink/
Incredibly useful tool for genome-wide analysis, organization, etc. Excellent documentation.

dbGaP: http://www.ncbi.nlm.nih.gov/gap/
Database of genotypes and phenotypes.

Acknowledgements

IGC-LONI (UCLA)

Paul Thompson (Advisor)

Jason L Stein

Neda Jahanshad

Omid Kohannim

Xue Hua

QTwin (Australia)

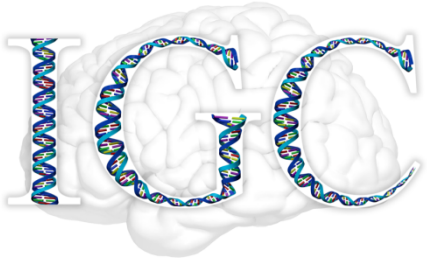
Sarah Medland

Margie Wright

Katie McMahon

Nick Martin

Greig de Zubicaray



Imaging Genetics Center

UCLA