# Tutorial on Genome-Wide Association Studies

William S. Bush PhD MS

Assistant Professor
Institute for Computational Biology
Department of Epidemiology and Biostatistics
Case Western Reserve University

School of Medicine
Case Western Reserve University

Institute for Computational Biology

# Acknowledgements

- Dana Crawford
- Holli Dilks-Hutchinson
- Marylyn Ritchie

# Key References

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

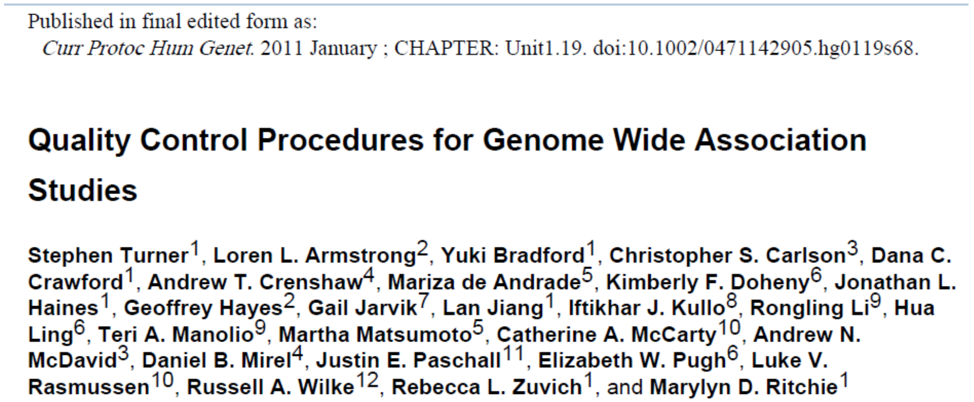
Education

## Chapter 11: Genome-Wide Association Studies

William S. Bush[1]*, Jason H. Moore[2]

[1] Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, Tennessee, United States of America, [2] Departments of Genetics and Community Family Medicine, Institute for Quantitative Biomedical Sciences, Dartmouth Medical School, Lebanon, New Hampshire, United States of America

- www.ploscollections.org/translationalbioinformatics



Published in final edited form as:
*Curr Protoc Hum Genet.* 2011 January ; CHAPTER: Unit1.19. doi:10.1002/0471142905.hg0119s68.

## Quality Control Procedures for Genome Wide Association Studies

Stephen Turner[1], Loren L. Armstrong[2], Yuki Bradford[1], Christopher S. Carlson[3], Dana C. Crawford[1], Andrew T. Crenshaw[4], Mariza de Andrade[5], Kimberly F. Doheny[6], Jonathan L. Haines[1], Geoffrey Hayes[2], Gail Jarvik[7], Lan Jiang[1], Iftikhar J. Kullo[8], Rongling Li[9], Hua Ling[6], Teri A. Manolio[9], Martha Matsumoto[5], Catherine A. McCarty[10], Andrew N. McDavid[3], Daniel B. Mirel[4], Justin E. Paschall[11], Elizabeth W. Pugh[6], Luke V. Rasmussen[10], Russell A. Wilke[12], Rebecca L. Zuvich[1], and Marylyn D. Ritchie[1]

# Overview

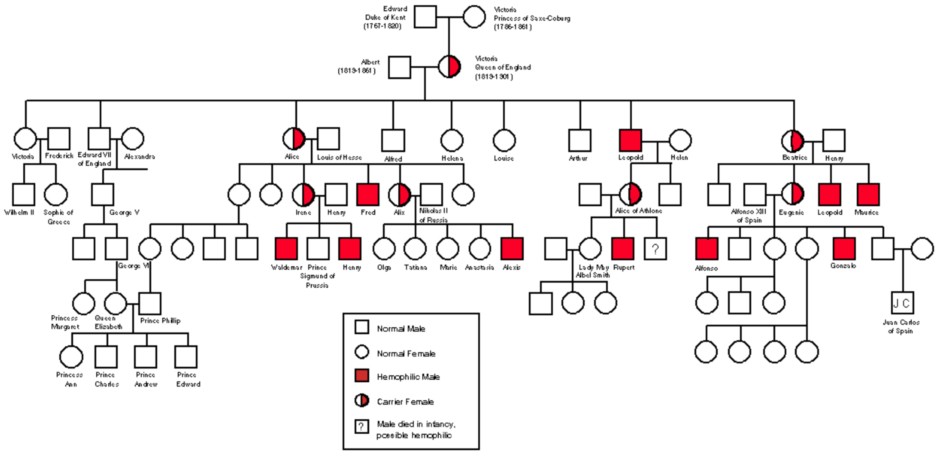
- Common Study Designs for GWAS
- Quality Control Procedures for GWAS Data
- Statistical Analysis
- Replication

# Goals of a Genetic Study

- Determine if there is a genetic component (heritability)
- Describe mode of inheritance (segregation analysis)
- Determine the effect size of the genetic component
- Identify the gene causing the disease

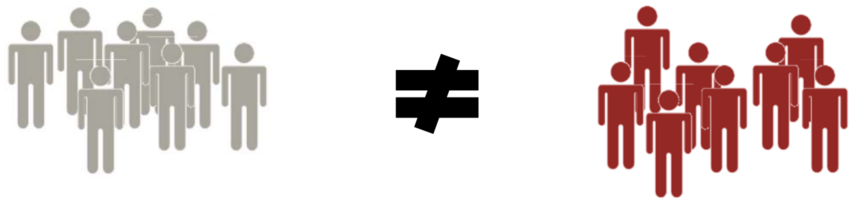
# Family Studies

- Allows estimation of genetic component
- Allows examination of mode of inheritance
- Difficult to collect
- Power derived from the number of families



# Population Studies

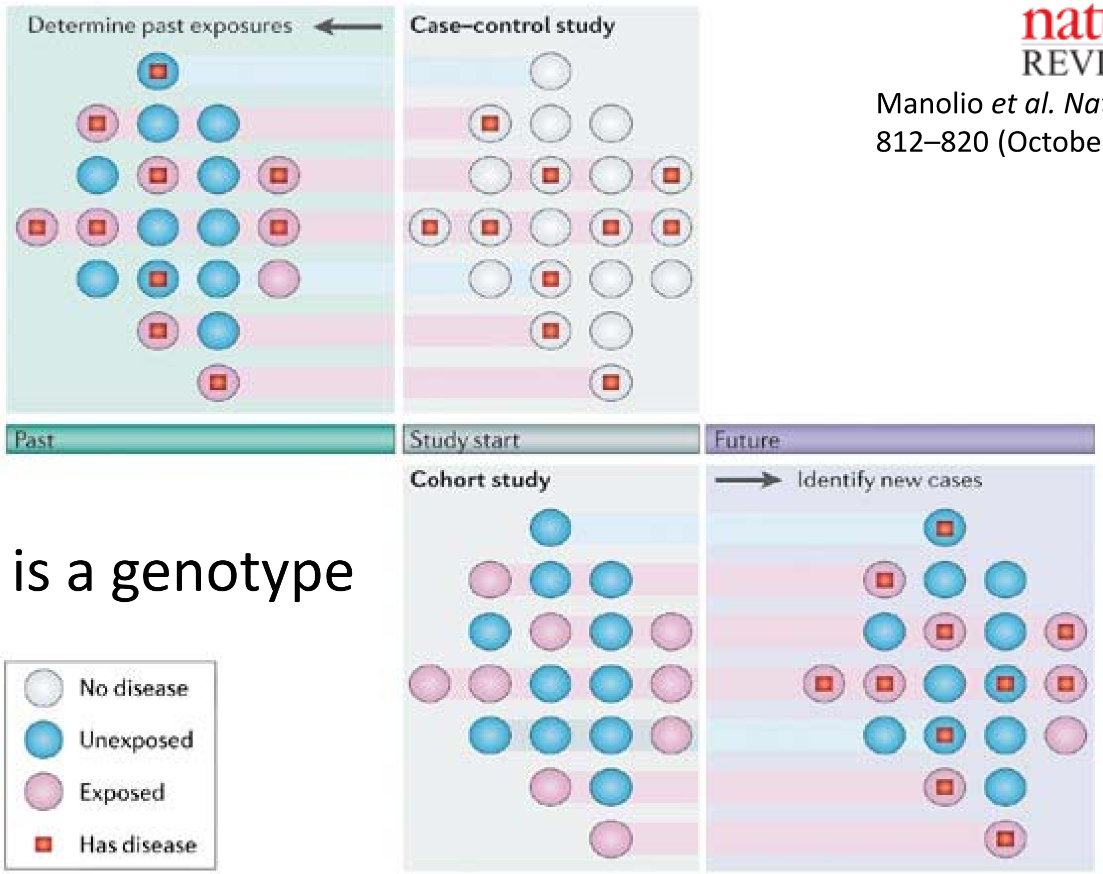
- Easier to collect
- Larger sample sizes
- Assumes there is a genetic component
- Power derived from ratio of controls to cases



# Retrospective vs. Prospective

nature REVIEWS GENETICS

Manolio *et al. Nature Reviews Genetics* 7, 812–820 (October 2006)



Exposure is a genotype

Nature Reviews | Genetics

# Choosing a Study Design

- What samples are available?
- Is a genetic component known?
- Details of the trait being studied (age at onset, disease frequency, penetrance, etc.)
- Interest in other factors of disease (environmental exposures, survival, effect size)

# Choosing a Study Design

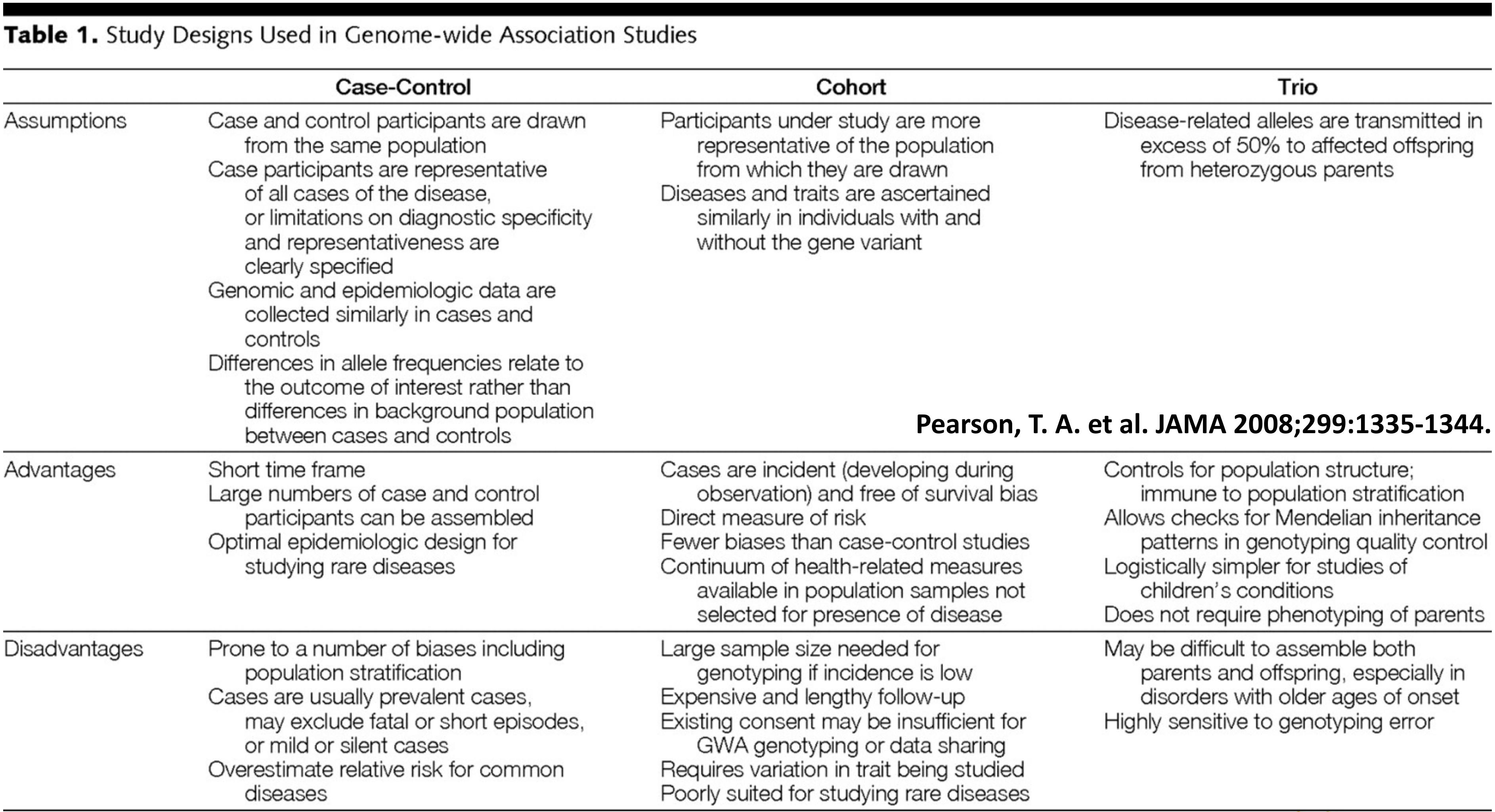


**Table 1.** Study Designs Used in Genome-wide Association Studies

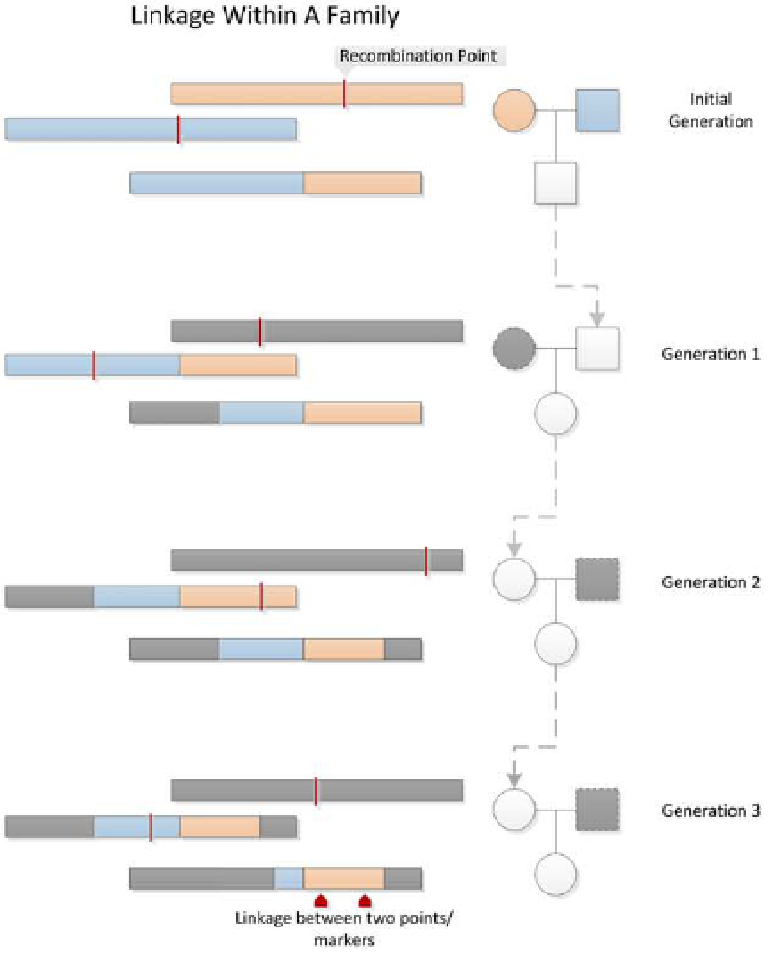
| | Case-Control | Cohort | Trio |
|---|---|---|---|
| Assumptions | Case and control participants are drawn from the same population<br>Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified<br>Genomic and epidemiologic data are collected similarly in cases and controls<br>Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls | Participants under study are more representative of the population from which they are drawn<br>Diseases and traits are ascertained similarly in individuals with and without the gene variant | Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents |
| Advantages | Short time frame<br>Large numbers of case and control participants can be assembled<br>Optimal epidemiologic design for studying rare diseases | Cases are incident (developing during observation) and free of survival bias<br>Direct measure of risk<br>Fewer biases than case-control studies<br>Continuum of health-related measures available in population samples not selected for presence of disease | Controls for population structure; immune to population stratification<br>Allows checks for Mendelian inheritance patterns in genotyping quality control<br>Logistically simpler for studies of children's conditions<br>Does not require phenotyping of parents |
| Disadvantages | Prone to a number of biases including population stratification<br>Cases are usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases<br>Overestimate relative risk for common diseases | Large sample size needed for genotyping if incidence is low<br>Expensive and lengthy follow-up<br>Existing consent may be insufficient for GWA genotyping or data sharing<br>Requires variation in trait being studied<br>Poorly suited for studying rare diseases | May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset<br>Highly sensitive to genotyping error |

**Pearson, T. A. et al. JAMA 2008;299:1335-1344.**

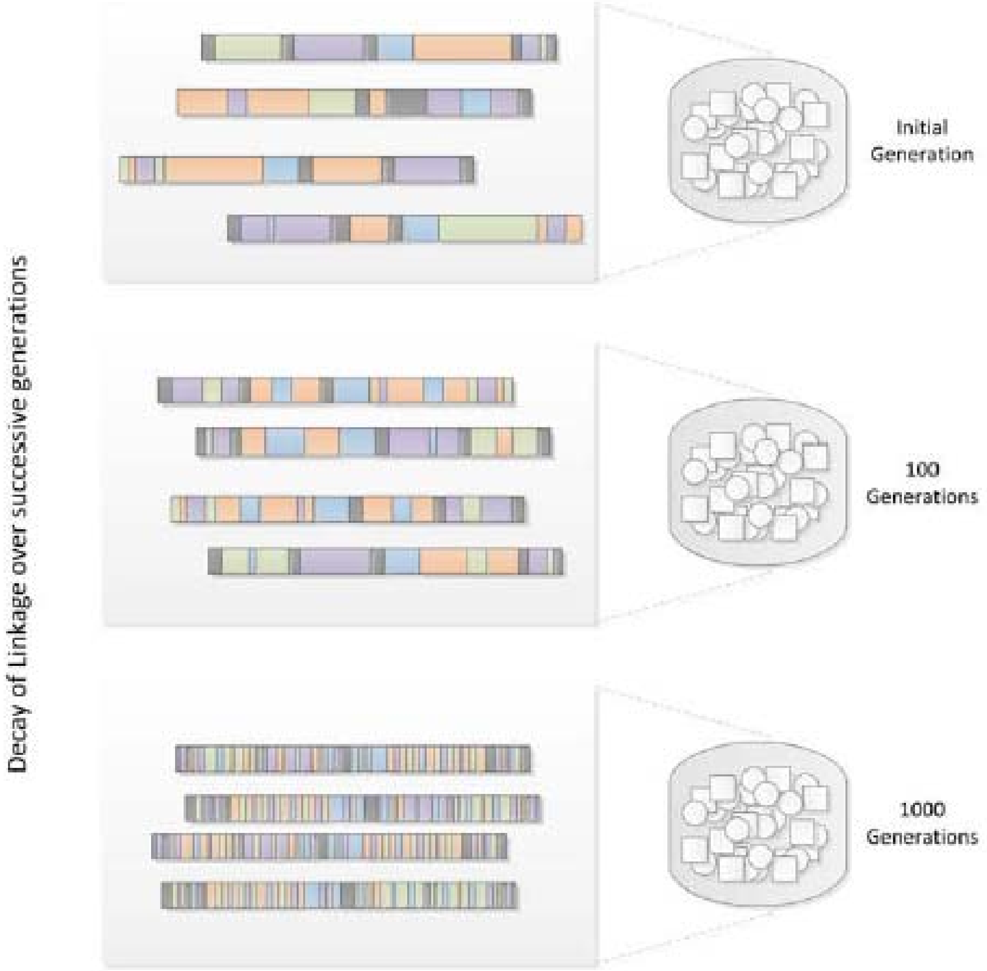
# Assumptions of GWAS

- Examines *only* the Common Disease – Common Variant hypothesis
- Relies on dense sets of genetic markers
- Exploits linkage disequilibrium to make "indirect associations"
- Goal: Identify markers with significant associations to disease

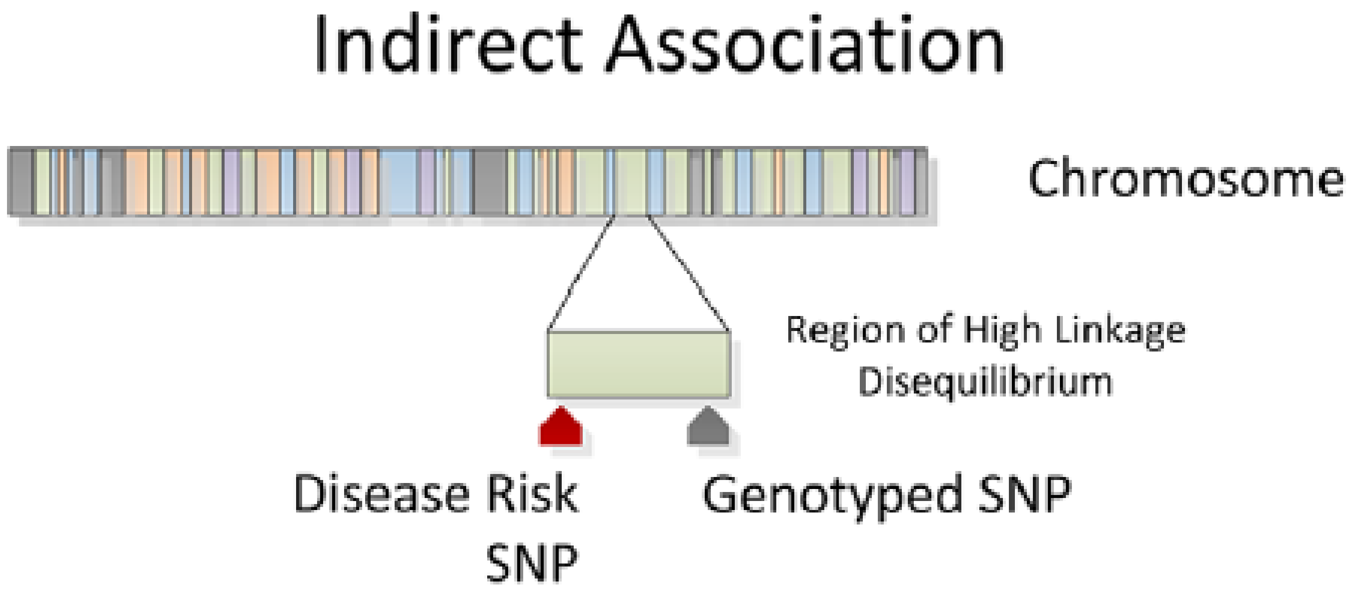
# Recombination



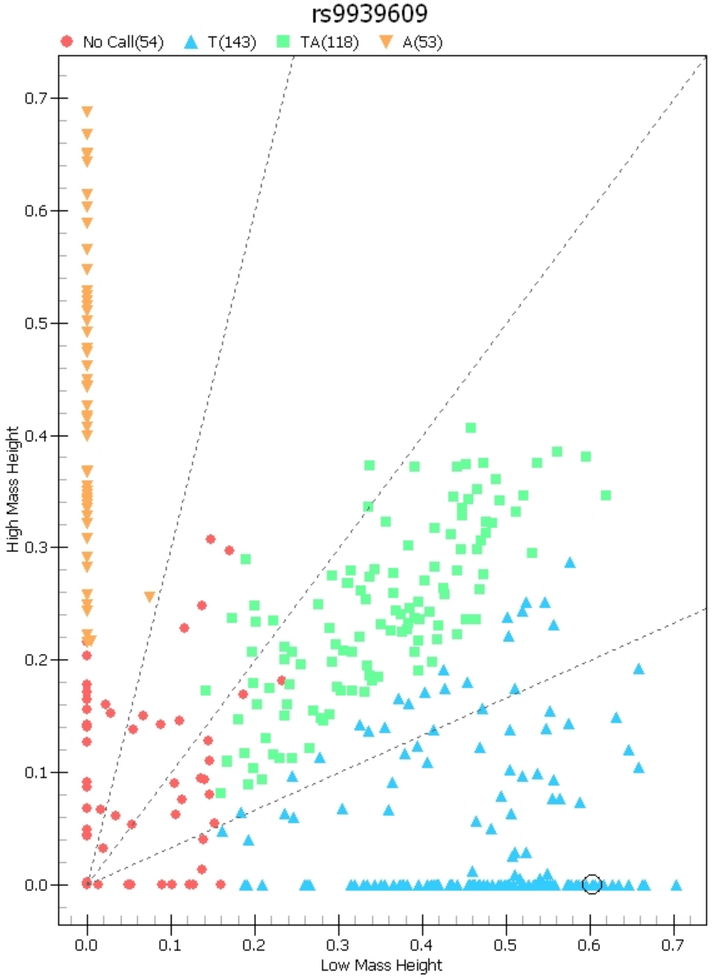
# Linkage Disequilibrium

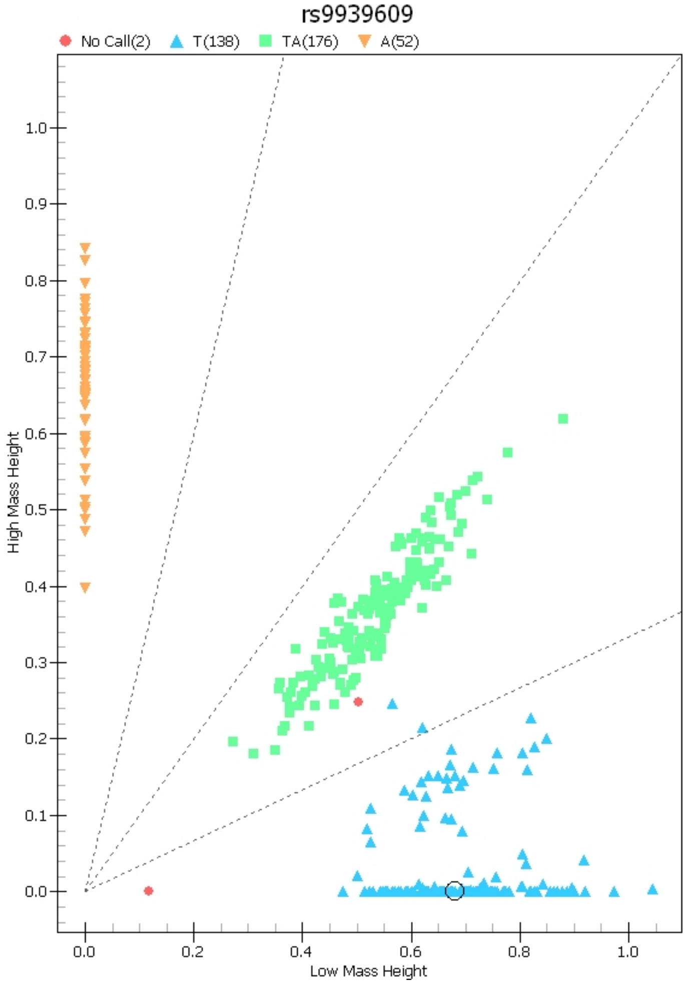


# Exploiting Linkage Disequilibrium



# How GWAS Data is Generated

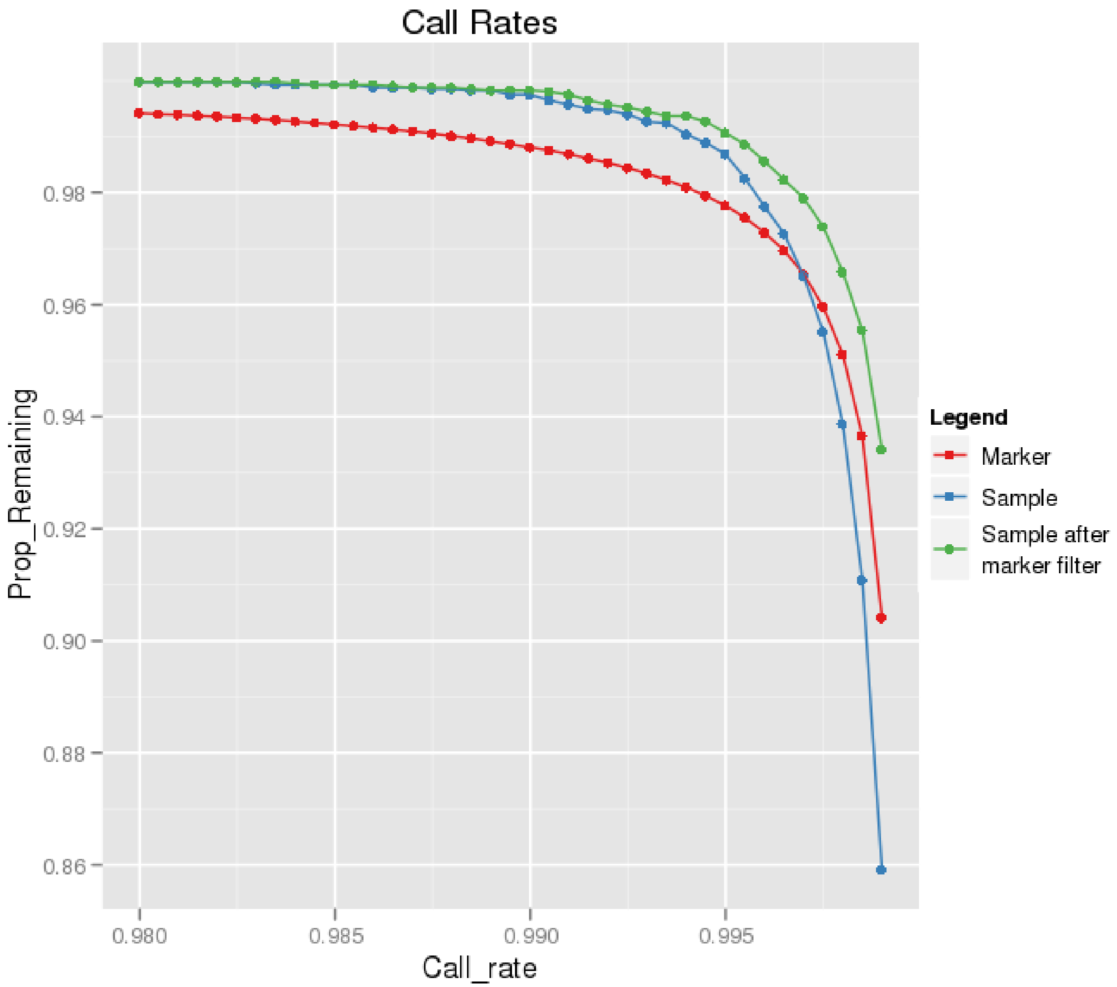




# Quality Control of GWAS Data

| Variable | Comments |
|---|---|
| Genotyping Call Rate | Low call rate often correlates with error. Some low call rate SNPs or samples may still be good. |
| Genotyping Quality | Worse quality score (GenCall) correlates strongly with error rate |
| Sex concordance | Check expectations for X marker heterozygosity and Y marker positive results. Can estimate error rate. |
| Sample Relatedness | Check for related samples (expected or unexpected) |
| Mendelian Inheritance Errors | For trio/family data, can identify problem samples and families. Can estimate error rate. |
| Replicate concordance | Check for consistent genotype calls in duplicate samples |
| Batch effects | Check for genotyping call differences due to plate |
| Hardy-Weinberg Equilibrium | Violation across all sample groups may indicate error, but can also be a good test of association |
| Population Stratification | Check for population substructure using the genome-wide data |

# Marker and Sample Call Rate



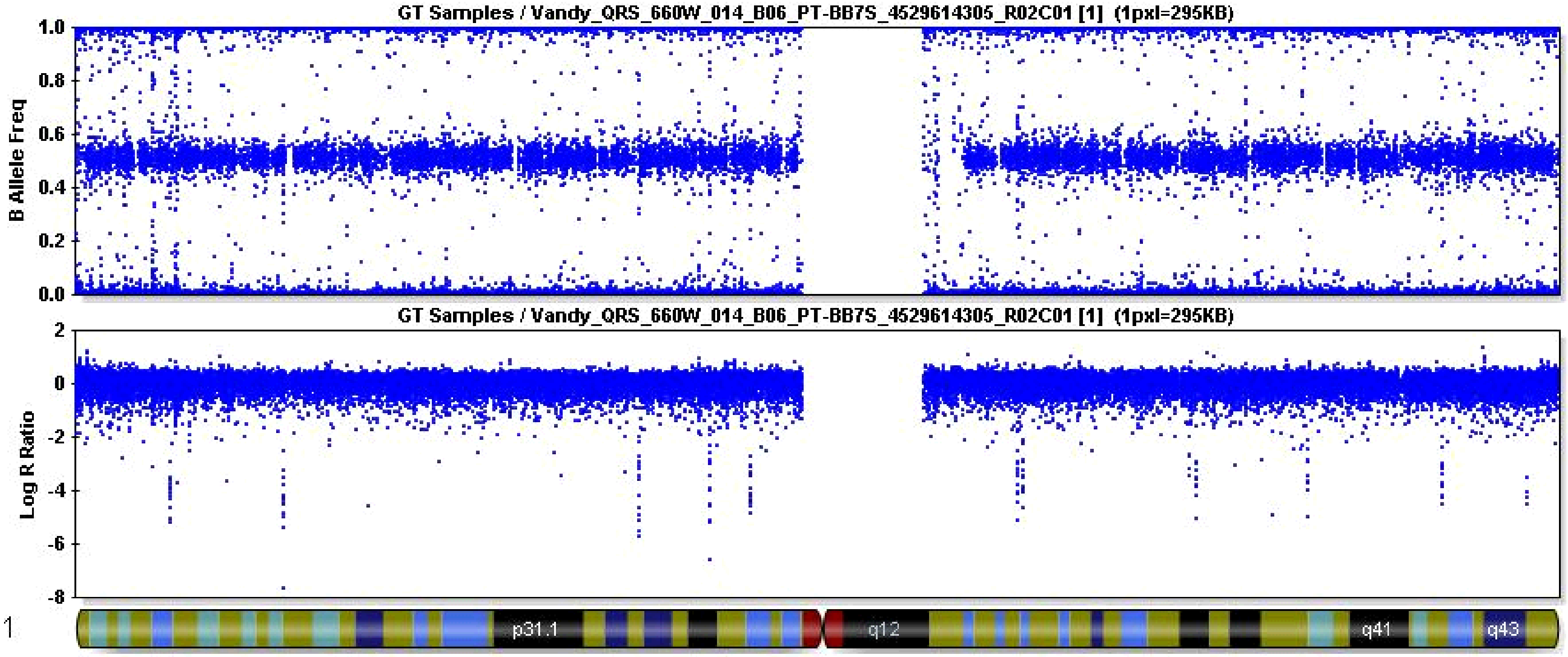
# Sex Concordance Check

| emerge_id | Pedsex | SNPsex | PLINK_F | Note |
|---|---|---|---|---|
| 16230834 | 2 | 0 | 0.4746 | CIDR comment after review of B allele freq and Log R ratio plots for all chromosomes: This sample has large loss-of-heterozygosity (LOH) blocks on X (and other autosomes). The sample is definitely female (2 X chromosomes by intensities). |
| 16228083 | 2 | 0 | 0.2654 | Same as above |
| 16231930 | 2 | 0 | 0.4376 | Same as above |
| 16233764 | 2 | 0 | 0.2603 | Same as above |
| 16221112 | 2 | 0 | 0.2048 | XX/XO mosaic not caught by initial check completed by CIDR |
| 16222319 | 2 | 0 | 0.7452 | Annotation by CIDR at data release: Appears to be XX/XO mosaic |
| 16228204 | 2 | 1 | 1 | Annotation by CIDR at data release: Appears to be XX/XO mosaic |
| 16233113 | 1 | 0 | 0.4752 | Annotation by CIDR at data release: Appears to be XXY |
| 16214881 | 1 | 2 | 0.136 | Annotation by CIDR at data release: Appears to be XXY/XY mosaic |

- Female: pedsex=2, SNPsex=2
- Male: pedsex= 1, SNPsex=1
- A male call is made if the F (actual X chromosome inbreeding estimate) is more than 0.8; a female call is made if the F is less than 0.2.

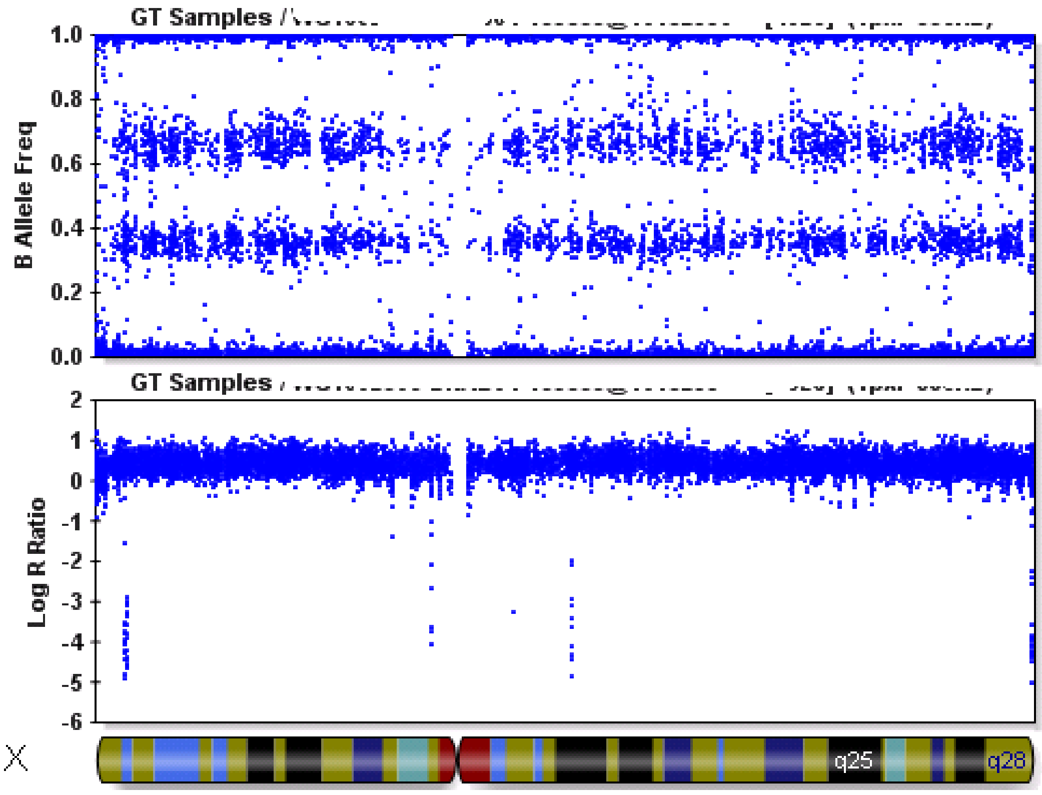
# Sex Concordance Check

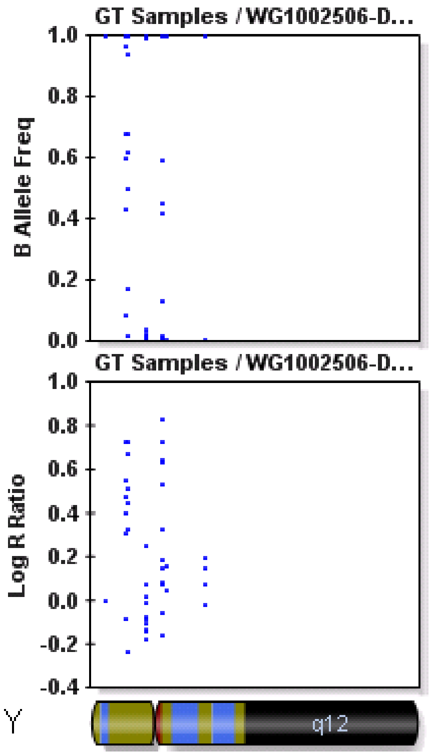
- Normal Chromosome 1



# Sex Concordance

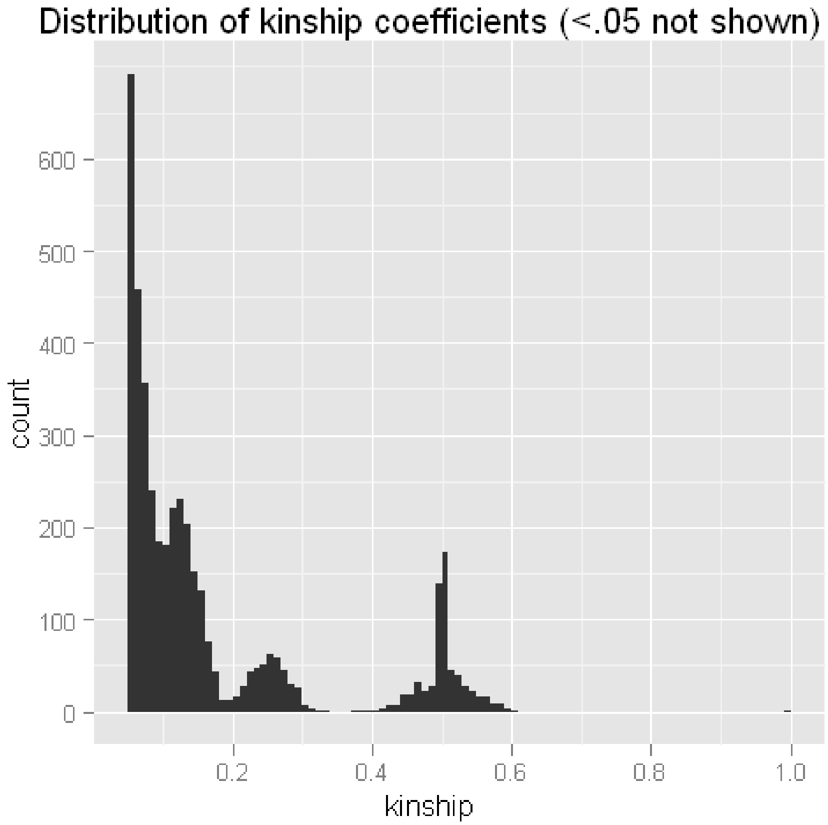
- Possible XXY/XY Mosaic



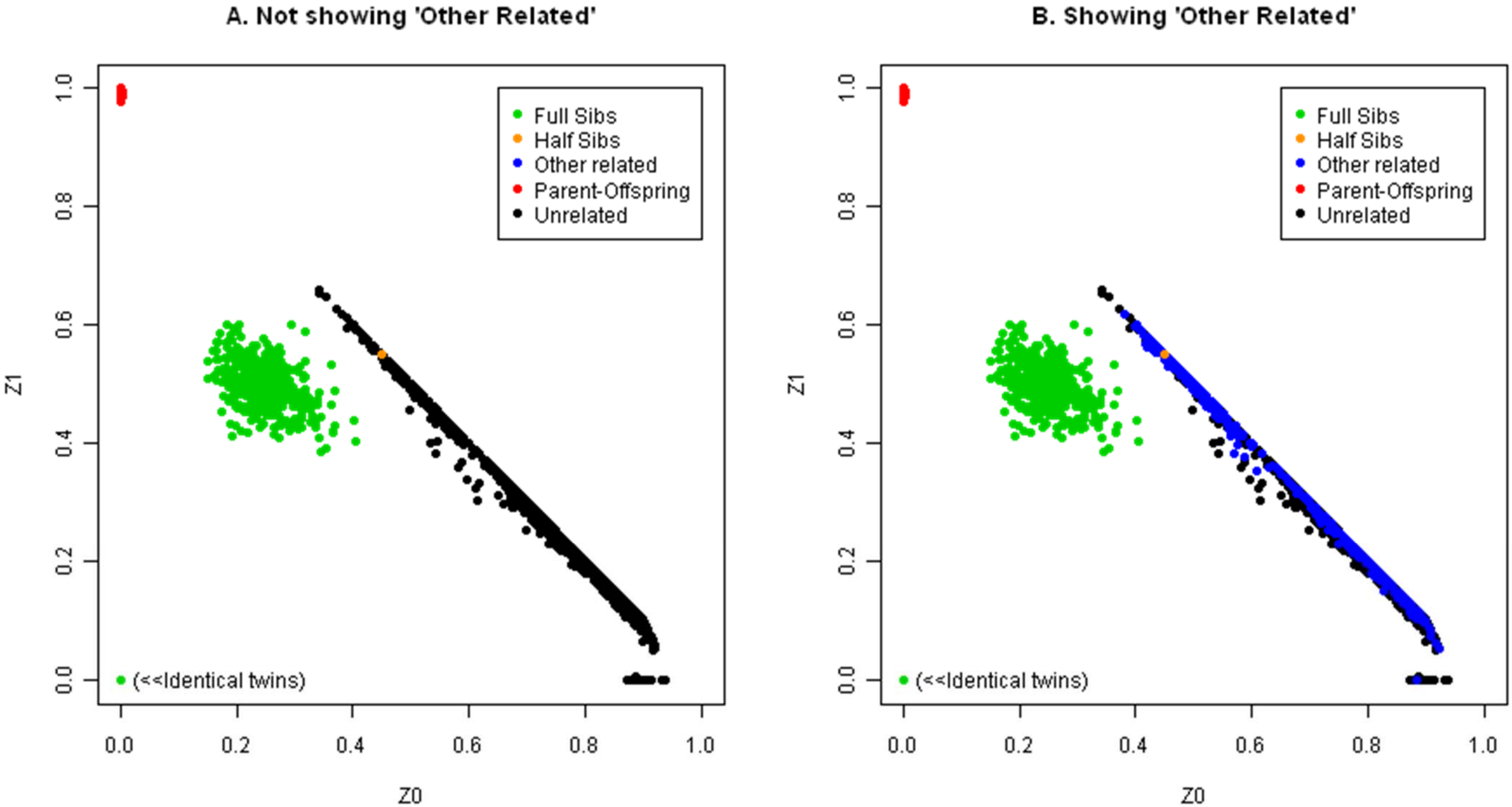


# Sample Relatedness

| Z0 | Z1 | Z2 | Kinship | Relationship |
|---|---|---|---|---|
| 0.0 | 0.0 | 1.0 | 1.0 | MZ twin or duplicate |
| 0.0 | 1.0 | 0.0 | 0.50 | Parent-offspring |
| 0.25 | 0.50 | 0.25 | 0.50 | Full siblings |
| 0.50 | 0.50 | 0.0 | 0.25 | Half siblings |
| 0.75 | 0.25 | 0.0 | 0.125 | Cousins |
| 1.0 | 0.0 | 0.0 | 0.0 | Unrelated |



# Sample Relatedness


A. Not showing 'Other Related'

B. Showing 'Other Related'

http://www.sph.umich.edu/csg/abecasis/publications/11524377.html

# Mendelian Inheritance Errors

- Even with Case/Control data, HapMap trios are typically plated with study samples for QC

| Number Mendelian Errors | Number SNPs pre QC | Number SNPs post marker QC |
|---|---|---|
| 0 | 558821 | 552346 |
| 1 | 1519 | 1353 |
| 2 | 97 | 64 |
| 3 | 5 | 1 |

# Sample Replicate Concordance

| emerge | Samp1 | samp2 | discordant | total | concordance_rate |
|---|---|---|---|---|---|
| 16231453 | A | B | 171 | 558882 | 0.99969 |
| 16223704 | A | B | 137 | 557783 | 0.99975 |
| 16216270 | A | B | 133 | 559711 | 0.99976 |
| 16230108 | A | B | 69 | 559341 | 0.99987 |
| 16224359 | A | B | 67 | 558868 | 0.99988 |
| 16234120 | A | B | 43 | 560202 | 0.99992 |
| 16232463 | A | B | 42 | 560355 | 0.99992 |
| 16234233 | A | B | 33 | 560384 | 0.99994 |
| 16216349 | A | B | 30 | 559345 | 0.99994 |
| 16215309 | A | B | 12 | 560041 | 0.99997 |
| 16224779 | A | B | 7 | 560412 | 0.99998 |
| 16231724 | A | B | 5 | 560427 | 0.99999 |
| 16233841 | A | B | 4 | 560519 | 0.99999 |
| 16221647 | A | B | 2 | 560457 | 0.99999 |
| 16230404 | A | B | 2 | 560309 | 0.99999 |
| 16226433 | A | B | 2 | 560500 | 0.99999 |
| 16234367 | A | B | 2 | 560373 | 0.99999 |
| 16224635 | A | B | 1 | 560560 | 0.99999 |
| 16219214 | A | B | 1 | 560535 | 0.99999 |
| 16231219 | A | B | 1 | 560547 | 0.99999 |
| 16220060 | A | B | 0 | 560580 | 1 |

# Hardy Weinberg Equilibrium

## All individuals

| threshhold | below | exp_below | excess_below |
|---|---|---|---|
| 0.05 | 37690 | 28022 | 9668 |
| 0.01 | 12774 | 5604 | 7170 |
| 0.001 | 4766 | 560 | 4206 |
| 1.00E-04 | 2949 | 56 | 2893 |
| 1.00E-05 | 2337 | 5 | 2332 |
| 1.00E-06 | 2004 | 0 | 2004 |
| 1.00E-07 | 1785 | 0 | 1785 |

## All cases

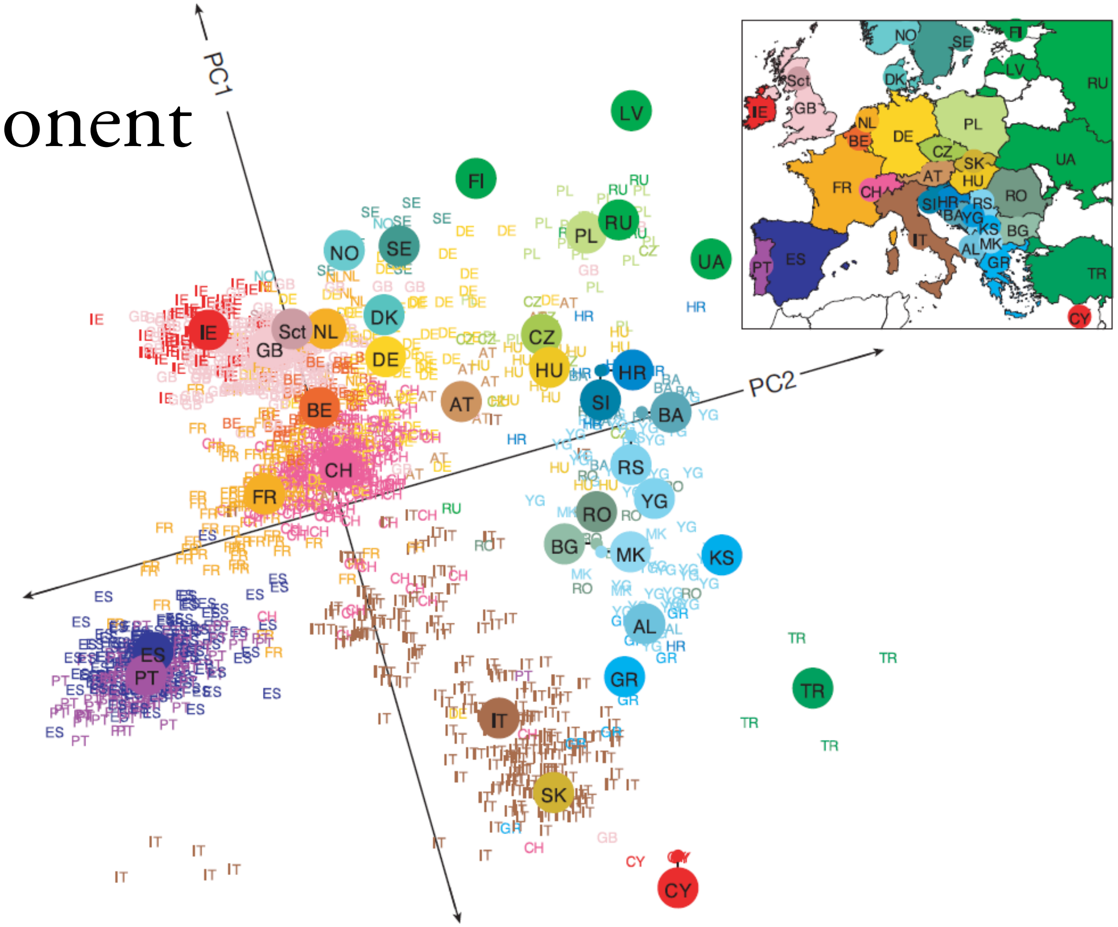
| threshold | below | exp_below | excess_below |
|---|---|---|---|
| 0.05 | 34646 | 28022 | 6624 |
| 0.01 | 10843 | 5604 | 5239 |
| 0.001 | 3642 | 560 | 3082 |
| 1.00E-04 | 2194 | 56 | 2138 |
| 1.00E-05 | 1792 | 5 | 1787 |
| 1.00E-06 | 1563 | 0 | 1563 |
| 1.00E-07 | 1394 | 0 | 1394 |

## All controls

| threshold | below | exp_below | excess_below |
|---|---|---|---|
| 0.05 | 30557 | 28022 | 2535 |
| 0.01 | 8859 | 5604 | 3255 |
| 0.001 | 2614 | 560 | 2054 |
| 1.00E-04 | 1517 | 56 | 1461 |
| 1.00E-05 | 1180 | 5 | 1175 |
| 1.00E-06 | 982 | 0 | 982 |
| 1.00E-07 | 860 | 0 | 860 |

# Population Stratification

- Principal Component Analysis (PCA)
- Can cause confounding



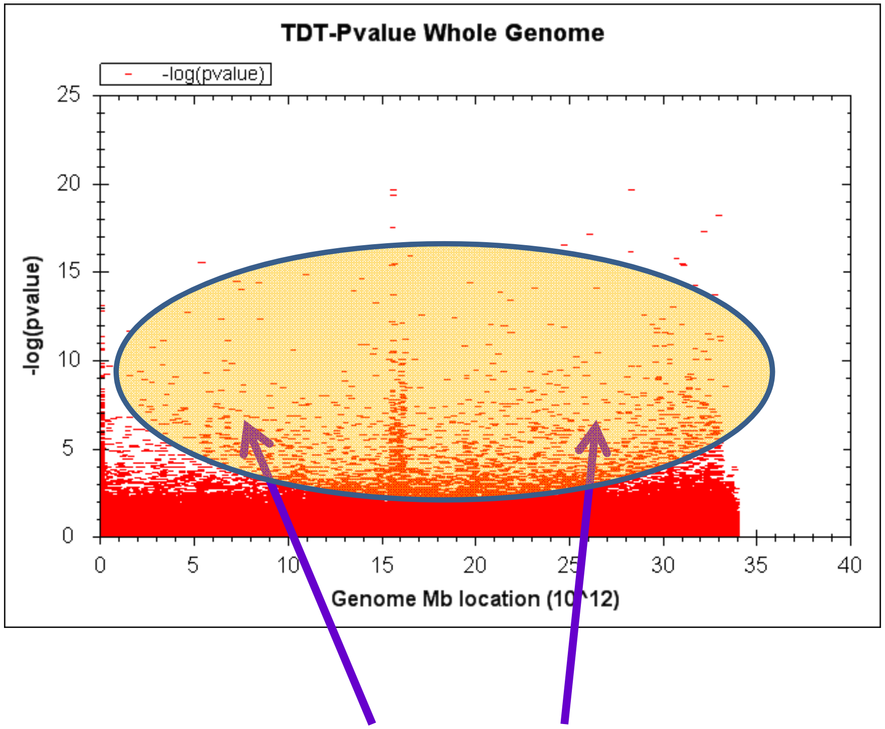
Genes Mirror Geography in Europe, Novembre, Nature Genetics, 2008

# Batch Effects

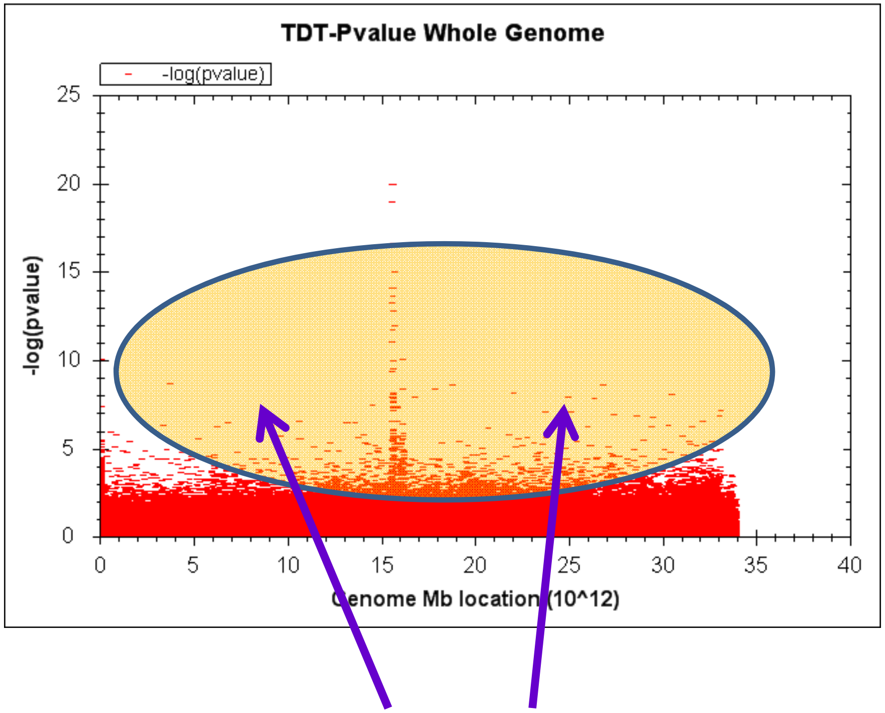
- Evidence that associations can result due to allele frequency difference due to plate effects
- Careful consideration when creating plate maps
  - Plate cases and controls together
  - Randomize by race, gender, age, BMI, others…
- After genotyping look for plate effects
  - MAF differences by plate
  - Call rate by plate
  - Association tests (one plate versus all others)

# Importance of QC

**Pre-QC Thresholds**



**Post-QC Thresholds**



Many false positives disappear after QC

# GWAS Analysis

- Consider 500,000 SNPs across the human genome
- Each SNP has its own statistical test
- Each SNP has a different statistical power (depending on allele frequency)

# Analysis of GWAS Data

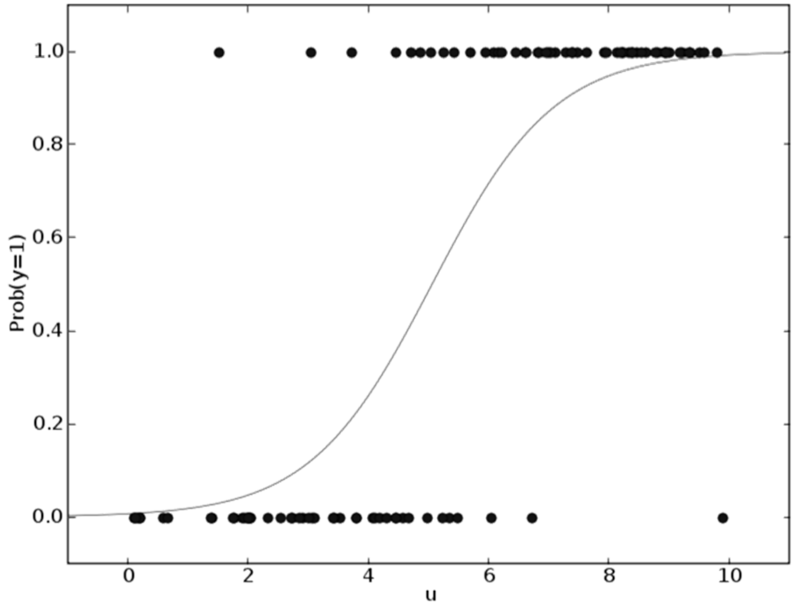
- Basic Statistical Methods are usually applied
- Linear Regression (continuous trait)
- Logistic Regression (dichotomous trait)
- Adjustments are critical to avoid confounding

# Software Tools

- PLINK - http://pngu.mgh.harvard.edu/~purcell/plink/
- PLATO – https://ritchielab.psu.edu/software/plato-download
- R – http://www.r-project.org/
  - Bioconductor - http://www.bioconductor.org/

# Logistic Regression

- Examines differences between two groups (cases and controls)
- Transforms the Y-Axis of a typical regression using a logit function
- Produces a probability of case or control status (Odds Ratio)



# QQ Plot

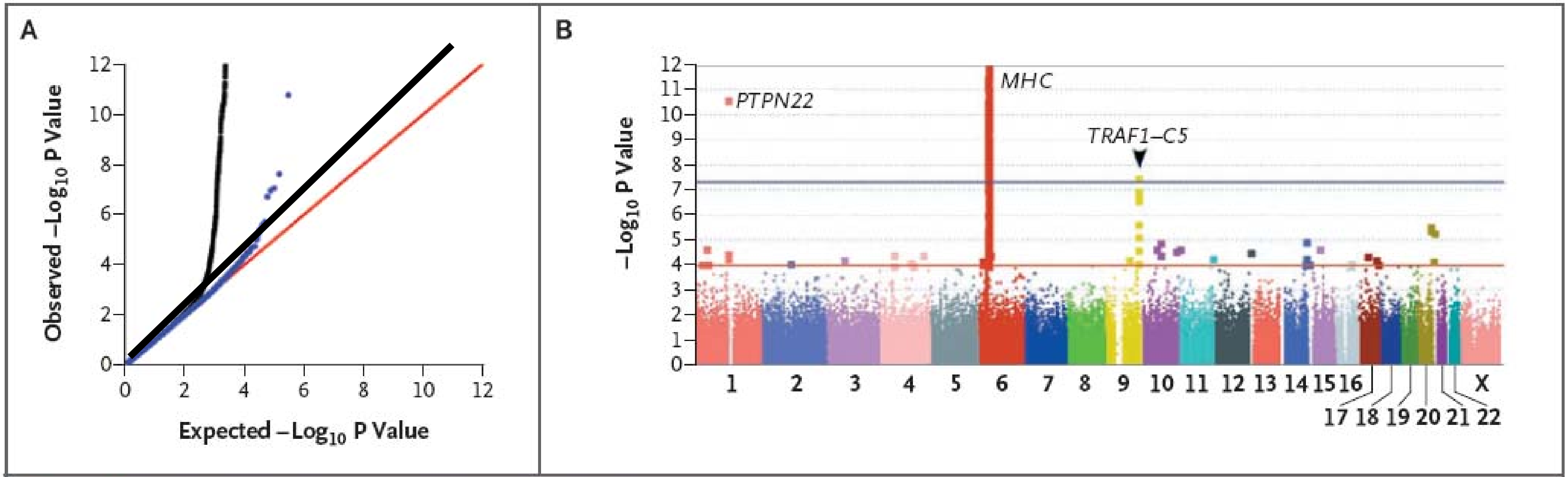


Figure 1. Results of the Genomewide Association Study.

- Systematic deviation from the line indicates population stratification / genomic inflation

# Multiple Testing

- Perform 500,000 analyses
  - Type I error set at 5%, we can expect 25,000 false positive results
    - Bonferroni correction
    - False Discovery Rate (FDR)
    - Gene-based correction (principal components)
- “Genome-wide significance” is $p<10^{-8}$
  - Can be problematic for non-European Populations

# Winner's Curse

- Consider an item with a fixed value (pashmina)
- If there are ten American tourists bidding on the same item, the bids will average around the item's true value
- By definition, the winner will ALWAYS overpay (Dana)

# Winner's Curse in GWAS

- Similarly when running a GWAS and discovering a SNP association, you will OVERESTIMATE the strength of the association
- Power calculations use an effect size to know how many samples you need to detect this effect
- If the effect size is actually SMALLER than you think, you'll need MORE samples to see you effect again

# Replication

- Now required for consideration in top journals
- Second sample, preferably with larger sample sizes to increase power
- Ideally should be interchangeable with the first sample in every way
  - Need all the covariates you used in the first dataset

# Great GWAS Examples

- Multiple Sclerosis GWAS
  - Trio design, extensive QC
  - http://www.ncbi.nlm.nih.gov/pubmed/17660530
- Type II Diabetes GWAS
  - http://www.ncbi.nlm.nih.gov/pubmed/17463246?dopt=Abstract&holding=npg