Association Methods for Functional and Structural MRI

Thomas Nichols, PhD
Director, Modelling & Genetics
GlaxoSmithKline Clinical Imaging Centre

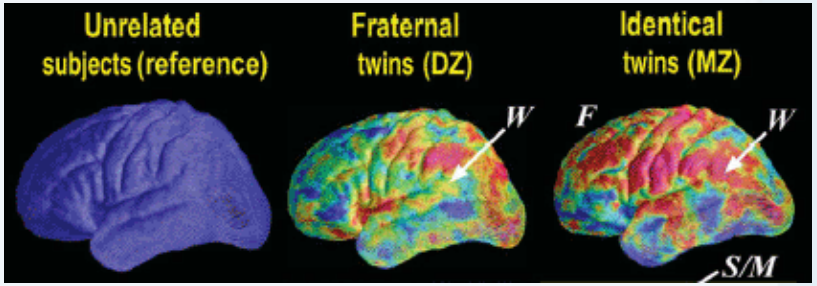
Motivation: Imaging Genetics in Drug Discovery

- Brain structure heritable
- Objective, reproducible phenotype
  - Important in psychiatry, where non-imaging measures are coarse, with poor reproducibility
- Sensitive
  - Brain anatomy/function closer to disease process than other measures
- Use to collaborate other findings
  - Use brain imaging to build confidence in marginal finding from whole-genome analyses

Brain Phenotype	h^2
Whole brain volume	0.78
Total gray matter volume	0.88
Total white matter volume	0.85

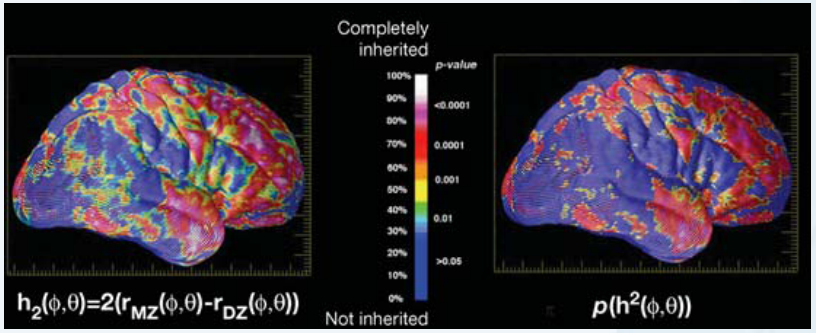
Glahn, Thompson, Blangero. Hum Brain Mapp 28:488-501, 2007

Thickness of Cortical GM (r^2)



Thompson et al, Nature Neuro, 4(12):1253-1258,. 2001

Heritability of GM Thickness (h^2 & corrected P-value)



Thompson & Toga, Annals of Medicine 34(7-8):523-36, 2002

Outline

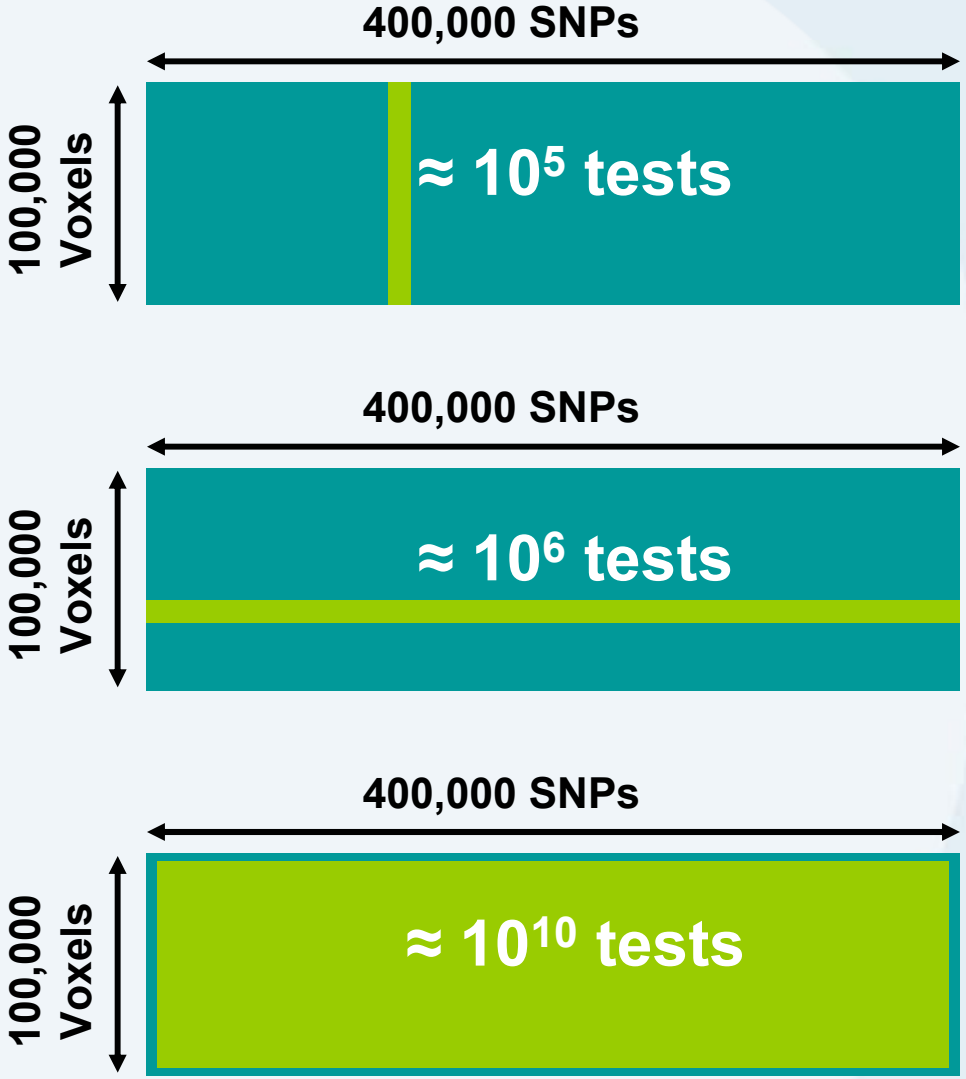
- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

Types of Imaging Genetics Analyses

- Brain Imaging already high-dimensional
  - ≈ 100,000 voxels
    - Highly correlated
- Genetic data also high-dimensional
  - ≈ 20 million known SNPs
    - The 0.5-1m tagging SNPs typically used are lightly correlated
  - ≈ 30,000 genes
- How to deal with all this multiplicity!?!

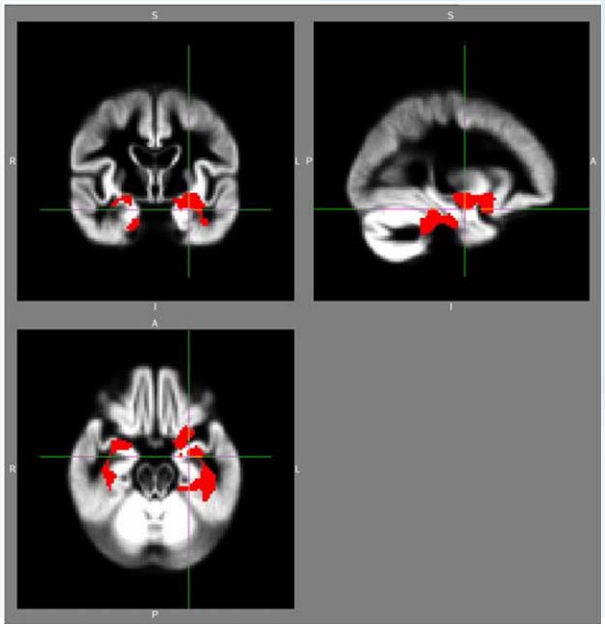
Types of Imaging Genetics Analyses

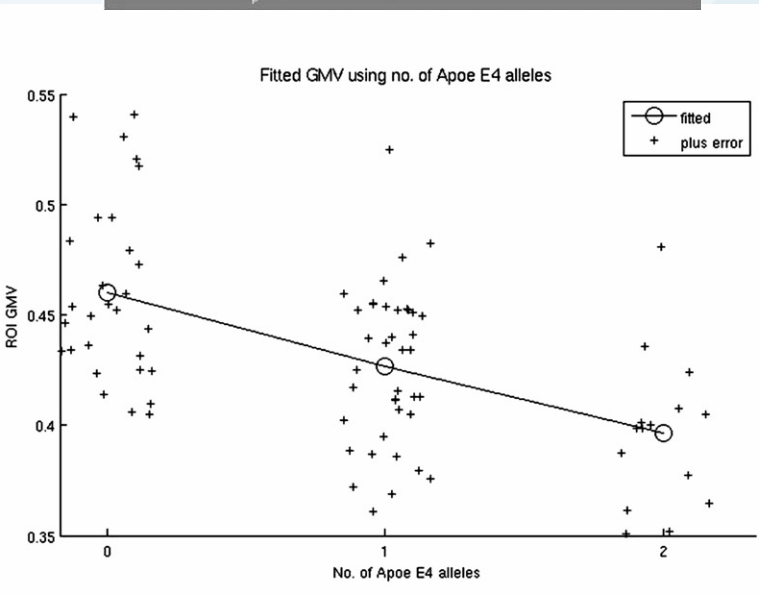
- Candidate SNP
  - Traditional imaging analysis w/ SNP predictor
- Region of Interest or 1 # summary
  - Traditional Whole-Genome Analysis
- Whole-Brain, Whole-Genome



Whole Brain, Candidate SNP Analyses

- One Genetic Marker selected *a priori*
  - Either single SNP, or single variant of a gene
- Example
  - VBM Association of GM & ApoE ε4 in Mild AD
    - Filippini et al (2009). Anatomically-distinct genetic associations of APOE ε4 allele load with regional cortical atrophy in Alzheimer's disease. NeuroImage 44:724–728





Imaging ROI, Whole Genome Analyses

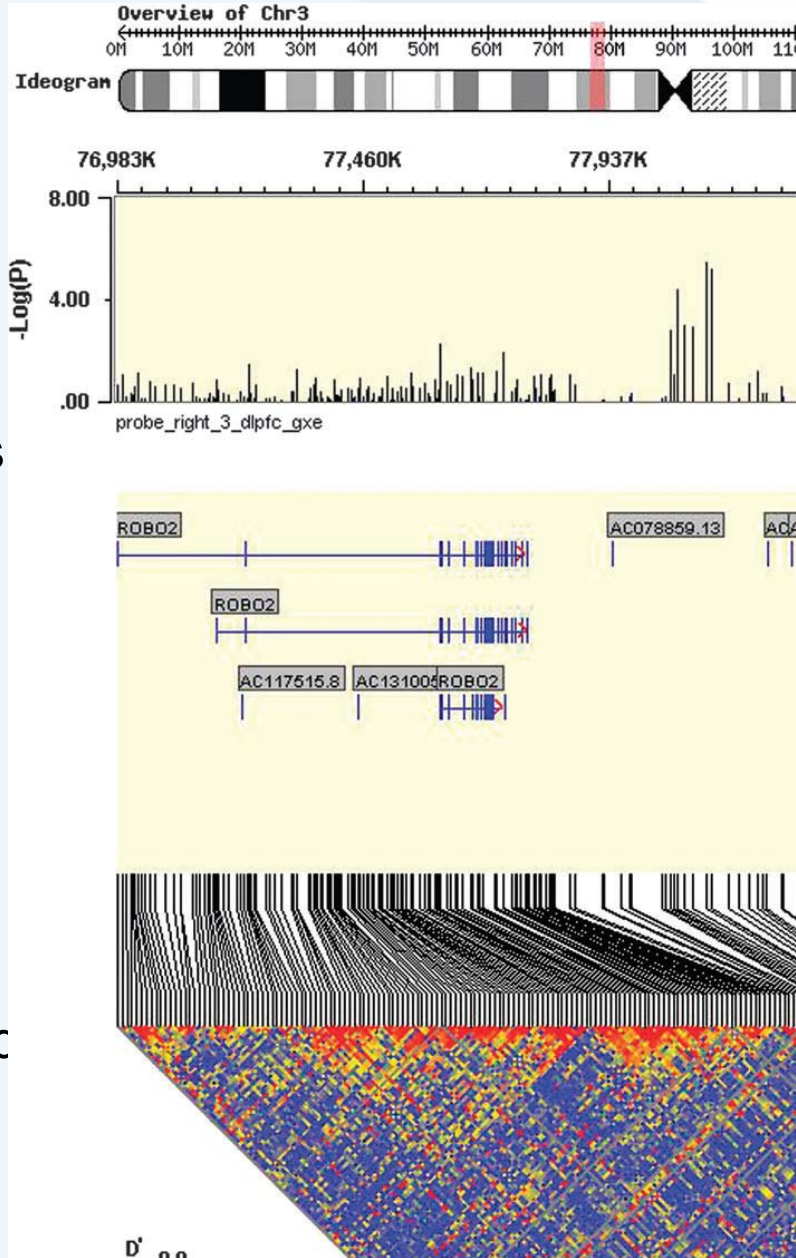
- One Imaging phenotype selected *a priori*
  - Either a ROI value (e.g. % BOLD change) or some single-number summary (e.g. total brain GM)
- Example
  - WGA Association in MS, n=794
  - Total brain volume results
    - No GWA sign.

SNP	Chrom	Position	GeneSymbol	Alleles	Minor allele frequency	Adj Geno Log *P*-value
Brain parenchymal volume						
rs4866550	5	3361312	IRX1	C/T	0.32	6.06
rs10078091	5	25530762	CDH10	A/G	0.27	5.91
rs368380	20	14762090	C20orf133	C/T	0.33	5.73
rs4473631	4	174876499	MORF4	A/C	0.22	5.55
rs1869410	2	5207954	SOX11	C/T	0.28	5.40
rs261902	12	32367994	BICD1	T/C	0.16	4.42

Baranzini et al. (2009). Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Human Molecular Genetics 2009 18(4):767-778.

Imaging ROI, Whole Genome Analyses

- None known that use no-dimension reduction
  - Typically, reduce imaging dim
  - Set of comprehensive ROI's
  - Reduced resolution voxel-wise analysis
- Example
  - Schizophrenia WGA with %BOLD fMRI quantitative trait (QT)
    - n=64 SCZ, n=74 matched controls
  - QT is % BOLD in DLPFC for Sternberg Item Recognition Paradigm
    - Tested for QT × {NC,SCZ} interaction
  - Found weak evidence for six genes at $\alpha<10^{-6}$ (ROBO1-ROBO2, TNIK, CTXN3-SLC12A2, POU3F2, TRAF, and GPC1)
  - Potkin et al. (2009), Schizophrenia Bulletin 35:96–108.



Outline

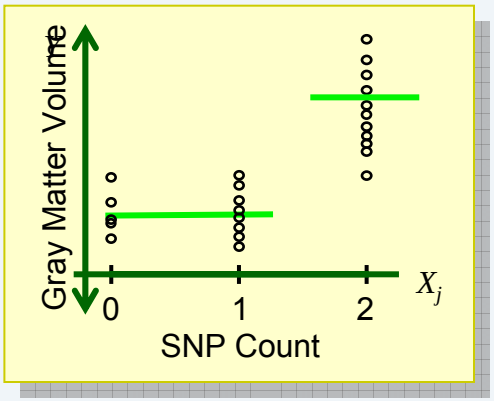
- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

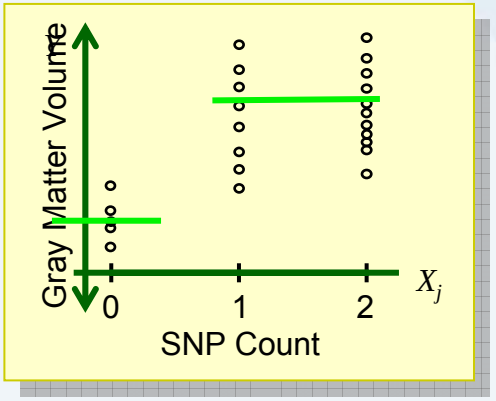
Modelling Imaging Data With Genetic Variables

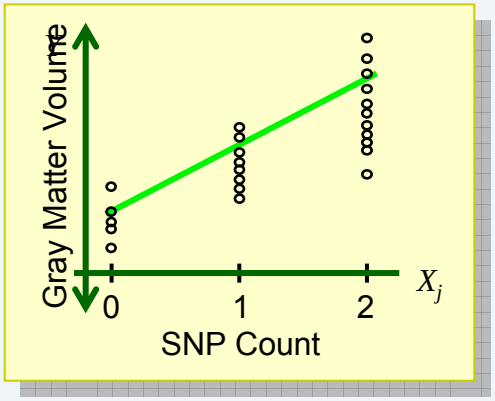
- Mass Univariate Modelling
  - Fit same univariate linear model at each voxel/ROI
- Quantitative Trait Multiple Regression
  - Linear model fit at each voxel
- Regressors
  - Genetic
  - Group (Case/Control)
  - Demographic / nuisance variables
  - etc

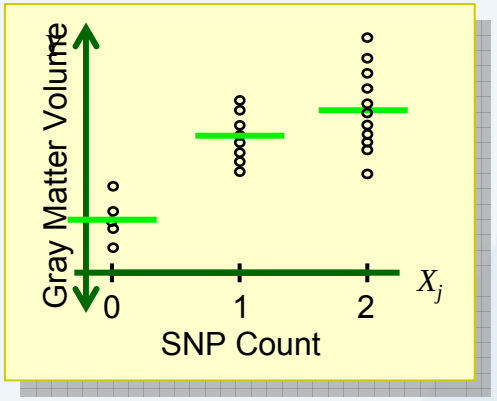
Genetic Models for SNP data

- Recessive
- Dominant
- Additive
- Genotypic





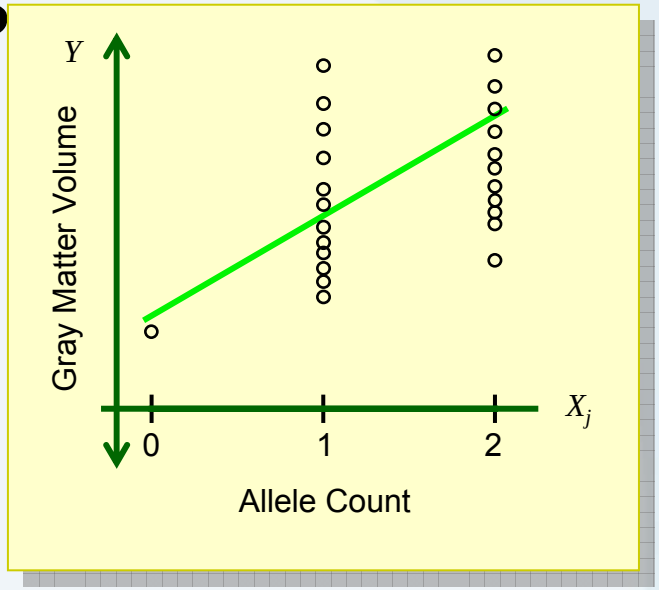




Genetic Models for SNP data: Power

- Q: What's the Optimal Model?
  A: The Correct One!
- True model unknown
  - Common disease, common variant hypothesis for complex diseases
  - Expect many genes contributing to risk
  - Don't expect to find one single SNP with simple Medelian influence
- To avoid yet further multiplicity, typical practice is to pick a one model
  - Fit additive, hope its additive
  - Additive seems like single best model for association studies: B Freidlin et al, Hum Hered, 53:146-152, 2002

Genetic Models for SNP data: Robustness



- Concerns about influence
  - When minimum allele frequency (MAF) too low, rare homozygotes may become influential
- Merge rare homozygotes with heterozygotes
  - Cutoff?
  - 5% MAF cutoff is common in GWAs, but corresponds to $0.05^2 = 0.25\%$ frequency!
    - 5% MAF, 100 subjects → < 1 rare homozygote expected!
  - 32% MAF cutoff → $0.32^2 = 10\%$ frequency
  - Or just set arbitrary limit (e.g. 10) below which rare homozygotes are merged with heterozygotes

Mass Univariate Modelling Nuisance Effects

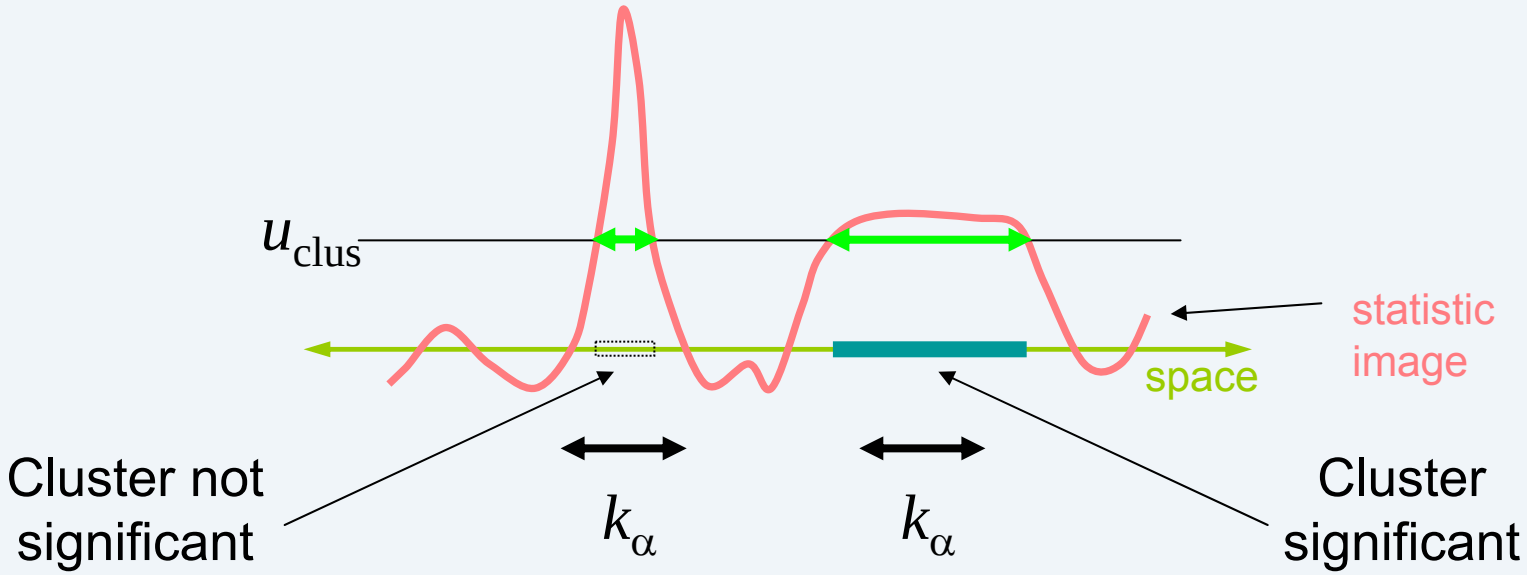
- Age & Gender
  - Substantial normal variation in GM w/ Age
- Total Gray matter (for VBM)
  - Discounts global changes to find localized changes
- Other
  - Site
  - Medication
  - Anything that is also related to the genetic effects

Outline

- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

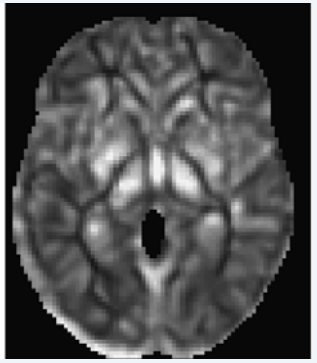
Inference On Images for Img.Gen… Nothing Special

- Voxel-wise
  - Reject Ho, point-by-point, by statistic magnitude
- Cluster-wise
  - Define contiguous blobs with arbitrary threshold u_{clus}
  - Reject Ho for each cluster larger than k_{α}



Cluster Inference & Stationarity

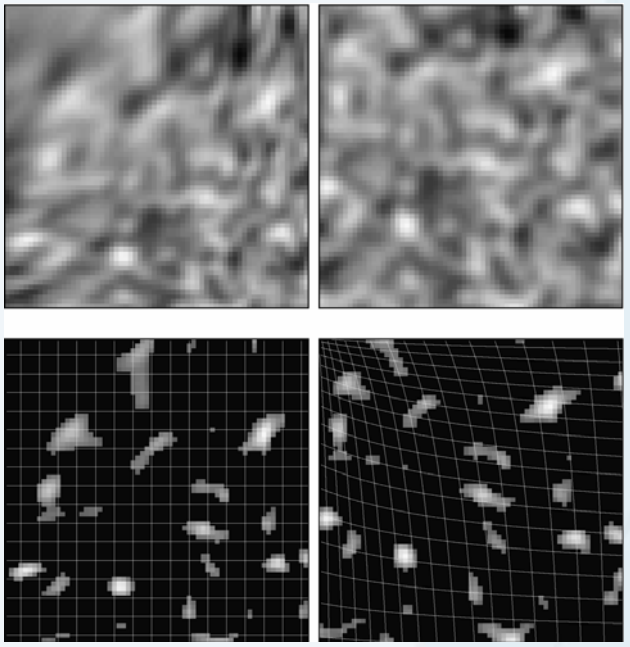
- Cluster-wise preferred over voxel-wise
  - Generally more sensitive
    Friston et al, NeuroImage 4:223-235, 1996
  - Spatially-extended signals typical
- Problem w/ VBM
  - Standard cluster methods assume stationarity, constant smoothness
  - Assuming stationarity, false positive clusters will be found in extra-smooth regions
  - VBM noise very non-stationary
- Nonstationary cluster inference
  - Must un-warp nonstationarity
  - Available as SPM toolbox
    - Hayasaka et al, NeuroImage 22:676– 687, 2004
    - http://fmri.wfubmc.edu/cms/software#NS
    - Also in Christian Gaser's VBM toolbox



VBM:
Image of
FWHM
Noise
Smoothness

Nonstationary noise…

…warped to stationarity



Inference on Images

- Must account for searching over space
  - 1 voxel / 1 ROI
    - No correction
  - k ROIs
    - Bonferroni (largish ROI should be fairly independent)
  - Whole brain, masked voxel-wise analysis
    - FWE, FDR correction for voxel-wise or cluster-wise analysis

Outline

- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

Inference Over the Genome

- Just with imaging, pay enormous power hit for un-constrained search
- 1 SNP
  - No correction
- 1 gene
  - For k tagging SNPs, Bonferroni OK
  - Better corrections available for dependent SNPs
- All SNPs, genes
  - Permutation methods, improved Bonferroni methods
  - FDR

One Inference Strategy: GSK CIC Candidate SNP Protocol

- Define strict primary outcome
  - For given gene, use single SNP
    - Best (large) association study significance, otw
    - Best nonsynonymous exonic available, otw
    - Best 5' intronic available
  - For each SNP, only consider main effect of gene
    - If fitting gene x group interaction, test for average effect
      - Any association is more likely than a disease-specific association
      - Even if disease-specification association, opposing sign of effect unlikely w/ VBM
  - 1-number summary per gene
    - Minimum nonstationary cluster FWE-corrected P-value for association (1 DF F-stat)
  - Bonferroni correction for number of genes
- Primary outcomes then have strong FWE control
  - Over brain, over genes
  - $(1-\alpha)100\%$ confidence of no false positives anywhere
- Secondary outcomes
  - Interactions, sub-group results
  - Use same FWE-inferences, but mark as post-hoc

Inference Over the Genome: Combining SNPs

- To pool SNPs within genes, typically separate models are fit & P-values are combined…
  - Tippett's Method (1931)
    - Minimum P-value
  - Fisher's Method (1950)
    - Based on product of P-values, equivalently $-2 \times \sum_i \log P_i$
  - Stouffer's Method (1949)
    - Scaled Average Z, $\text{Avg}(Z) \times \sqrt{n} \sim N(0,1)$, $Z = \Phi(1-P)$
- Same approaches used to combine gene inferences within networks

See: Poster #178 SU-PM, TE Nichols, "Comparison of Whole Brain Multiloci Association Methods"

Inference Over the Genome: Haplotypes

- Haplotypes
  - Set of closely linked genetic markers
  - Tend to be inherited together
  - Example
    - 3 SNPs within a gene, alleles: A/T, A/T, C/G
    - This could give rise to $2^3 = 8$ possible haplotypes: AAC, TAG, TAC, AAG, ATC, TTG, TTC, AAG
    - Fit regression model 8 regressors, use F-test to find any haplotype variation
- Should be more sensitive then separate models, but high-DF F-tests are often have low power
  - Unless small number of SNPs, SNP-combining probably better

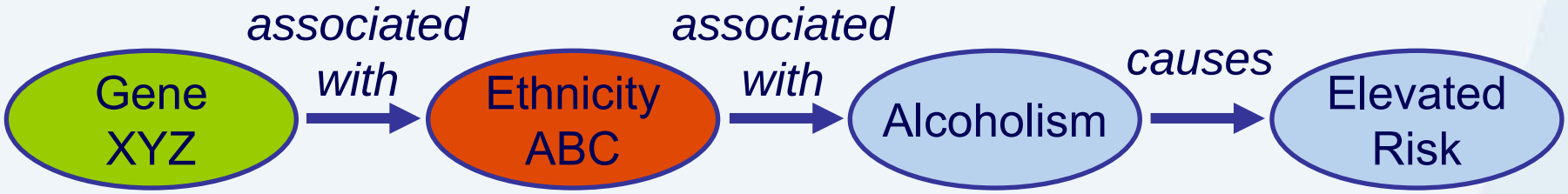
Outline

- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

Population Substructure

- When sample is a mix of ethnicities, can find spurious correlations
- Example: Coronary Artery Disease
  - Find association btw gene XYZ & heart attack incidence. Conclude?

    - Great!
  - Or…



    - Oop's… I've only discovered that gene XYZ is an ancestry marker!

Population Substructure

- Solution
  - “Admixture modelling” or PCA-based methods (“eigen-strat”)
  - Methods find large scale patterns of genetic variation that typify different sub-groups of your population
  - Can enter these patterns as nuisance variables to discount such variation creating false positives
- Problem with the solutions
  - Need large sample sizes (1,000's) to adequately deal with this
  - Remains potential source of false positive risk for typical tiny imaging genetic sample sizes
- Pragmatic solution
  - Work closely with genetics colleagues to define ethnically homogeneous study groups
  - Build imaging sample as subset of large (1000+) association samples, get population stratification covariates based on entire sample

Statistical Validity vs. Face Validity

- Statistically Inference
  - Optimally sensitive results are obtained from modelling all data jointly
  - A positive result is an inference on the population sampled
- Current Statistical Genetics Practice
  - One study a publication does not make
  - Any positive result must be **replicated** in an independent population
    - Result of high incidence of unreplicable early findings in GWAS
    - Also possible population substructure problems

Statistical Validity vs. Face Validity

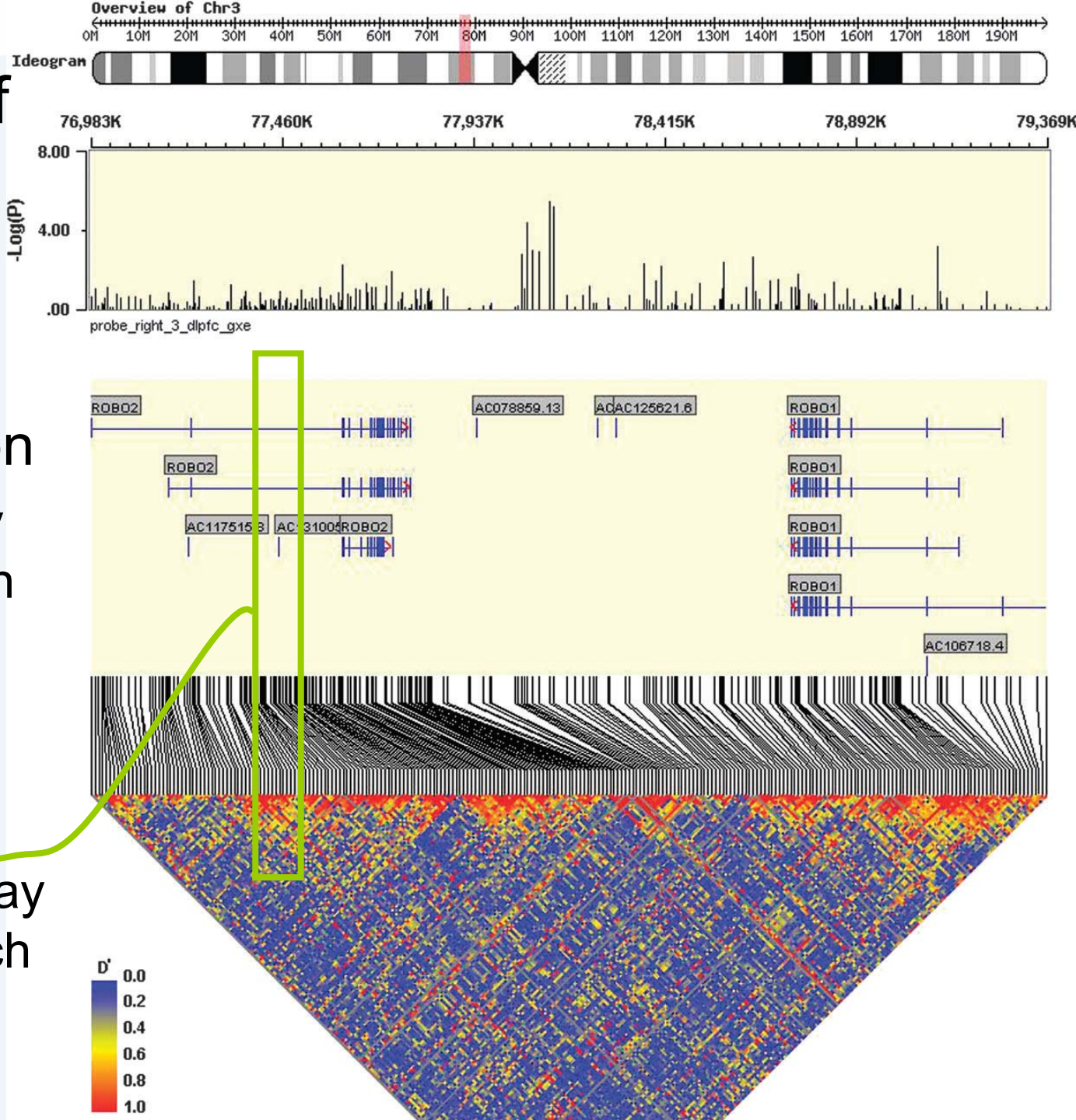
- Replication is desirable
- In defence of imaging genetics
  - In genetics, FWE significance in a GWAS study is almost never seen
    - Typical is a fixed rule-of-thumb GWAS $\alpha = 5 \times 10^{-7}$
    - Imaging literature is rife with uncorrected inferences, but whole brain corrected significance is seen
  - All GWAS intuition is on a *categorical* phenotype, “Case” or “Control”
    - Quantitative phenotype, especially one derived from a *designed experiment* (i.e. fMRI) may well have better power

Further Limitations

- Basic stats quiz, A or B?
  - A: “This genetic variant causes more gray matter in MTL”
  - B: “This genetic variant explains differences in grey matter in MTL”
  - (Causality vs Causation)
- Remember even more sources of false positives
  - Data quality, outliers
    - Check plots of intriguing results for outliers
  - Linkage Disequilibrium (LD) & Mis-localization
    - Significant SNP can inside Gene X’s exon, but in LD 2 or 3 other genes!!
  - Gene networks
    - Other genes in tightly regulated network may give similar results
    - Non-unique effect

Challenges of Localization

- Results for ROBO2-ROBO1 region
  - Note near by genes in high LD regions
  - If a strong association were found here ● no way to know which gene responsible



Outline

- Types of Imaging Genetics Analyses
- Models for Genetic Effects
- Inference Over the Brain
- Inference Over the Genome
- Limitations
- Conclusions

Conclusions

- Understand the Genetic Models
  - Additive default choice
- Understand the Limitations
  - Population substructure, need for replication
- Massive Multiple Testing Problem
  - Limit search whenever possible, over the brain & genes/SNPs
- Befriend a geneticist!
  - No way to good science with out a tight collaboration