Genetic Association Studies

Matthew C. Cowperthwaite, PhD
NeuroTexas Institute at St. David's HealthCare
Texas Advanced Computing Center, The University of Texas at Austin

Disclosures

Nothing to declare.

Goals

- Understand the basic history of quantitative genetics
- Attain a basic understanding of the genetic terminology
- Understand the basic statistical framework of genetic association studies
- Gain insight into how genetic information can provide additional power to imaging research

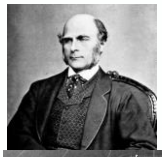
The Dawn of Genetics

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Traits

Discrete traits



Quantitative traits

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

A Great Debate

Are the mechanisms of inheritance the same?

Mendelians
Traits evolve by few large mutations

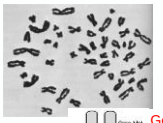
Biometricians
Traits evolve by many small mutations

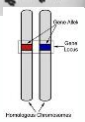
NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Genetic Terminology

Human Karyotype





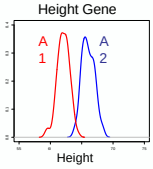
Gene := heritable "unit", corresponding to region of DNA (the *locus*)

Allele := specific form of a gene, i.e. variant

Flower Color Gene

Purple Allele White Allele

Height Gene

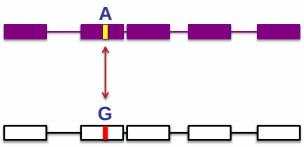


Height

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Single Nucleotide Polymorphisms



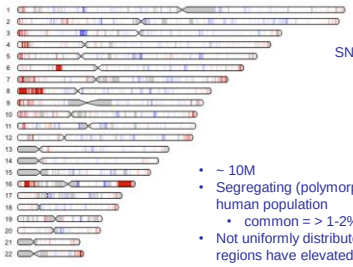
Single-base mutation at a locus

Polymorphic – variant is segregating in the human population

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Many SNPs in the Human Genome



SNP Density

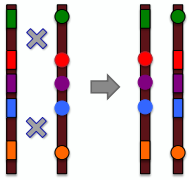
- ~ 10M
- Segregating (polymorphic) in the human population
  - common = > 1-2%
- Not uniformly distributed (red regions have elevated SNP density)

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

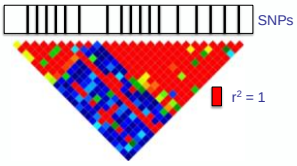
Surveying the Genome with Tag SNPs

Recombination



Linkage Dysequilibrium
- Non-independent inheritance

Tag SNPs
- SNPs that represent regions of high LD

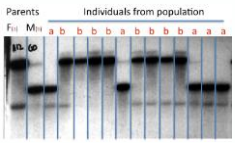


NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

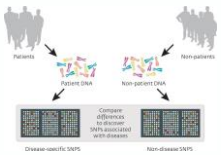
GWA Studies

Classic genetic linkage analysis



- Low throughput
- Bias – geneticist often looking at one (or a few) genes with some prior knowledge

Genome-Wide Analysis

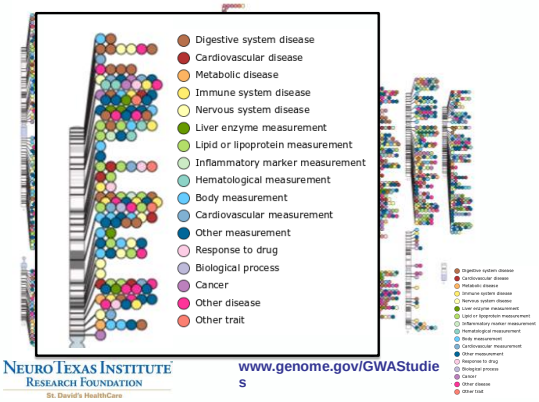


- High throughput
- Less bias – survey SNPs that represent most of the genome

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

SNP:Disease Associations



NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

www.genome.gov/GWAStudies

SNP Genotyping

Microarray-based Assays ("chips")

Affymetrix GenomeWide 6.0
~ 1 M tag SNPs

Illumina HumanHap 550
~ 1 M tag SNPs

DNA Sequencing-based Assays

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

Leading Neuroscience Research

Sequencing Based SNP Discovery

In a sequencing based method, we sequence genomes or genes of interest and compare our sequence data to known databases of variation

Next-Gen Sequencing (NGS)
- Costs are plummeting
- Exome (all protein coding genes) ~ $2K
- Whole Genome ~ $6-8K

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

Leading Neuroscience Research

Arrays versus Sequencing

- "Chips"
  - Cheaper (for now…)
  - Survey common variants (> 2% within population)
  - Not identifying the causal variant(s); most introns and intergenic
- Sequencing
  - Identify common and rare variants
  - May identify new mutations and causal variant(s)
  - Can also get coding sequence of many genes

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

Leading Neuroscience Research

GWA Study Design

- Trait or phenotype of interest
  - Discrete – categorical or case-control study
  - Continuous – quantitative study
- Power to detect significant associations
- Chip-based assay or NGS genotyping

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

GWAS Power

Image of typical Affymetrix GW6.0 hybridization
~ 1M SNPS, 11X redundancy

Large multiple-testing load:

- Bonferroni:
  $q_{(i)} = \alpha/N = 0.05/10^6 \approx 10^{-8}$
- FDR and derivatives:

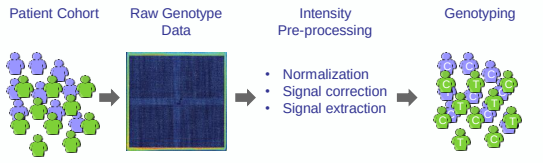
$$q_{(i)} = \frac{Np_{(}i)}{i}c(N)$$

Overcoming the load:
- Large N, but expensive and time-consuming
- Test fewer SNPs, for example two-stage GWAS

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Array Data Processing



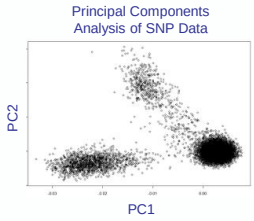
Affy: Birdsuite
Illimuna: BeadStudio

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Additional Data Considerations

- Genotype curation
  - Missingness
  - Hardy-Weinberg Equilibrium
  - Odd Mendelian Inheritance Patterns
- Population Stratification



NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

GWAS Study Design: Case-Control

Cases
(affected with some disease)

Controls
(not affected)

	Case	Control
C	8	5
T	2	5

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Case-Control Statistics

- Count frequency of each *allele* in case and control population
- Only identify significant differences in frequency
- No models of dominance, recessiveness, etc.

	Case	Control
C	16	10
T	4	10

Chi-Square Test

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Fisher's Exact Test

$$p = \frac{\binom{N_{cases}}{N_{cases_C}}\binom{N_{ctl}}{N_{ctl_C}}}{\binom{N}{N_C}}$$

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Case-Control Statistics

- Model frequencies of *genotypes*
- Can model interactions between alleles (e.g. dominance)

	Case	Control
C/C	6	3
C/T	3	4
T/T	1	3

Cochran-Armitage Trend Test

$$T \equiv \sum_{i}^{(CC,CT,TT)} w_i(N_{case_i}R_{ctl} - N_{ctl_i}R_{case})$$

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

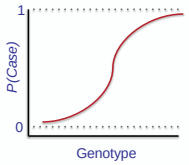
Leading Neuroscience Research

Case-Control Statistics

- Modeling probability of case
- Potential for prediction and/or classification
- Covariates can be included

Logistic Regression

$$log(\frac{P}{1-P}) = \beta_1 x + \beta_0$$

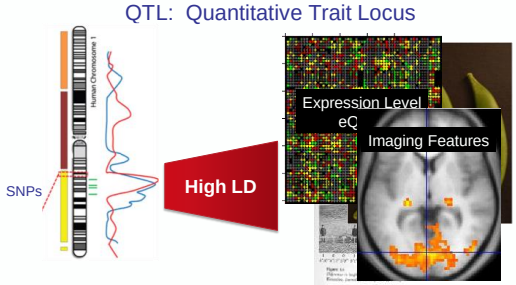


NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Quantitative Association Studies

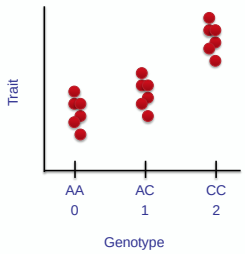
QTL: Quantitative Trait Locus



NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Quantitative GWAS Statistics

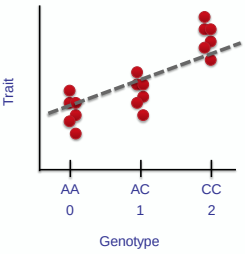


Spearman's correlation Coefficient (*rho*)

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i = difference in ranks between genotype and trait values

Quantitative GWAS Statistics



Linear Regression

$$Y = \beta_i X + \beta_0$$

- Models
- Covariates
- Perhaps less robust

GWAS Software

- Pre-processing:
  - Affy PowerTools/Birdsuite @ Affymetrix
  - Illumina BeadStudio for Illumina data
  - RMA adjustment – many implementations in MATLAB, R/Bioconductor, and more
  - Eigenstrat - http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html
- Associations
  - PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/
  - GWASTools: http://www.bioconductor.org/packages/devel/bioc/html/GWASTools.html
  - Merlin: http://www.sph.umich.edu/csg/abecasis/merlin/index.html

GWAS Data

- Raw data:
  - dbGAP: http://www.ncbi.nlm.nih.gov/gap
  - dbSNP: http://www.ncbi.nlm.nih.gov/SNP/
  - Other repositories:
    - NIH validated: http://gds.nih.gov/02dr2.html
    - ADNI: http://www.adni-info.org/
    - TCGA: http://cancergenome.nih.gov/
    - Ask researchers for source data…
- Meta-data
  - GWAS Catalog: http://www.genome.gov/gwastudies/

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Alzheimer's Disease

- Most common form of dementia
  - Familial – defects in APP processing
  - Sporadic – mutations in APOE-4, CLU, CR1, …
- Loss of memory and other cognitive functions

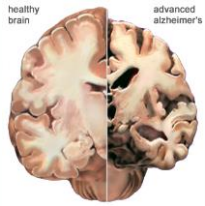


Image: www.alz.org

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

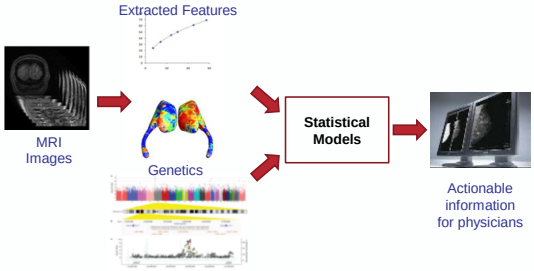
Identifying Risk Factors for Alzheimer's Disease

- Diagnostic image-based biomarkers for development of AD
- Identifying genes and pathways that can improve power of imaging biomarkers
- Improve CAD systems for early detection and tracking progression of Alzheimer's disease

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

AD Imaging Genomics



NEUROTEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Image-Genomics Study

- Meta-analysis of AD GWAS studies identified10 candidate genes
  - Replicated in multiple studies
  - APOE/TOMM40, ABCA7, BIN1, CD2AP, CD33, CLU, CR1, EPHA1, MS4A4/MS4A6A, PICALM
  - 66 candidate SNPs
- Quantitative phenotypes
  - Imaging features (e.g. hippocampus volume)
  - CSF biomarkers

NEUROTEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

Leading Neuroscience Research

Example Associations

Gene	SNPs	Imaging Feature
ESR1 (Estrogen receptor 1)	rs9341052, rs9341052	Ventricular enlargement
BIN1 (Bridging integrator 1)	rs17014923, rs749008	Third-ventricle enlargement
SORCS1 (Sortilin-related VPS10 domain receptor 1)	rs6584777	Left-inferior lateral ventricle enlargement
EPHA1 (EPH receptor A1)	rs4726618	Left-inferior lateral ventricle enlargement

NEUROTEXAS INSTITUTE RESEARCH FOUNDATION St. David's HealthCare

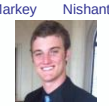
Leading Neuroscience Research

Acknowledgements

TACC


Mia Markey


Nishant Verma


Blake Printy

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

St. David's FOUNDATION IMPACT FUND

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

Leading Neuroscience Research

Matthew C. Cowperthwaite, PhD

matthew.cowperthwaite@stdavids.com

(512) 544-8059

For more information:
www.ntiresearch.org

NEURO TEXAS INSTITUTE RESEARCH FOUNDATION
St. David's HealthCare

Leading Neuroscience Research