National Human Genome Research Institute

National Institutes of Health

U.S. Department of Health and Human Services

Lecture 5: Why Can't Anyone Reproduce the Findings of My "Discovery" Study?

U.S. Department of Health and Human Services

National Institutes of Health

National Human Genome Research Institute

Teri A. Manolio, M.D., Ph.D.
Director, Office of Population Genomics
Senior Advisor to the Director, NHGRI,
for Population Genomics

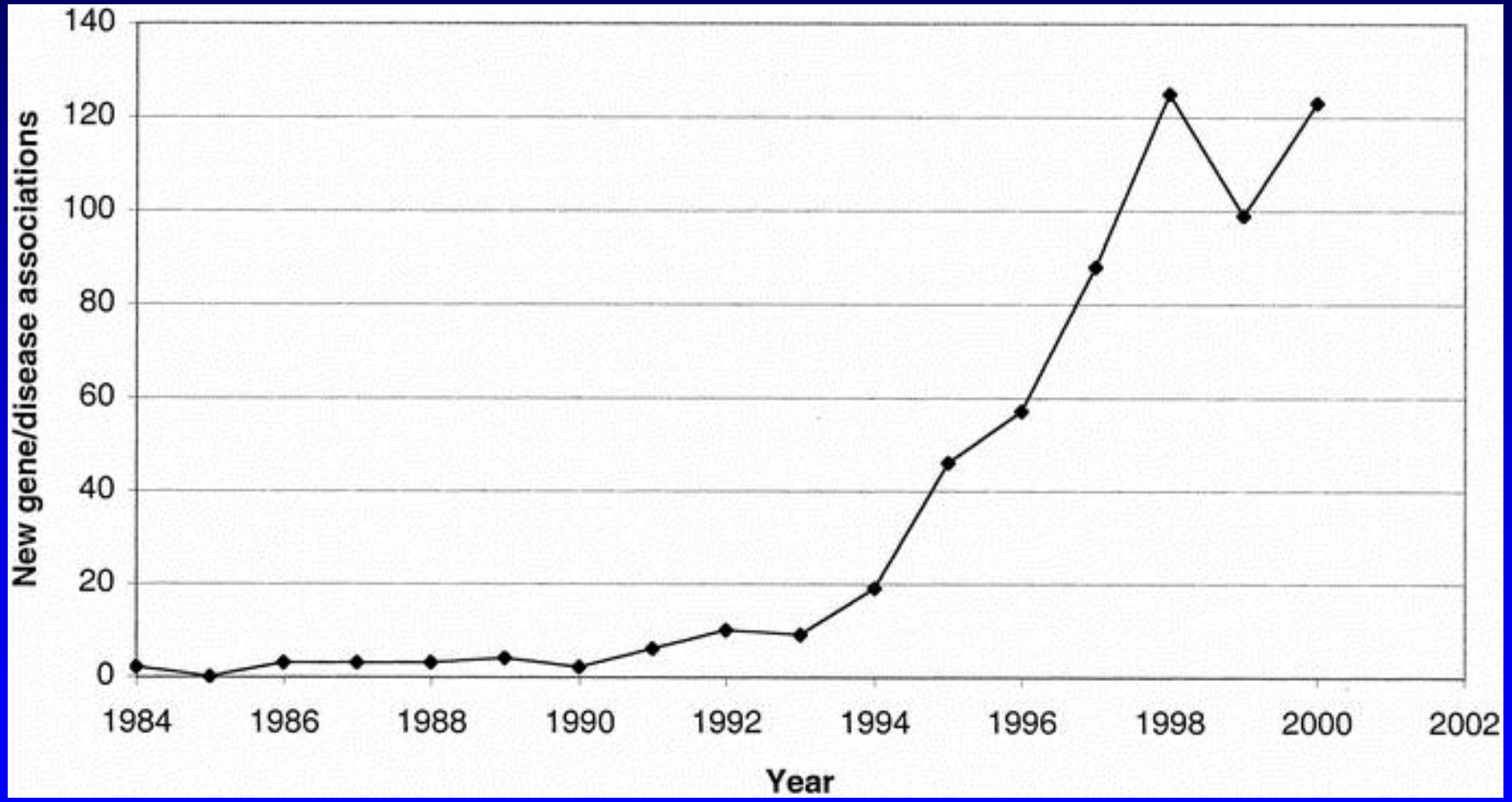
July 18, 2008

Sidney Harris, http://www.sciencecartoonsplus.com/gallery.htm.

Learning Objectives

- Understand why replication is important in genetic association studies
- Define and apply consensus criteria for replication
- Identify possible causes of failure to replicate
- Examine genotyping data quality and its potential role in failure to replicate

Number of New, Significant Gene-Disease Associations by Year, 1984 - 2000



Hirschhorn J et al., *Genet Med* 2002; 4:45-61.

Of 600 Gene-Disease Associations, Only 6 Significant in ≥ 75% of Identified Studies

Disease/Trait	Gene	Polymorphism	Frequency
DVT	F5	Arg506Gln	0.015
Graves' Disease	CTLA4	Thr17Ala	0.62
Type 1 DM	INS	5' VNTR	0.67
HIV/AIDS	CCR5	32 bp Ins/Del	0.05-0.07
Alzheimer's	APOE	Epsilon 2/3/4	0.16-0.24
Creutzfeldt-Jakob Disease	PRNP	Met129Val	0.37

Hirschhorn J et al., *Genet Med* 2002; 4:45-61.

Reports For and Against Associations of Variants with Carotid Atherosclerosis

POLYMORPHISM	PRESENT	ABSENT	SUMMARY
ACE I/D	13 with D; 1 with I	18	favors none
Apo E	8 with ε4, 2 with ε2	9	equivocal
AGT M235T	0	8	none
AGTR1 A1166C	0	7	none
MTHFR	7 with T, 1 with non-T	8	equivocal
PON 1 Q192R	3 with R	10	none
PON 1 L55M	5 with L (subgroups)	1	weak
NOS 3 G894T	1 with T	4	none
MMP3 -1516 5A/6A	4 with 6A	0	association
IL-6 G-174C	1 with G	3	none

Manolio et al., *ATVB* 2004; 24:1567-77.

nature genetics
volume 22 no. 1 may 1999

Freely associating

May 1999

Editorial: Once and Again—Issues Surrounding Replication in Genetic Association Studies

PERSPECTIVE
The Future of Association Studies: Gene-Based Analysis and Replication

B

Editorial

Replication Publication

Statistical false positive or true disease pathway?

John A Todd

Nat Genet July 2006

Need for Consensus on What Constitutes Replication

- Replication held as *sine qua non*
- Multiple approaches to replication: functional studies, fine mapping, etc
- Avalanche of GWA and candidate gene studies now and in near future
- Likelihood of single study establishing an association is low until sample sizes increase sufficiently and analytical methods improve substantially
- Common problem of how to interpret confusing and spurious findings

NCI/NHGRI Replication Working Group
November 10, 2006

Goncalo Abecasis
David Altshuler
Joan E. Bailey-Wilson
Michael Boehnke
Eric Boerwinkle
Lisa D. Brooks
Lon R. Cardon
Stephen Chanock
Francis S. Collins
Mark Daly
Peter Donnelly
Joseph F. Fraumeni
Nelson B. Freimer
Daniela S. Gerhard
Chris Gunter
Alan E. Guttmacher
Mark S. Guyer
Joel Hirschhorn
Josephine Hoh
Robert Hoover
David Hunter
C. Augustine Kong
Teri Manolio
Kathleen R. Merikangas
Cynthia C. Morton
Lyle J. Palmer
John P. Rice
Jerry Roberts
Charles Rotimi
Gilles Thomas
Kyle Vogan
Sholom Wacholder
Ellen M. Wijsman

NCI/NHGRI Replication Working Group November 10, 2006: Objectives

To propose best practices for design, conduct and publication of replication studies to follow up notable findings, particularly in GWAs.

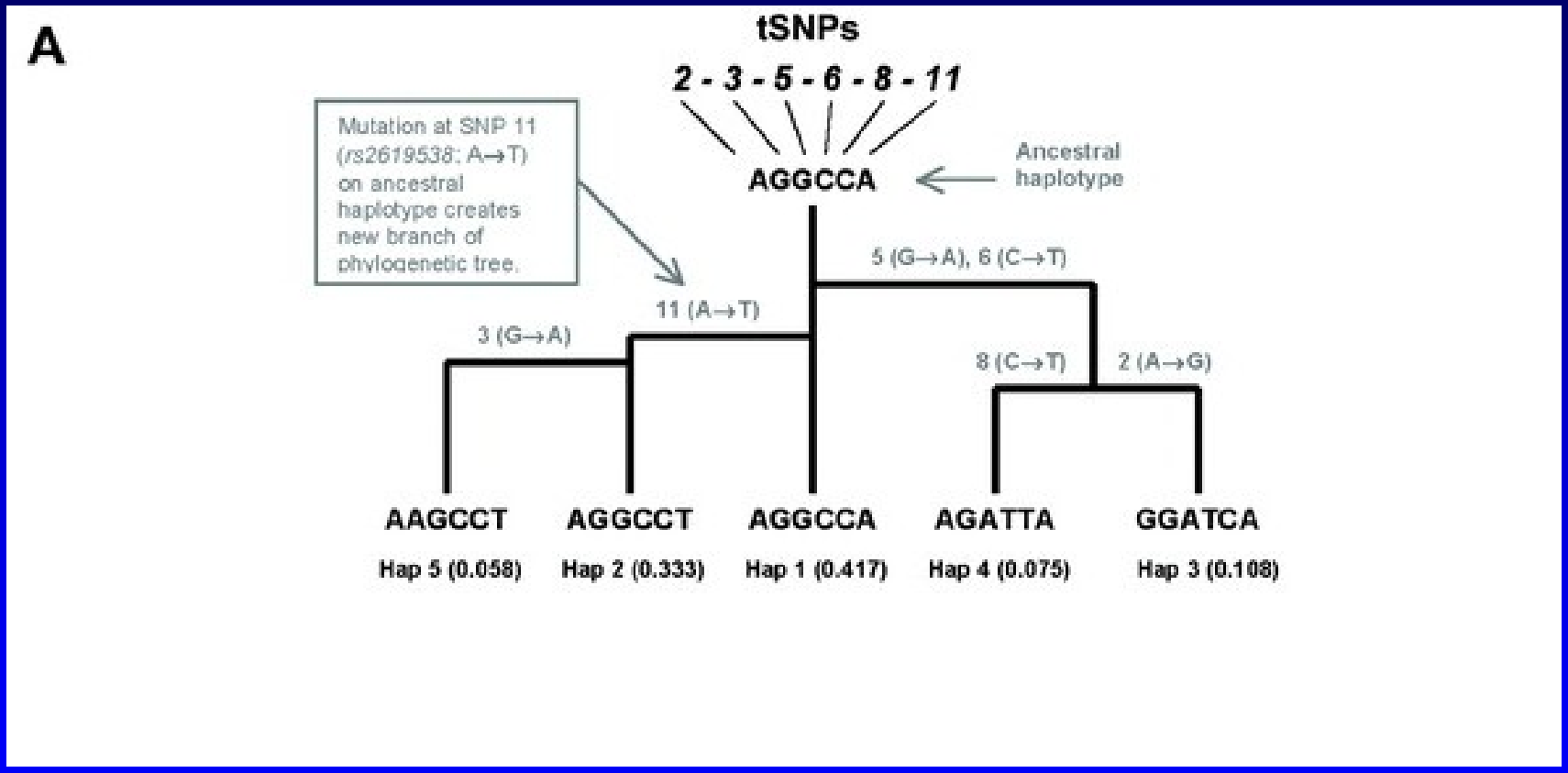
- Assess validity and limitations of any single genetic association study
- Define criteria for establishing replication in genetic association studies
- Develop “points to consider” for publication of high quality genotype-phenotype association reports

Case in Point: *DTNBP1* and Schizophrenia

- First identified as putative schizophrenia-susceptibility gene in Irish pedigrees
- Reported confirmation in several replication studies in independent European samples but reported risk alleles and haplotypes appeared to differ between studies
- Comparison among studies difficult because different marker sets used by each group
- HapMap data and all identified polymorphisms typed in CEPH samples to produce high density reference map

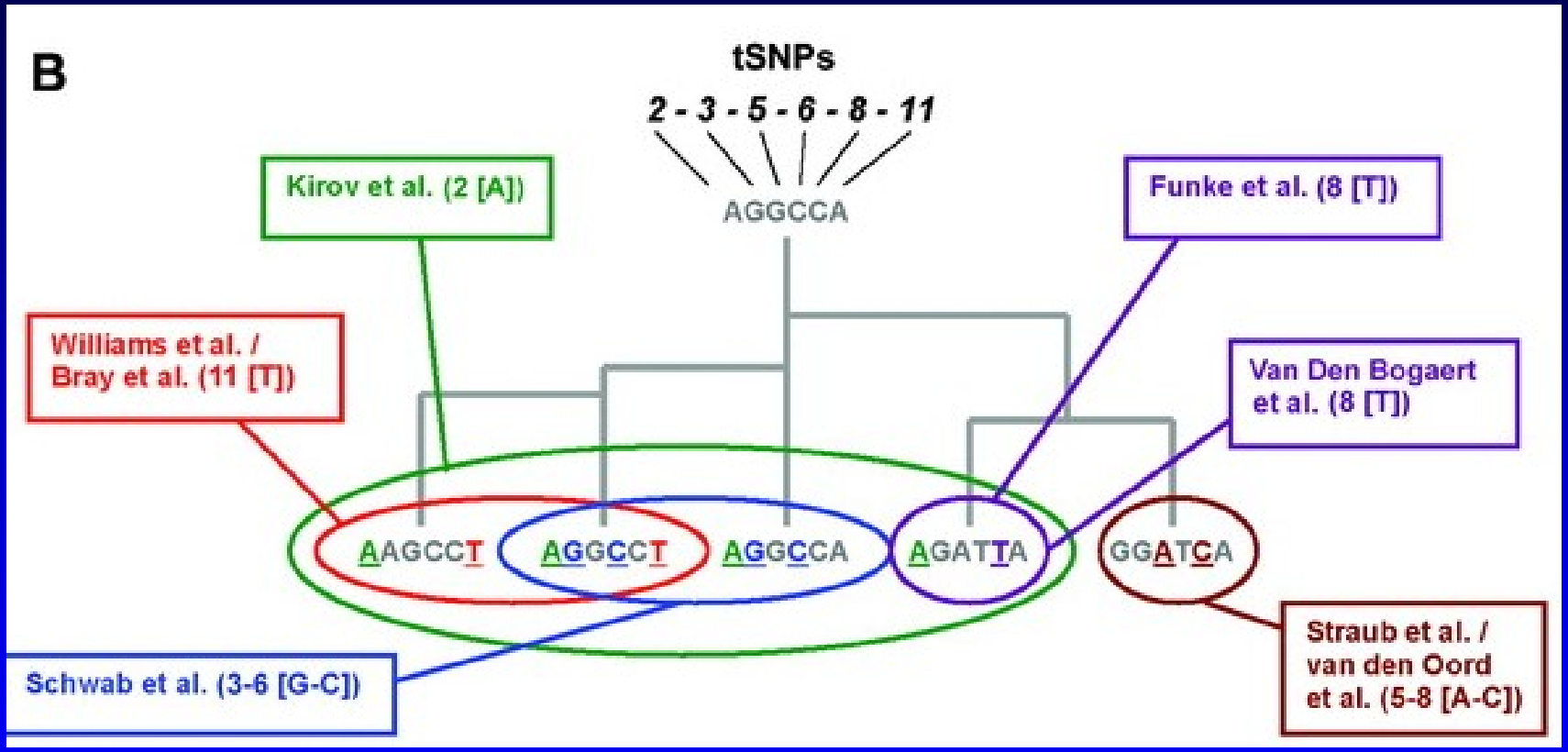
Mutsuddi et al., *Am J Hum Genet* 2006; 79:903-909.

Phylogenetic Tree of Five Common Haplotypes of *DTNBP1*



Mutsuddi et al., *Am J Hum Genet* 2006; 79:903-909.

Positively Associated Haplotypes Differ in All Six Studies



Each common DTNBP1 haplotype was tagged by association signal of at least one study, implying there is not one common variant contributing to schizophrenia risk at DTNBP1 locus

Mutsuddi et al., *Am J Hum Genet* 2006; 79:903-909.

How NOT To Do A Replication Study

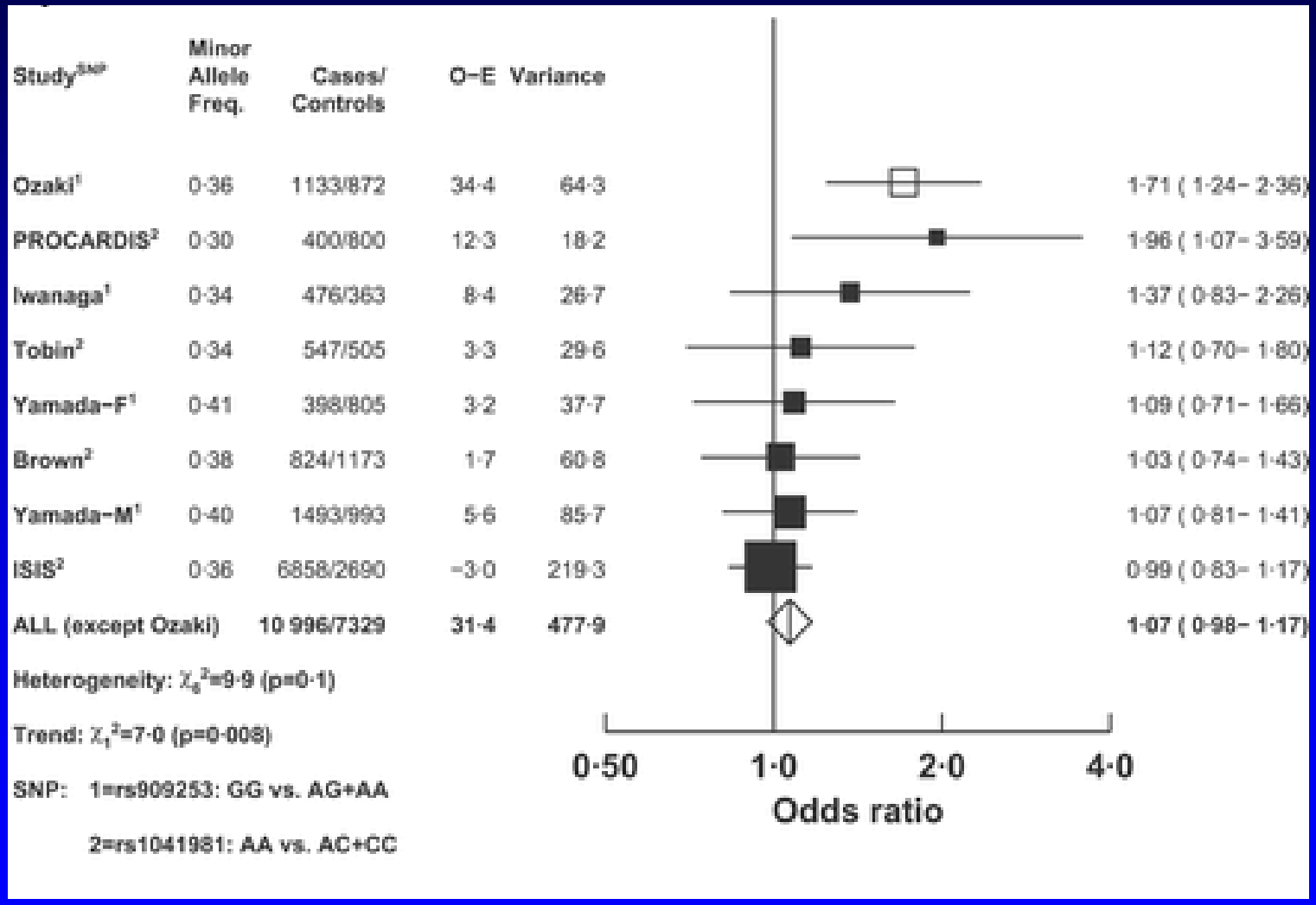
- Use a different phenotype
- Use different markers
- Mix fine-mapping and replication
- Use different analytic methods (haplotype vs. single marker)
- Use different populations

Associations between MI and LTA SNPs

	Cases (1,133)	Control1 (1,006)	Control2 (872)	P-Value Control1	P-Value Control2	Odds Ratio [95% CI] Control 1	Odds Ratio [95% CI] Control 2
Exon 1: 10G → A							
GG	37	38	39	3.3×10^{-6}	6.9×10^{-6}	1.78	1.79
GA	45	51	49				
AA	19	12	12				
Intron 1: 252A → G							
AA	37	37	40	2.2×10^{-5}	1.8×10^{-5}	1.69	1.75
AG	45	51	49				
GG	18	12	12				
Exon 3: 804C → A							
CC	37	37	39	3.3×10^{-6}	7.3×10^{-6}	1.78	1.79
CA	45	51	49				
AA	19	12	12				

Ozaki et al., *Nat Genet* 2002; 32:650-54.

Odds ratio (CI) for CHD Associated with *LTA* Genotypes in ISIS and Other Studies



Clarke et al., *PLoS Genet* 2006; 2:e107.

Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data

Sebastian Zöllner and Jonathan K. Pritchard

Am J Hum Genet 2007; 80:605-15.

Genomewide association studies are now a widely used approach in the search for loci that affect complex traits. After detection of significant association, estimates of penetrance and allele-frequency parameters for the associated variant indicate the importance of that variant and facilitate the planning of replication studies. However, when these estimates are based on the original data used to detect the variant, the results are affected by an ascertainment bias known as the "winner's curse." The actual genetic effect is typically smaller than its estimate. This overestimation of the genetic effect may cause replication studies to fail because the necessary sample size is underestimated. Here, we present an approach that corrects for the ascertainment bias and generates an estimate of the frequency of a variant and its penetrance parameters. The method produces a point estimate and confidence region for the parameter estimates. We study the performance of this method using simulated data sets and show that it is possible to greatly reduce the bias in the parameter estimates, even when the original association study had low power. The uncertainty of the estimate decreases with increasing sample size, independent of the power of the original test for association. Finally, we show that application of the method to case-control data can improve the design of replication studies considerably.

- Initial report of association almost always over-estimates magnitude of association, particularly if sample size is small
- Place more faith in large OR derived from very large studies than from very small studies

WTCCC, *Nature* 2007; 447:661-78.

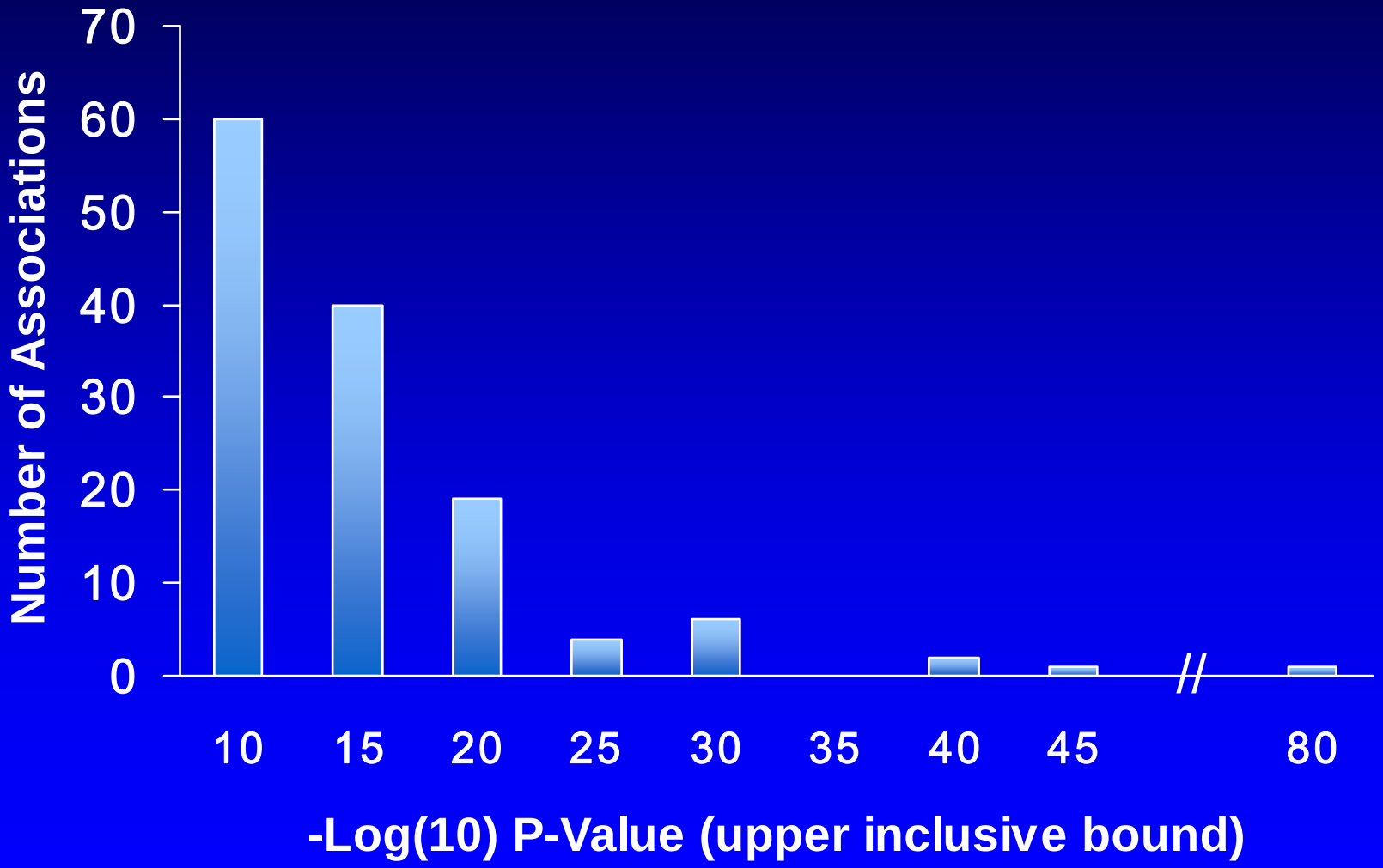
Definition of Robust Initial Finding

- Sufficient statistical power to observe reported effect, which will vary by magnitude of observed effect
- Highly significant analysis using stable method
- Consistent findings using simple, straight-forward analytic approach
- Consistent findings in:
  - Epidemiologically sound study
  - Overall and within key subgroups
  - Same or very similar phenotypes

Importance of Significance Level

- Should we promulgate a specific number– NO, but in general, smaller is better
- General agreement: range is very broad, higher threshold for difficult to measure phenotype
- Beware of the very smallest
- If significance depends on analytic method or multiple comparison correction, BEWARE
- If significance or association depends on phenotype definition, BEWARE
- Randomize the phenotypes and report number significant at that level
- Biologic information may be useful *A PRIORI* but *a posteriori* can come up with almost anything

- Log_{10} P-Values of Discrete Associations

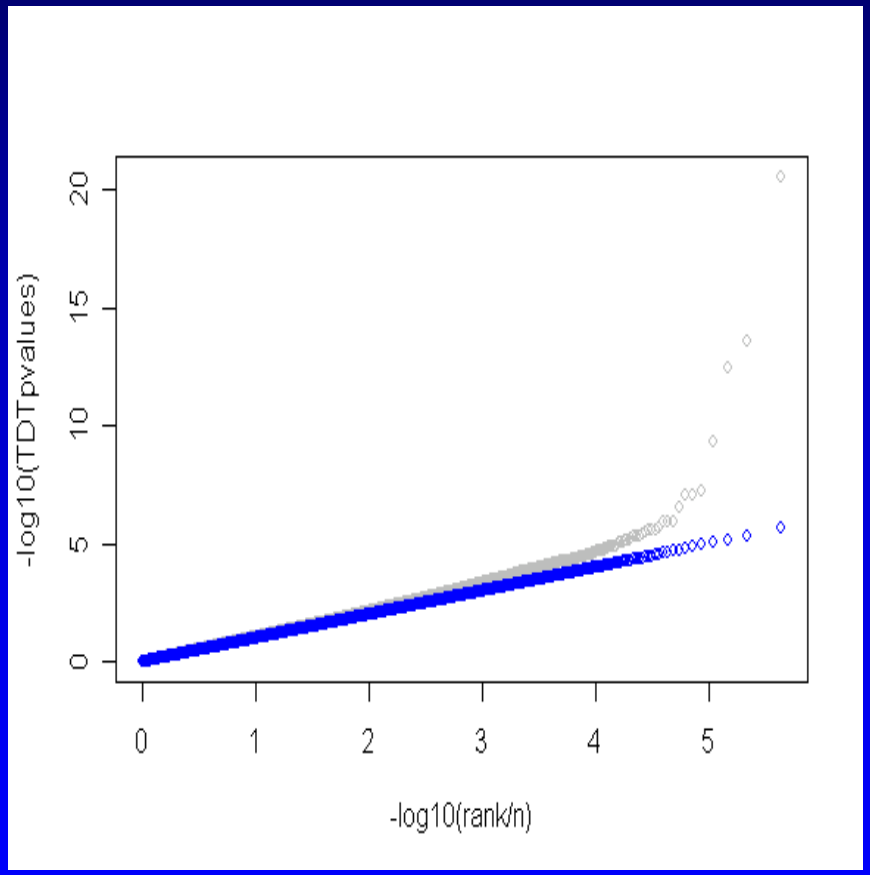


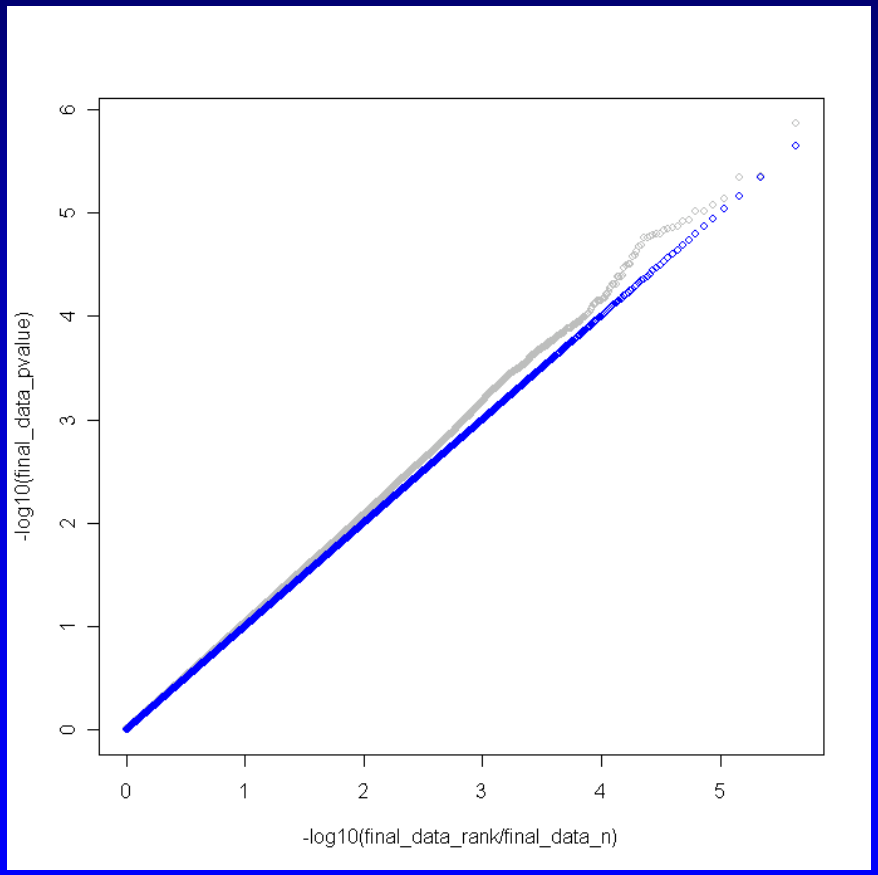
http://www.genome.gov/gwastudies/

Importance of Genotyping Quality

- Report results of known study sample duplicates, HapMap or other standard duplicates
- Replicate small number of “significant” SNPs with second technology at some late stage
- May not be needed if nearby SNPs in strong LD show same results
- Strong caveats are needed regarding fallibility of genotyping
  - Results can change based on genotype calling algorithm
  - QC filters and consistency of results after applying them must be described

Q-Q Plots Before and After Elimination of SNPs with Low Call Rate and Low MAF





Courtesy J Paschall, NCBI

Consensus Criteria for Positive Replication

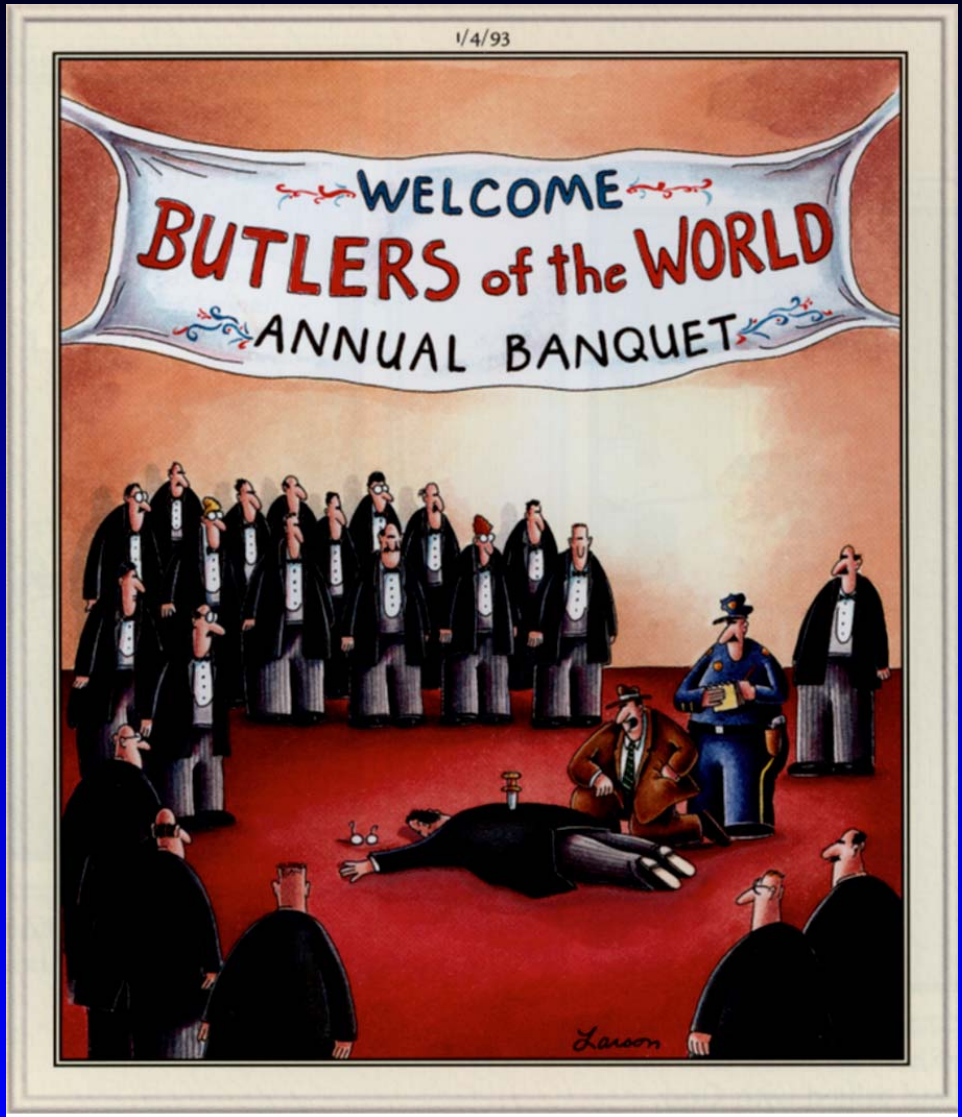
- Sufficient sample size to distinguish proposed effect from no effect convincingly
- Same or very similar trait
  - Extension to related trait may increase confidence, such as dichotomized obesity and continuous BMI)
- Same or very similar population
  - Extension to other populations may also increase confidence, such as consistent association in populations of European, Asian, or even recent African ancestry

Consensus Criteria for Positive Replication

- Same inheritance model (dominant, co-dominant, recessive), though not necessarily same analytic method
- Same gene, same SNP (or SNP in complete LD with prior SNP, $r^2 \sim 1$), same direction as original finding
- Highly significant association
- N.B.: Initial study must adequately describe these parameters

Proposed Criteria for True Non-Replication or "Meaningful Negativity"

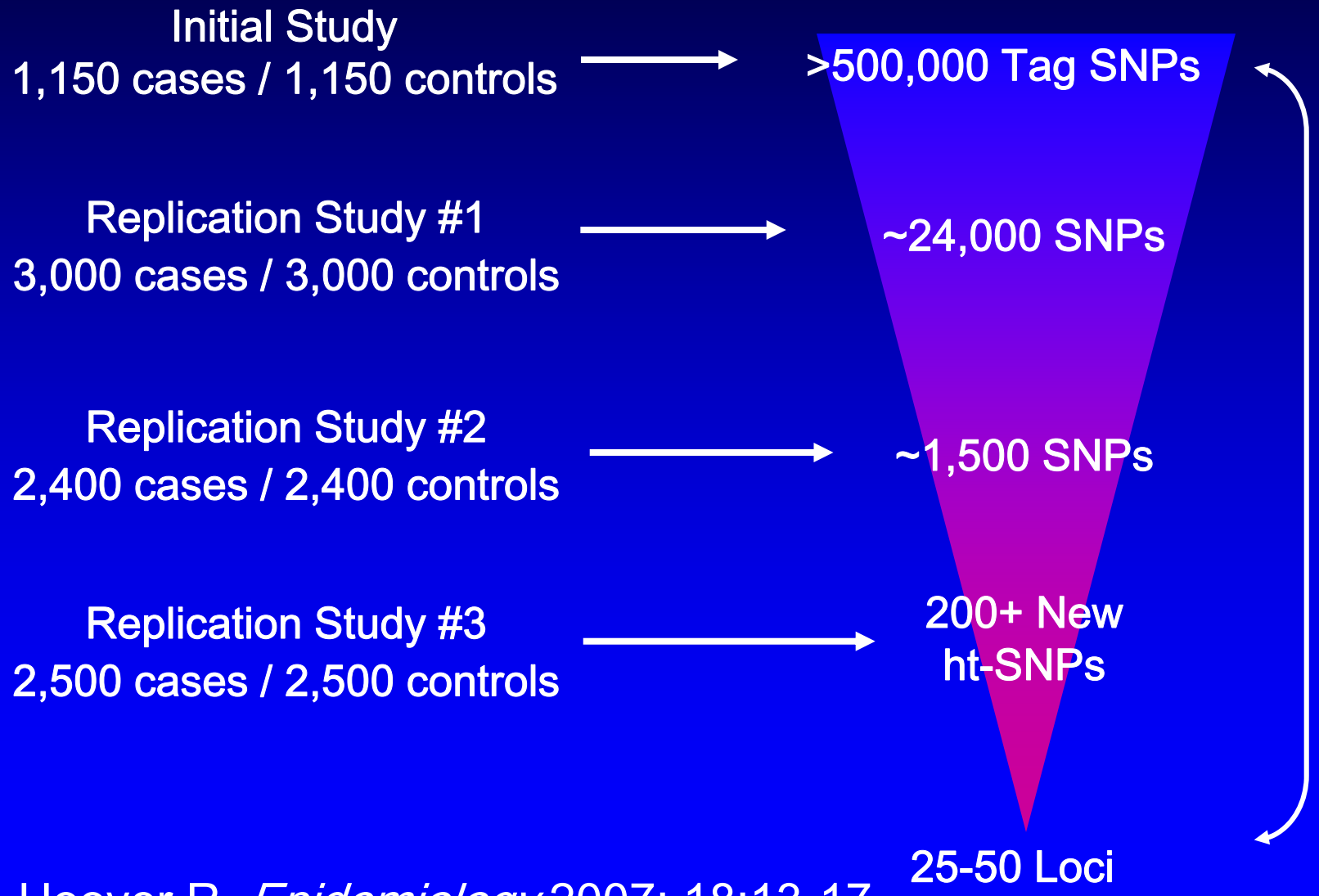
- Same as for positive replication (same trait, same gene, same SNP, same direction, same genetic model)
- Must be identical trait and population to claim non-replication
- Powered to appropriate effect size (account for "winner's curse")



"God, Collings, I hate to start a Monday with a case like this."

Larson, G. The Complete Far Side. 2003.

Replication Strategy for Prostate Cancer Study in CGEMS



Hoover R, *Epidemiology* 2007; 18:13-17.

Replication Strategy in Easton Breast Cancer Study

Stage	Cases	Controls	SNPs
1	408	400	266,722

Easton et al, *Nature* 2007; 447:1087-93.

Replication Strategy in Easton Breast Cancer Study

Stage	Cases	Controls	SNPs
1	408	400	266,722
2	3,990	3,916	13,023

Easton et al, *Nature* 2007; 447:1087-93.

Replication Strategy in Easton Breast Cancer Study

Stage	Cases	Controls	SNPs
1	408	400	266,722
2	3,990	3,916	13,023
3	23,734	23,639	31

Easton et al, *Nature* 2007; 447:1087-93.

Replication Strategy in Easton Breast Cancer Study

Stage	Cases	Controls	SNPs
1	408	400	266,722
2	3,990	3,916	13,023
3	23,734	23,639	31
Final			6

- ABCFS
- BCST
- COPS
- GENICA
- HBCS
- HBCP
- TBCS
- KConFab/AOCS
- KBCP
- LUMCBCS
- MCBCS
- MCCS
- MEC-W
- MEC-J
- NHS
- PBCS
- RBCS
- SASBAC
- SEARCH2
- SEARCH3
- SBCP
- SBCS
- CNIOBCS
- USRT

Easton et al, *Nature* 2007; 447:1087-93.

"And now Edgar's gone. ... Something's going on around here."

Larson, G. *The Complete Far Side*. 2003.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

* Selected for $p < 0.068$

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

* Selected for $p < 0.068$

SNP	Gene	Stage 1+2 P-value
rs4962416	*MSMB*	7×10^{-13}
rs10896449	11q13	2×10^{-9}
rs10993994	*CTBP2*	2×10^{-7}
rs10486567	*JAZF1*	2×10^{-6}

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

* Selected for $p < 0.068$

SNP	Gene	Stage 1+2 P-value	Initial Rank
rs4962416	*MSMB*	7×10^{-13}	24,223
rs10896449	11q13	2×10^{-9}	
rs10993994	*CTBP2*	2×10^{-7}	
rs10486567	*JAZF1*	2×10^{-6}	

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

* Selected for $p < 0.068$

SNP	Gene	Stage 1+2 P-value	Initial Rank
rs4962416	*MSMB*	7×10^{-13}	24,223
rs10896449	11q13	2×10^{-9}	2,439
rs10993994	*CTBP2*	2×10^{-7}	319
rs10486567	*JAZF1*	2×10^{-6}	24,407

Thomas et al, *Nat Genet* 2008; 40:310-15.

Replication Strategy in CGEMS Prostate Cancer Study

Stage	Cases	Controls	SNPs
1	1,172	1,157	527,869
2	3,941	3,964	26,958*

* Selected for $p < 0.068$

SNP	Gene	Stage 1+2 P-value	Initial Rank	Initial P-value
rs4962416	*MSMB*	7×10^{-13}	24,223	0.042
rs10896449	11q13	2×10^{-9}	2,439	0.004
rs10993994	*CTBP2*	2×10^{-7}	319	4×10^{-4}
rs10486567	*JAZF1*	2×10^{-6}	24,407	0.042

Thomas et al, *Nat Genet* 2008; 40:310-15.

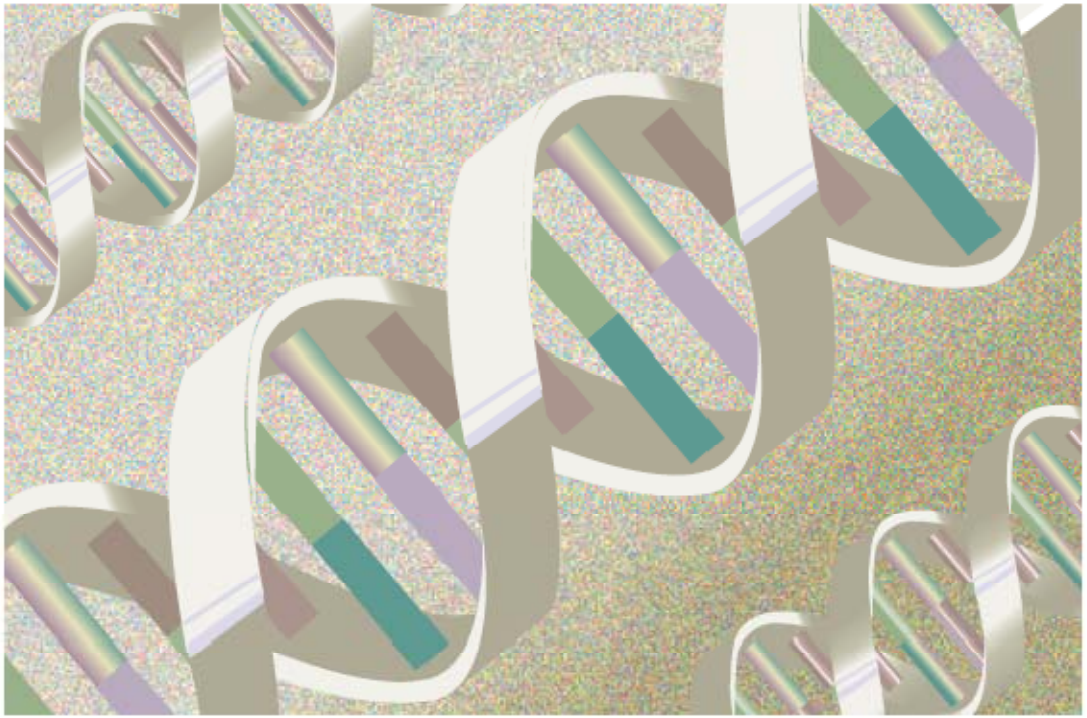
Summary Points: Replication

- False positives are huge potential problem
- Statistical corrections: Bonferroni, false discovery rate, false positive report probability
- False negatives are also important problem
- Replication is *sine qua non*
- Same inheritance model, same SNP, same direction, same or similar population
- Allow for smaller effect size (winner's curse)

Replicating genotype–phenotype associations

What constitutes replication of a genotype–phenotype association, and how best can it be achieved?

NCI-NHGRI Working Group on Replication in Association Studies



The study of human genetics has recently undergone a dramatic transition with the completion of both the sequencing of the human genome and the mapping of human haplotypes of the most common form of genetic variation, the single nucleotide polymorphism (SNP)[1–3]. In concert with this rapid expansion of detailed genomic information, cost-effective genotyping technologies have been developed that can assay hundreds of thousands of SNPs simultaneously. Together, these advances have allowed a systematic, even 'agnostic', approach to genome-wide interrogation, thereby relaxing the requirement for strong prior hypotheses.

So far, comprehensive reviews of the published literature, most of which reports work based on the candidate-gene approach, have demonstrated a plethora of questionable genotype–phenotype associations, replication of which has often failed in independent studies[4–7]. As the transition to genome-wide association studies occurs, the challenge will be to separate true associations from the blizzard of false positives attained through attempts to rep-

studies because of issues in either the initial study or the attempted replication[4–6,32,33]. Small sample size is a frequent problem and can result

conclusion from the literature because follow-up studies have not consistently analysed the same markers or those in perfect linkage dis-

Chanock et al., *Nature* 2007; 447:655-60.

Box 1 | Points to consider in genotype–phenotype association reports

This checklist is intended to serve as a guide for authors, journal editors and referees to allow clear and unambiguous interpretation of the data and results of genome-wide and other genotype–phenotype association studies.

Study information

- A detailed description of the study design and its implementation
- The source of cases and controls (or cohort members, if based on cohort design), including time period and location(s) of subject recruitment
- Methods for ascertaining and validating affected or unaffected status and reproducibility of classification
- Participation rates for cases, controls or cohort members
- Presentation of case and control selection in a flow chart, including exclusion points for missing and erroneous data (possibly as supplementary tables)
- Initial table comparing relevant characteristics (such as demographics, risk factors and exposures) of cases and controls
- Success rate for DNA acquisition, including comparisons of those with and without collection, extraction failures and exclusions due to inconsistent data

Genotyping and quality control procedures

- Sample tracking methods, such as bar-coding, to ensure accuracy of analysis
- Description of genotyping assays and protocols, particularly when new or applied in a non-standard method
- Description of genotyping calling algorithm
- Genotype quality control design for samples, including numbers, plating locations, selection criteria for:
  - External control samples from standard accepted sets (such as HapMap)
  - Internal control samples (duplicate samples; it should be specified whether these are from the same or different DNA collection, extraction or aliquot)
- Assay and DNA quality metrics by locus, sample, plate or 'batch'
- Validation of most critical results on an independent genotyping platform

Data issues

- Statement on availability of results and data so that, as far as possible, others can analyse them independently
- Links to supplemental online resources and database accession numbers

Results

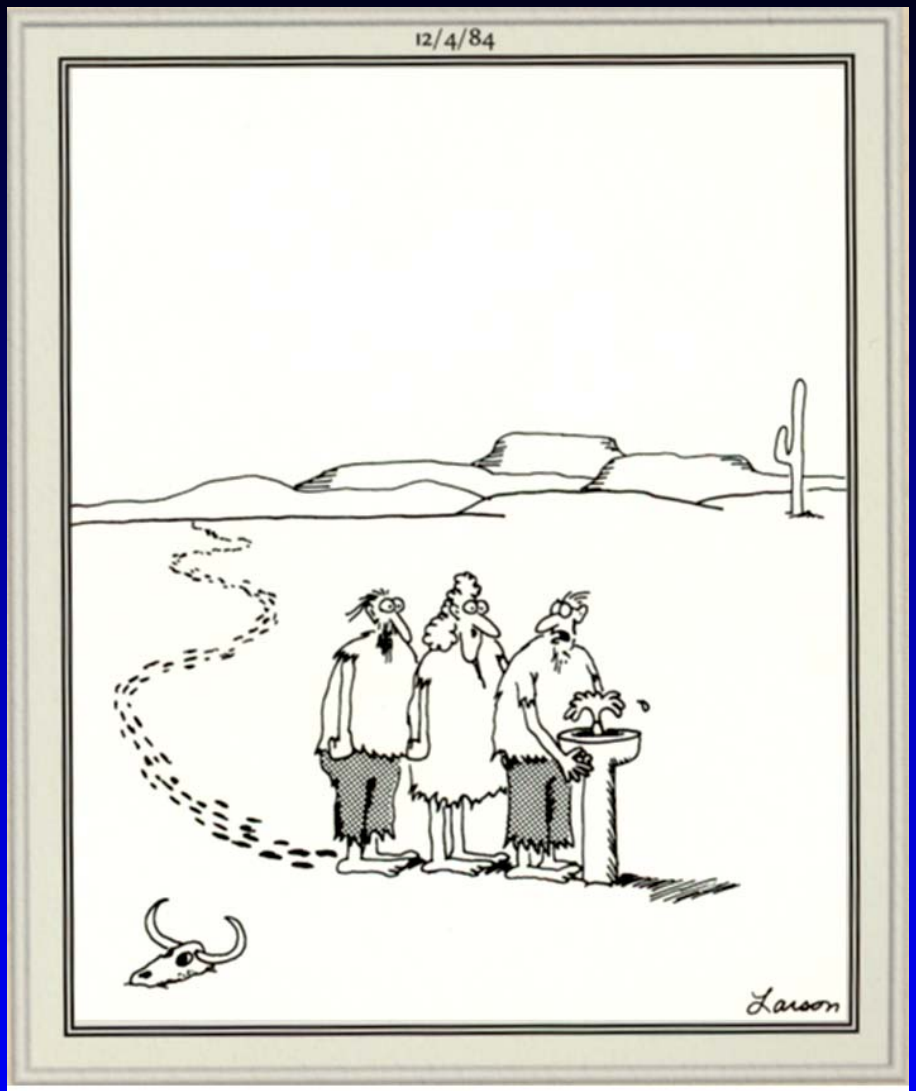
- Analysis methods in sufficient detail to reconstruct the analytical approach and reproduce all reported results
- Description of any pre-analysis weighting scheme for selecting variants for replication
- Simple single-locus and multi-marker (haplotype) association analyses
- Genetic models tested (unconstrained genotype effects — dominant, additive, multiplicative or trend)
- Graphical display of genotype clustering for assays of high interest
- Verification of results at highly correlated loci
- Discussion of choice of threshold for significance and the statistical basis for any adjustment for multiple testing and the relationship to overall study power
- Significance of any known 'positive controls' (that is, loci established in previous genetic associations)
- Consistency of results before and after application of quality control filters

Chanock et al., *Nature* 2007; 447:655-60.

Inclusion of Standard Genotyping Quality Control Analyses

- Average value of chi-square and full distribution
- Q-Q plots of chi square and p-values
- Genotyping cluster plots for SNPs of interest
- Signal at nearby or correlated SNPs
- Genotype QC filters applied, including HWE, call rates, MAF
- Testing for plate or batch effects
- Description of calling algorithm
- Confirmation of top hits on different platform

Chanock et al., *Nature* 2007; 447:655-60.



"Now just hold your horses, everyone. ... Let's let it run for a minute or so and see if it gets any colder."

Larson, G. *The Complete Far Side*. 2003.

Quality Control of SNP Genotyping: Samples

- Identity with forensic markers (Identifiler)
- Blind duplicates
- Gender checks
- Cryptic relatedness or epidemic twinning
- Degradation/fragmentation
- Call rate (> 80-90%)
- Heterozygosity: outliers
- Plate/batch calling effects

Chanock et al., *Nature* 2007; Manolio et al., *Nat Genet* 2007

Quality Control of SNP Genotyping: SNPs

- Duplicate concordance (CEPH samples)
- Mendelian errors (typically ≤ 1)
- Hardy-Weinberg errors (often $> 10^{-5}$)
- Heterozygosity (outliers)
- Call rate (typically > 98%)
- Minor allele frequency (often > 1%)
- Validation of most critical results on independent genotyping platform

Chanock et al., *Nature* 2007; Manolio et al., *Nat Genet* 2007

COMMENTARY

New models of collaboration in genome-wide association studies: the Genetic Association Information Network

The GAIN Collaborative Research Group

The Genetic Association Information Network (GAIN) is a public-private partnership established to investigate the genetic basis of common diseases through a series of collaborative genome-wide association studies. GAIN has used new approaches for project selection, data deposition and distribution, collaborative analysis, publication and protection from premature intellectual property claims. These demonstrate a new commitment to shared scientific knowledge that should facilitate rapid advances in understanding the genetics of complex diseases.

GAIN Collaborative Group, *Nat Genet* 2007; 39:1045-51.

Coverage, Call Rates, and Concordance of Perlegen and Affymetrix Platforms on HapMap Phase II

Metric	Perlegen		Affymetrix/Broad	
Number of SNPs	480,744		439,249	
Coverage	Single Marker	Multi-Marker	Single Marker	Multi-Marker
CEU	0.90	0.96	0.78	0.87
CHB + JPT	0.87	0.93	0.78	0.86
YRI	0.64	0.78	0.63	0.75
Average call rate	98.9%		99.3%	
Concordance				
Homozygous genotypes	99.8%		99.9%	
Heterozygous genotypes	99.8%		99.8%	

GAIN Collaborative Group, *Nat Genet* 2007; 39:1045-51.

Sample and SNP QC Metrics for Affymetrix 5.0 and 6.0 Platforms in GAIN

Metric	5.0	% fail	6.0	% fail
Total Samples	1,829	--	2,289	--
Passing QC	1,817	0.44	2,192	4.24
≥ 98% call rate	1,815	0.55	2,257	1.40

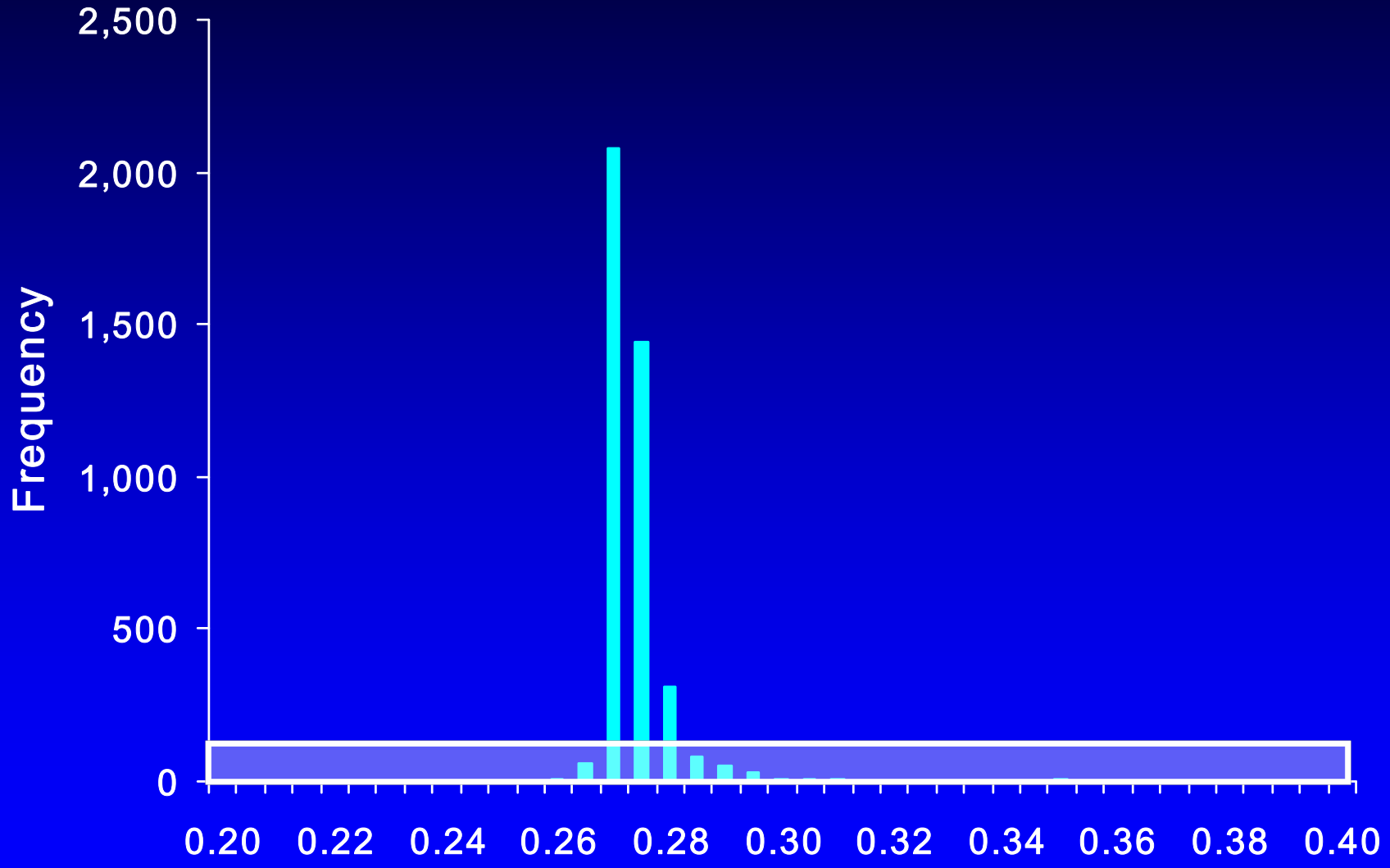
Courtesy, J Paschall, NCBI

Sample and SNP QC Metrics for Affymetrix 5.0 and 6.0 Platforms in GAIN

Metric	5.0	% fail	6.0	% fail
Total Samples	1,829	--	2,289	--
Passing QC	1,817	0.44	2,192	4.24
≥ 98% call rate	1,815	0.55	2,257	1.40
Total SNPs	457,645	--	906,660	--
Passing QC	429,309	6.19	845,814	6.70
MAF ≥ 1%	457,466	0.04	888,234	2.03
≥ 98% call rate	419,810	8.27	821,942	9.34
≥ 95% call rate	439,272	4.01	873,856	3.61
HWE < 10 -6	455,899	0.38	904,275	0.26
≤ 1 Mendel error	417,722	8.72	899,721	0.01
≤ 1 Duplicate error	454,820	0.01	892,103	0.02

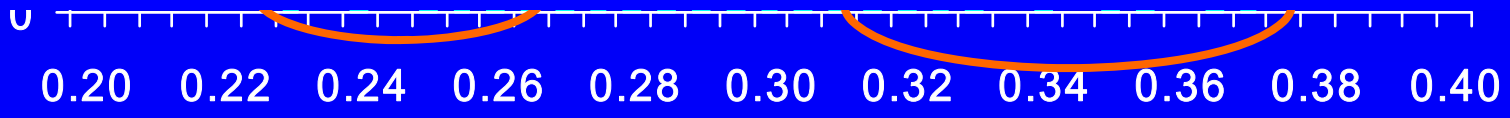
Courtesy, J Paschall, NCBI

Sample Heterozygosity in GAIN



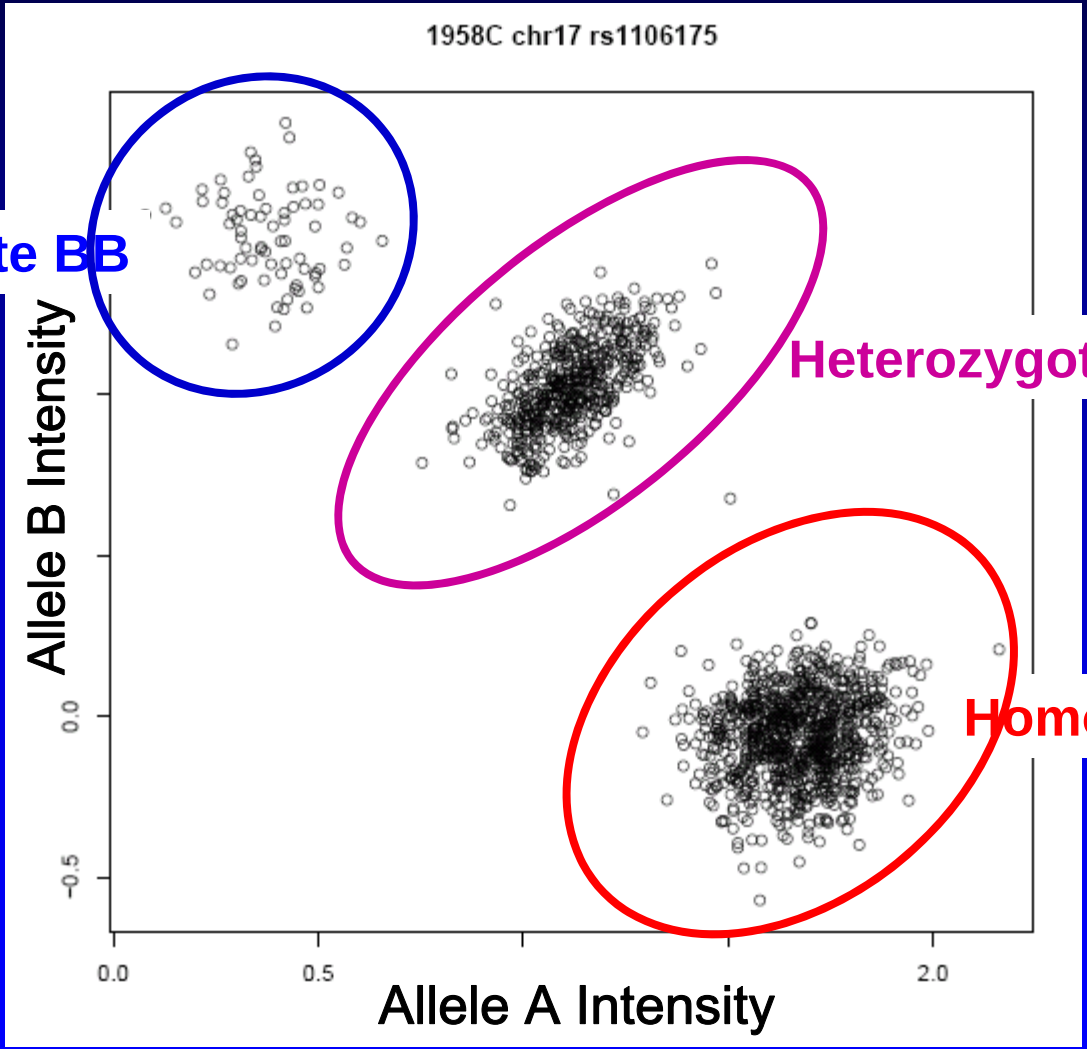
Courtesy, J Paschall, NCBI

Sample Heterozygosity in GAIN



Courtesy, J Paschall, NCBI

Automated Genotype Calling



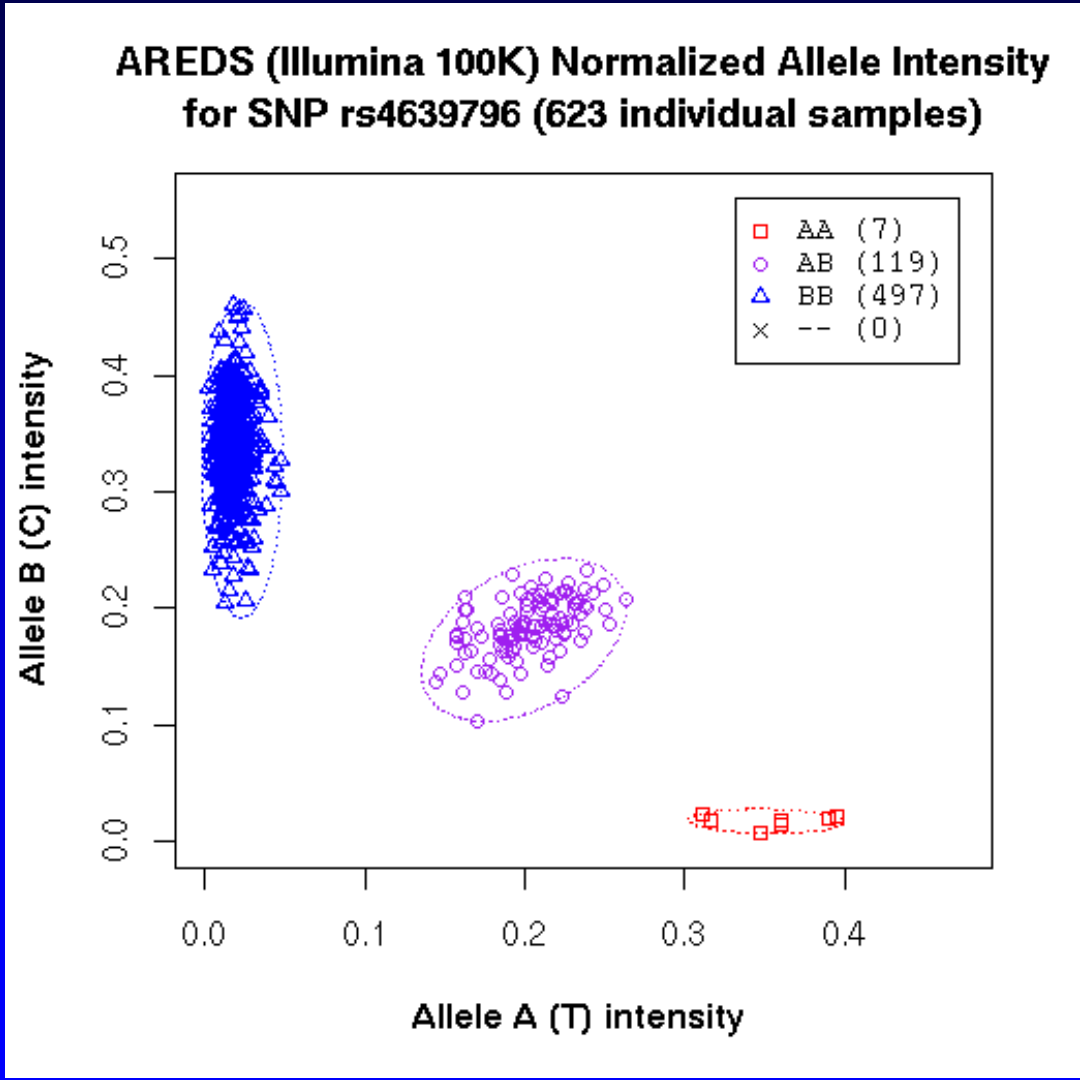
Homozygote BB

Heterozygote AB

Homozygote AA

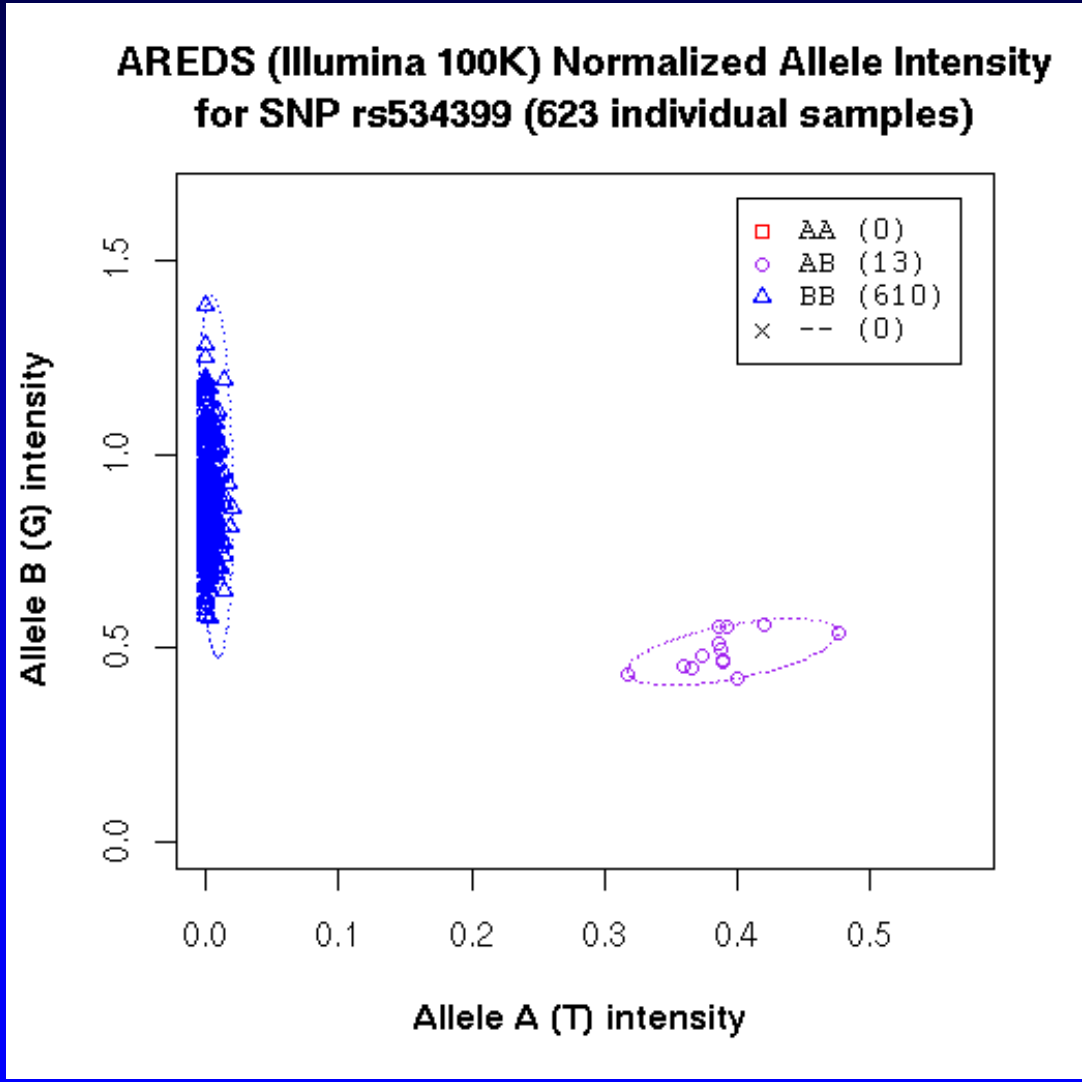
Courtesy, L Cardon, GSK

Signal Intensity Plots for rs4639796 in AREDS



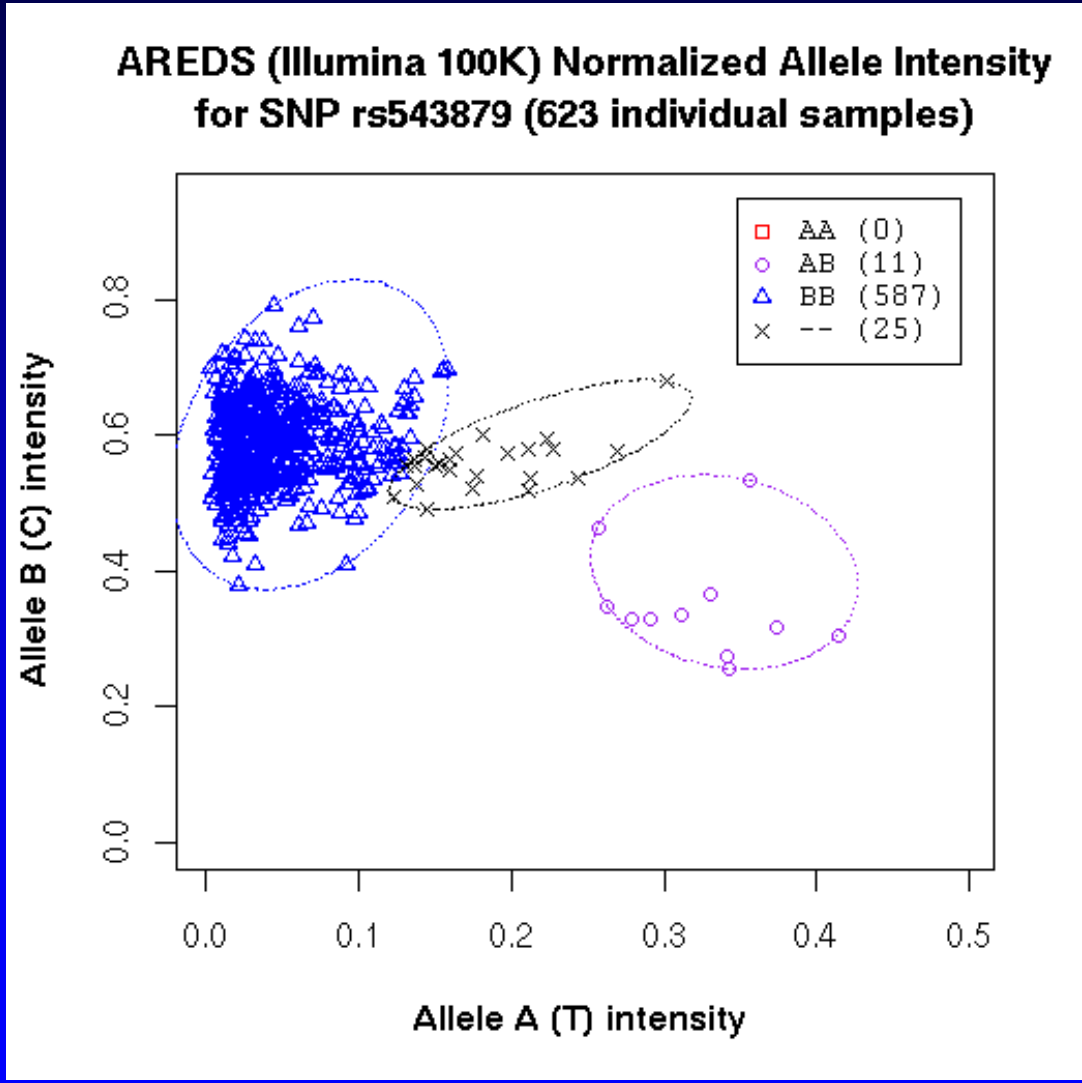
http://www.ncbi.nlm.nih.gov/sites/entrez

Signal Intensity Plots for rs534399 in AREDS



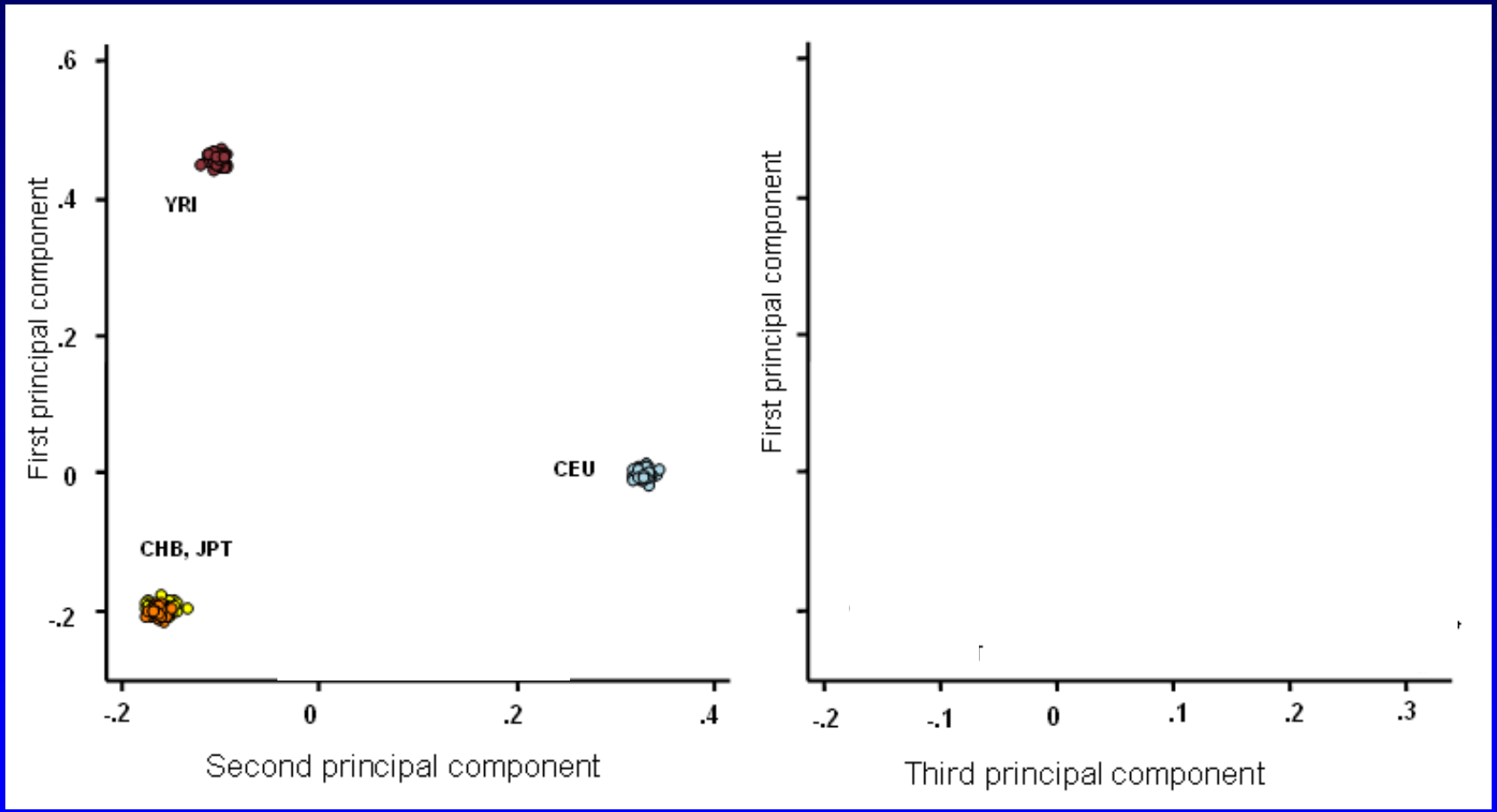
http://www.ncbi.nlm.nih.gov/sites/entrez

Signal Intensity Plots for rs572515 in AREDS



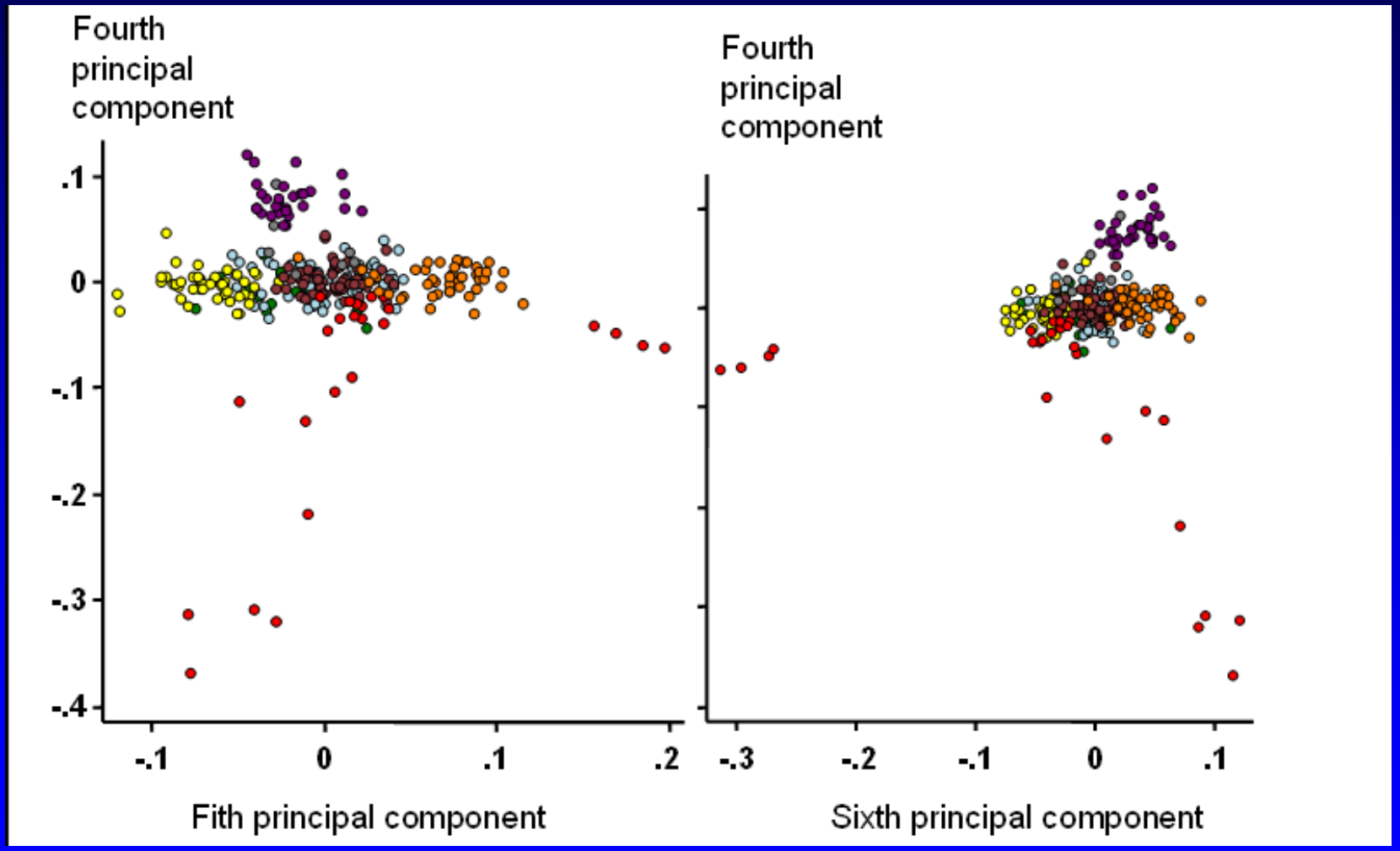
http://www.ncbi.nlm.nih.gov/sites/entrez

Principal Component Analysis of Structured Population: First to Third Components



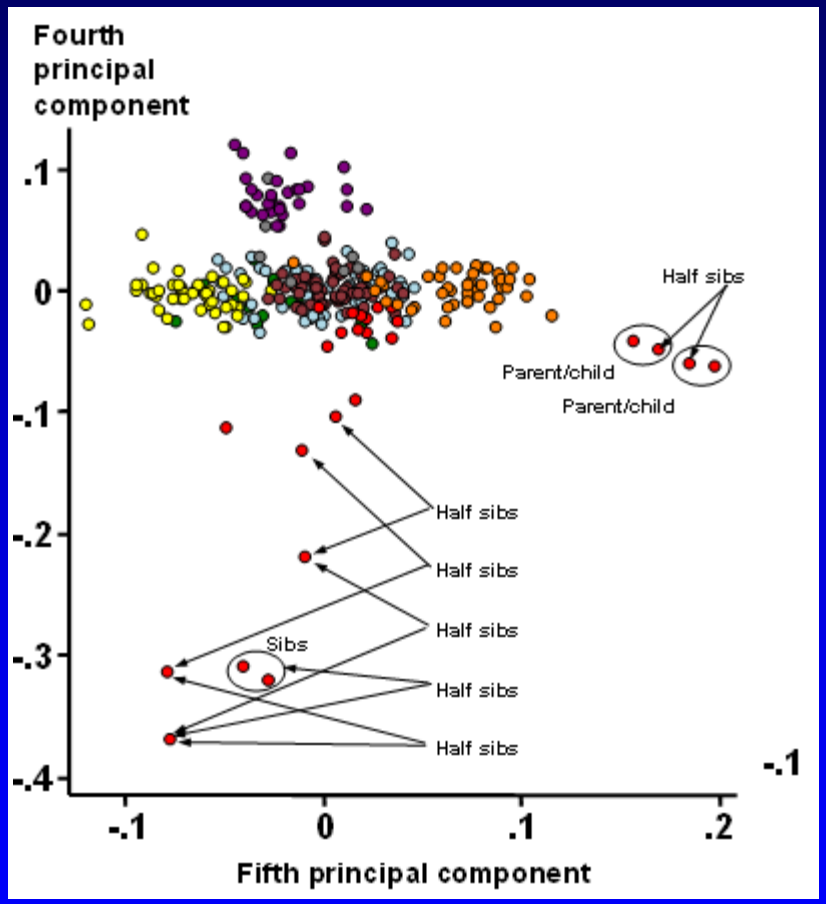
Courtesy, G. Thomas, NCI

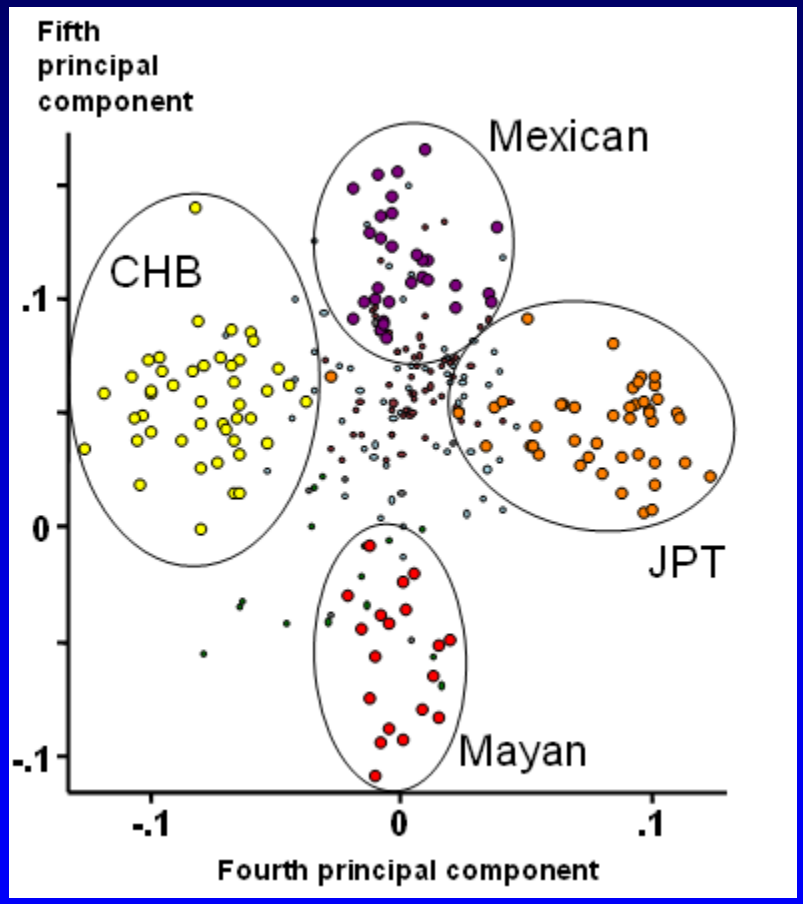
Principal Component Analysis of Structured Population: Fourth and Fifth Components



Courtesy, G. Thomas, NCI

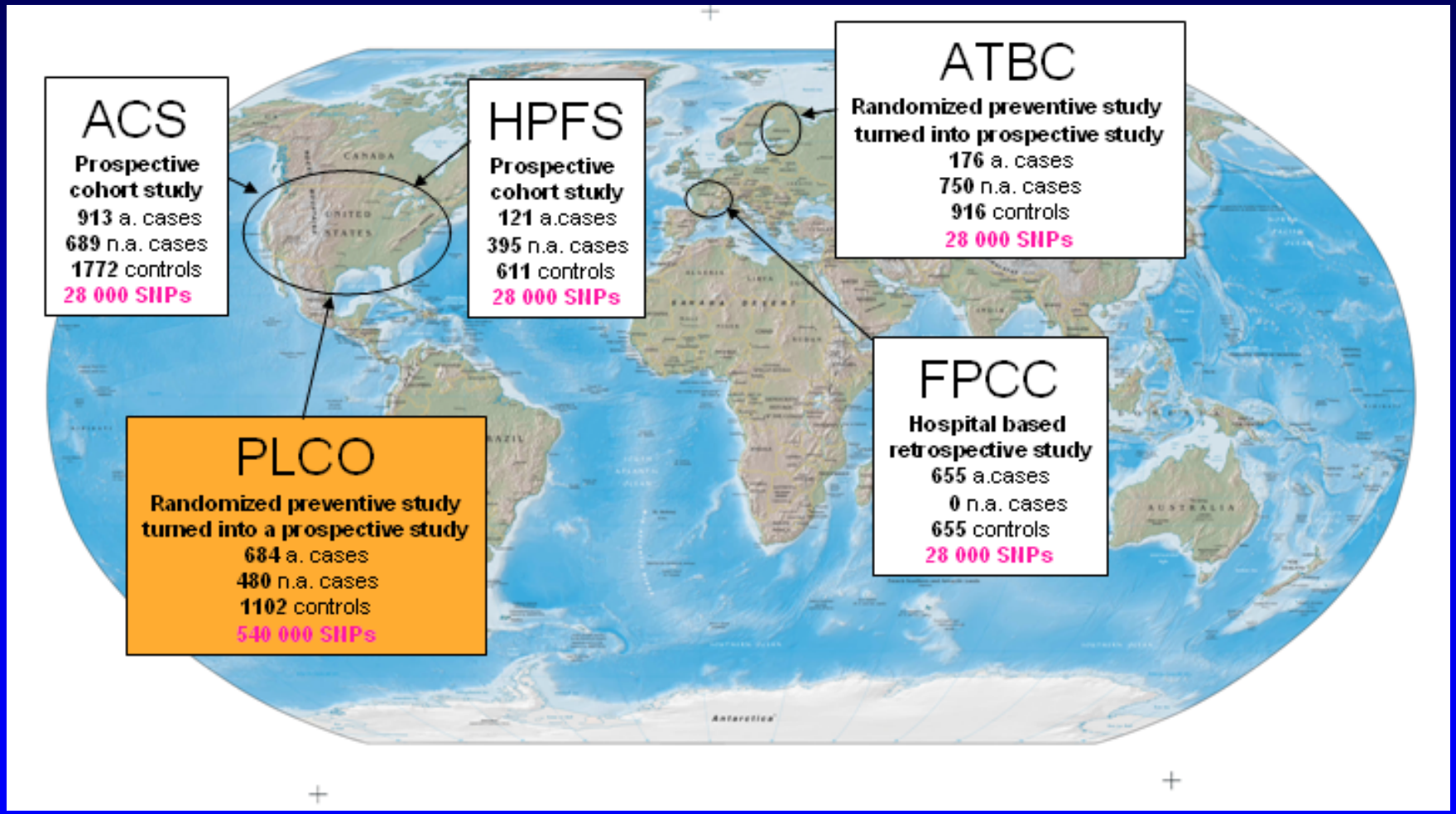
Influence of Relatedness on Principal Component Analysis





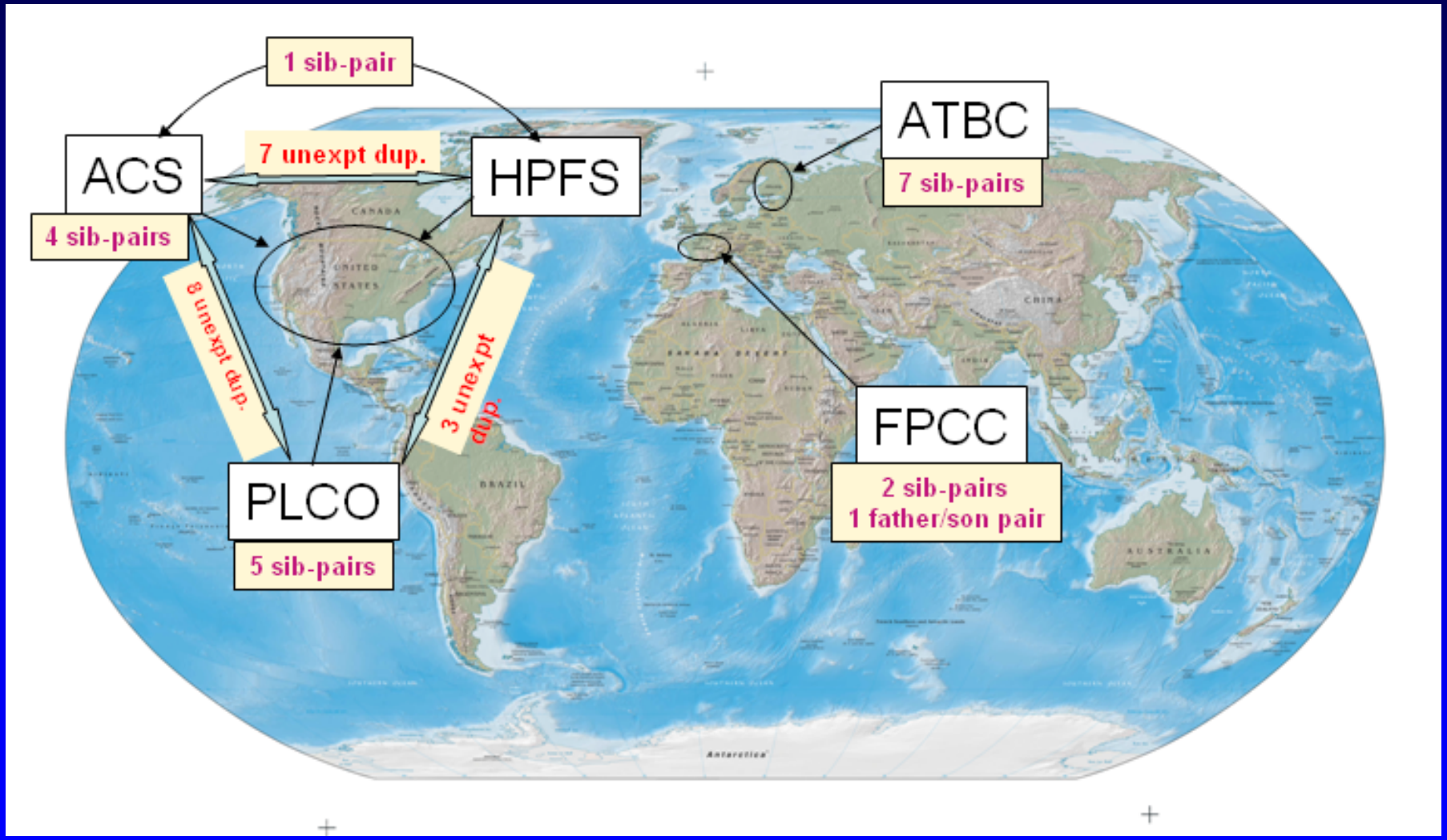
Courtesy, G. Thomas, NCI

Cryptic Relatedness in Multi-Center Studies



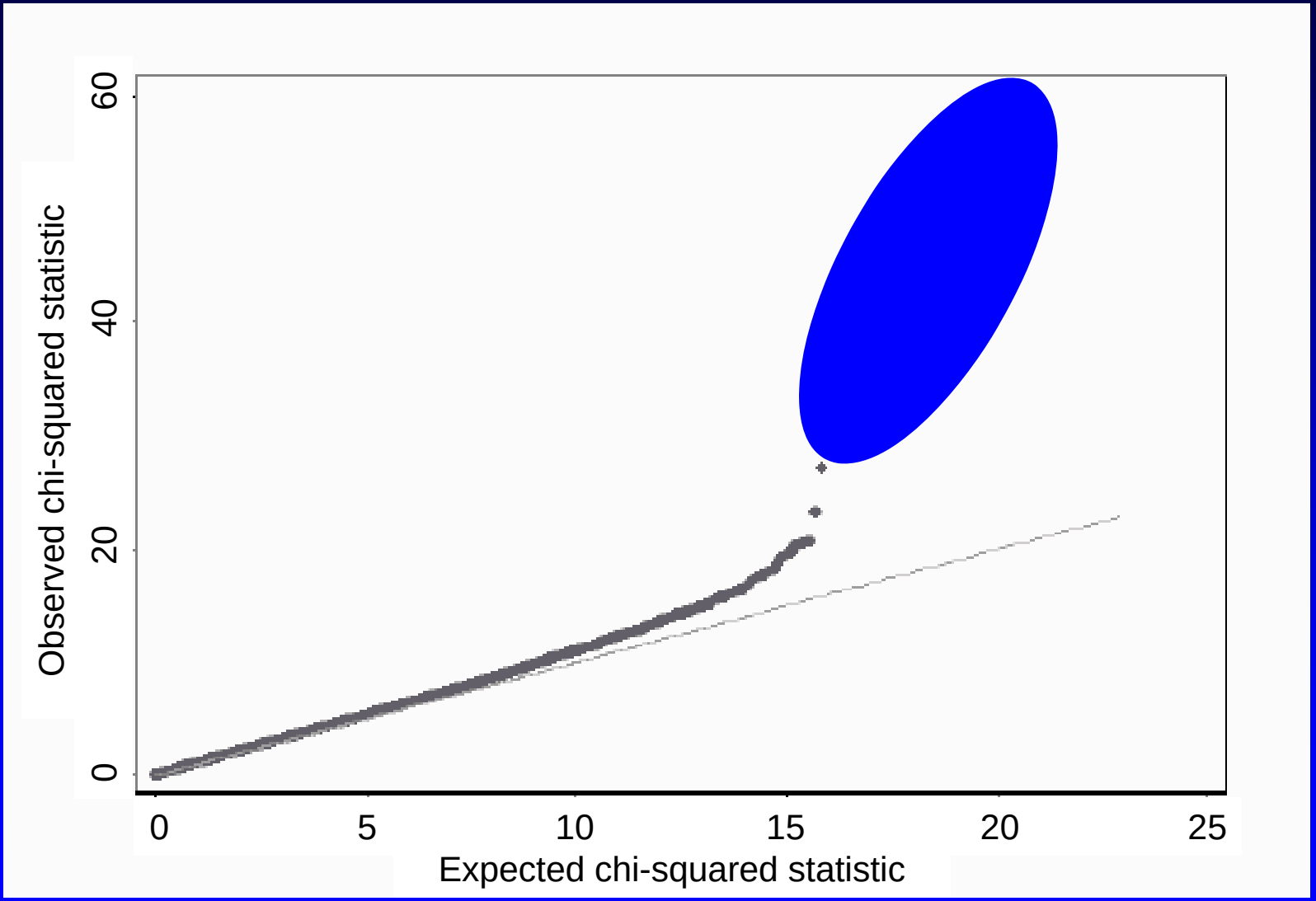
Courtesy, G. Thomas, NCI

Cryptic Relatedness in Multi-Center Studies



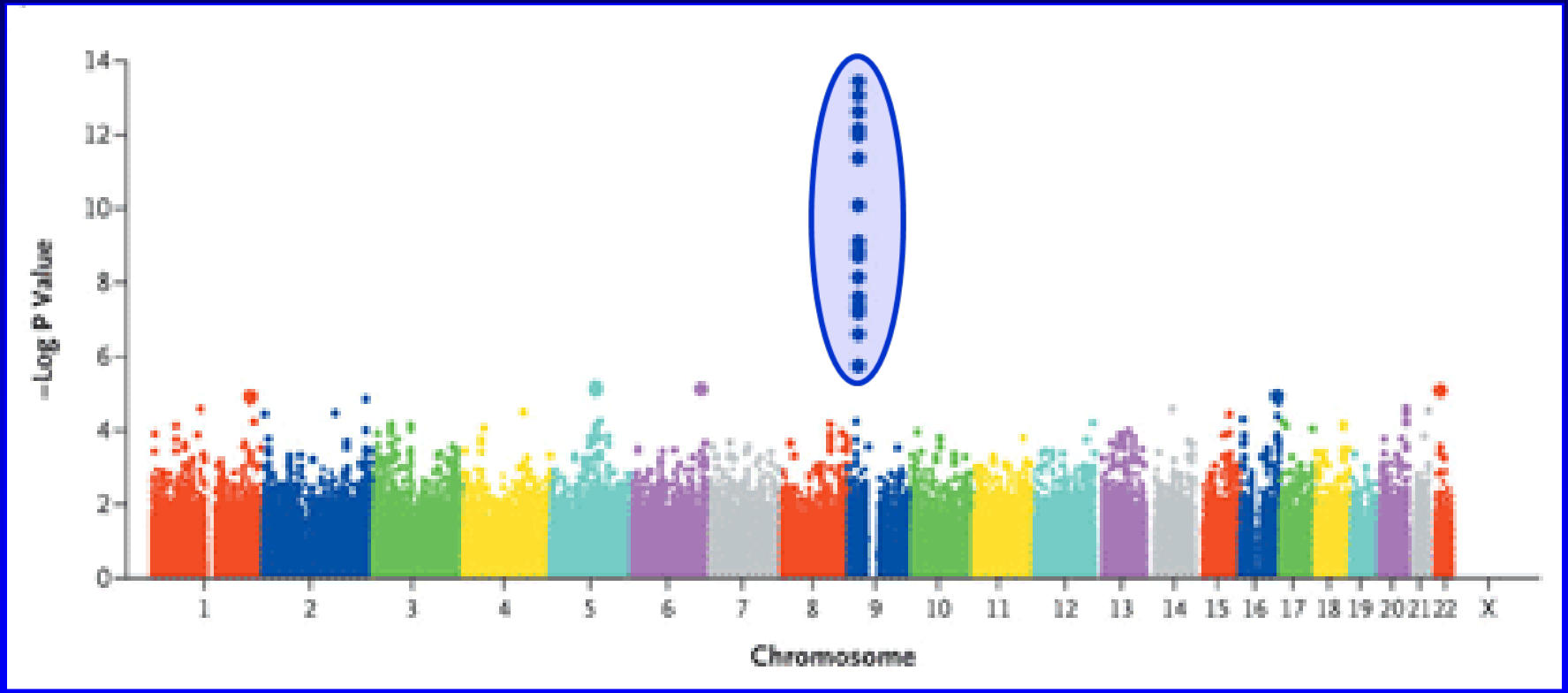
Courtesy, G. Thomas, NCI

Q-Q Plot for Myocardial Infarction



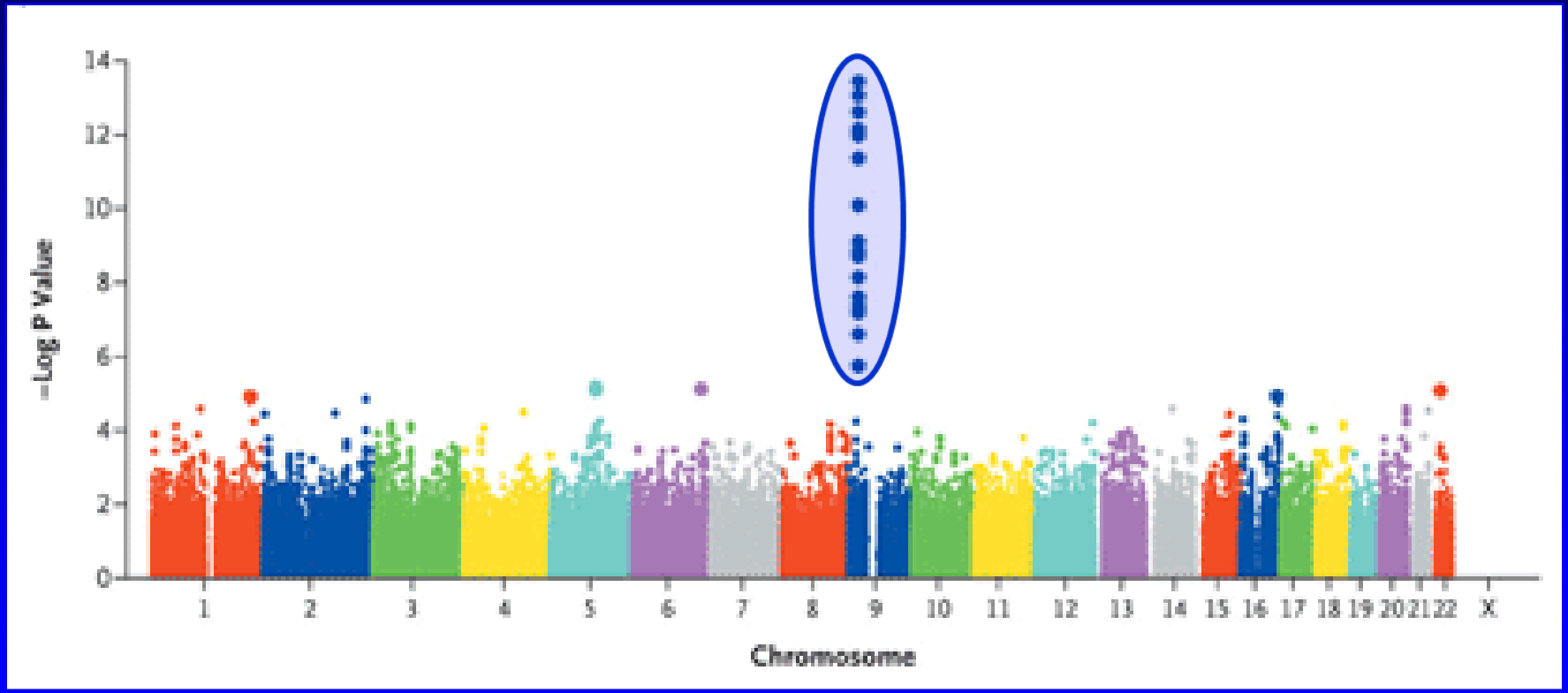
Samani N et al., *N Engl J Med* 2007; 357:443-53.

$-\text{Log}_{10}$ P Values for SNP Associations with Myocardial Infarction



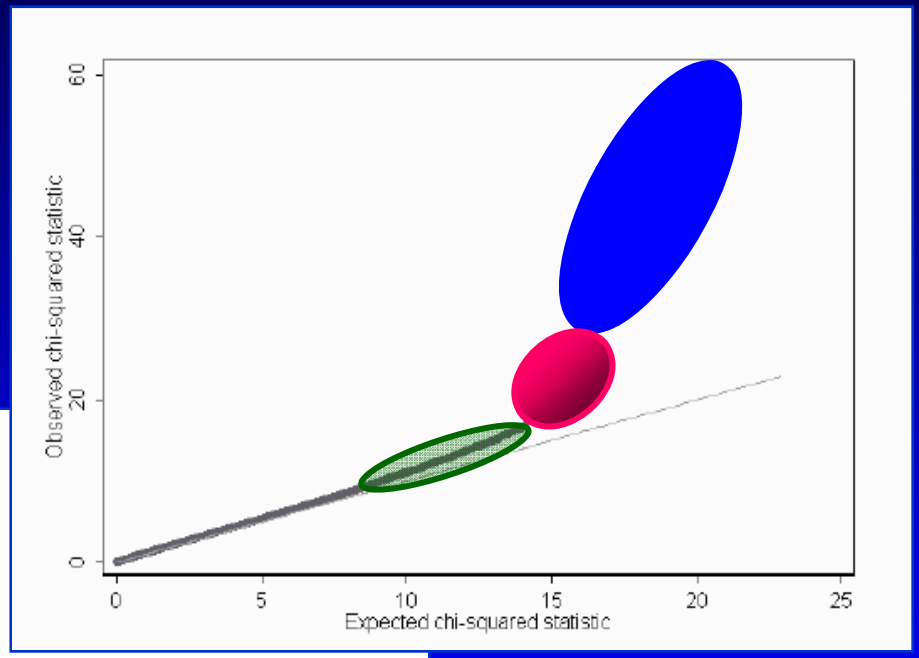
Samani N et al., *N Engl J Med* 2007; 357:443-53.

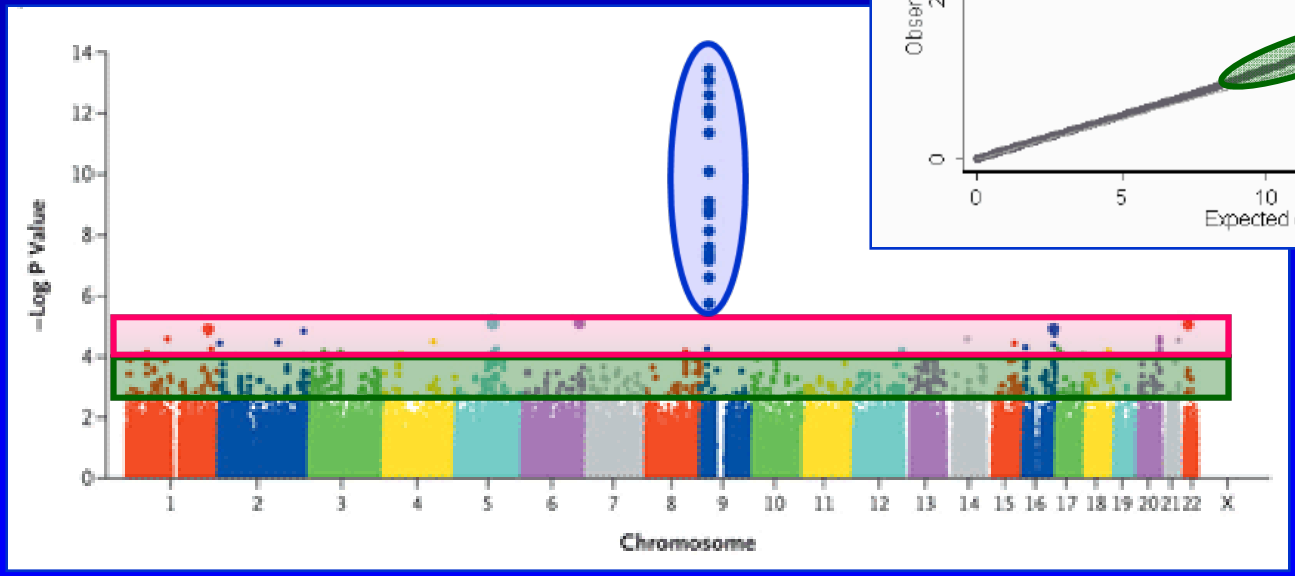
$-\text{Log}_{10}$ P Values for SNP Associations with Myocardial Infarction



Samani N et al., *N Engl J Med* 2007; 357:443-53.

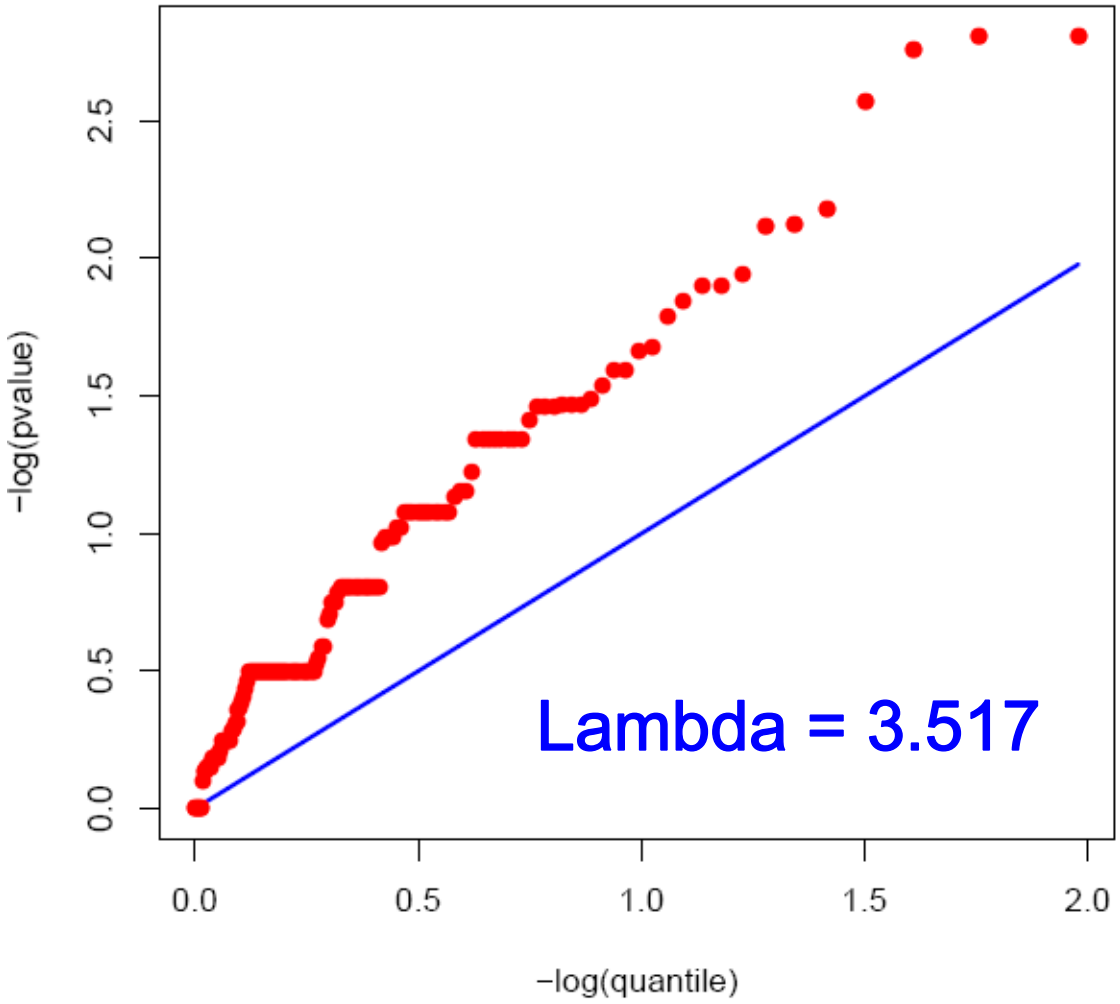
SNP Associations with 1,928 MI Cases and 2,938 Controls from UK





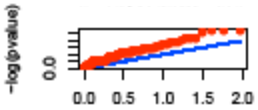
Samani N et al., *N Engl J Med* 2007; 357:443-53.

Q-Q Plot for 143 SNPs with Call Rate 90-95%; MAF 0-1%

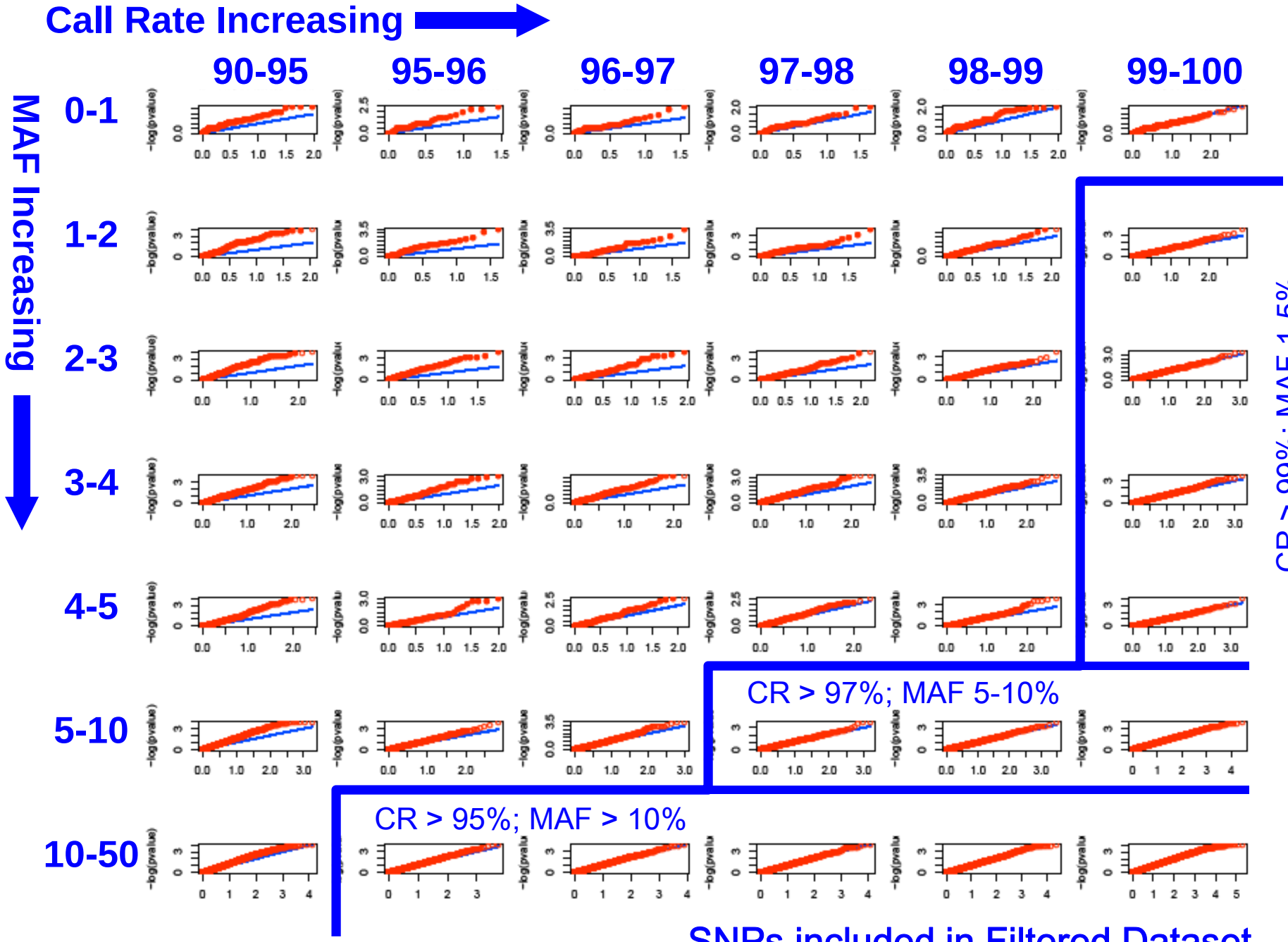


Courtesy, G Abecasis and J Paschall

90-95



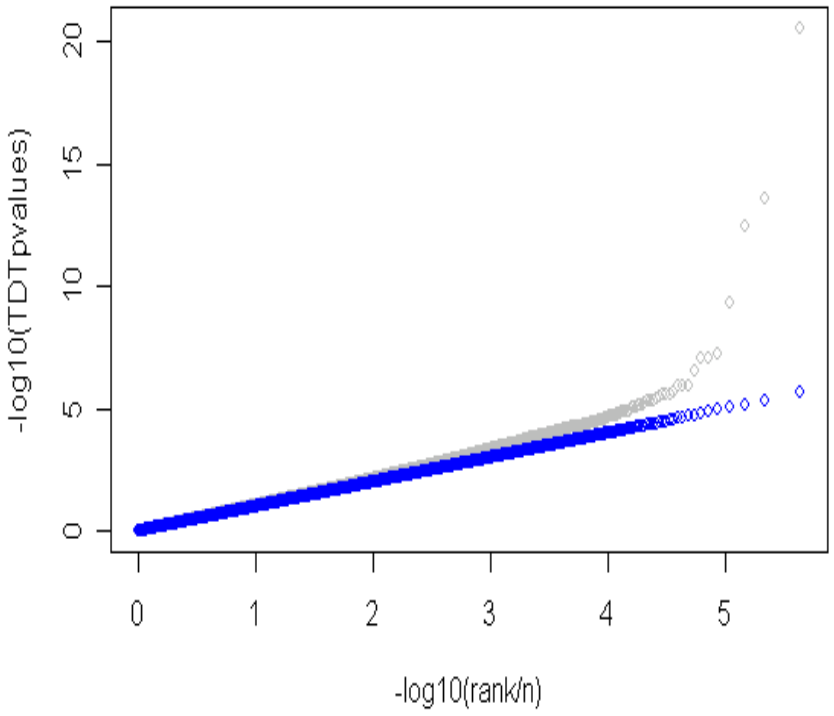
Courtesy, G Abecasis and J Paschall



Courtesy, G Abecasis and J Paschall

Q-Q Plots Before and After Elimination of SNPs with Low Call Rate and Low MAF

MAF > 1%, Call rate > 90%

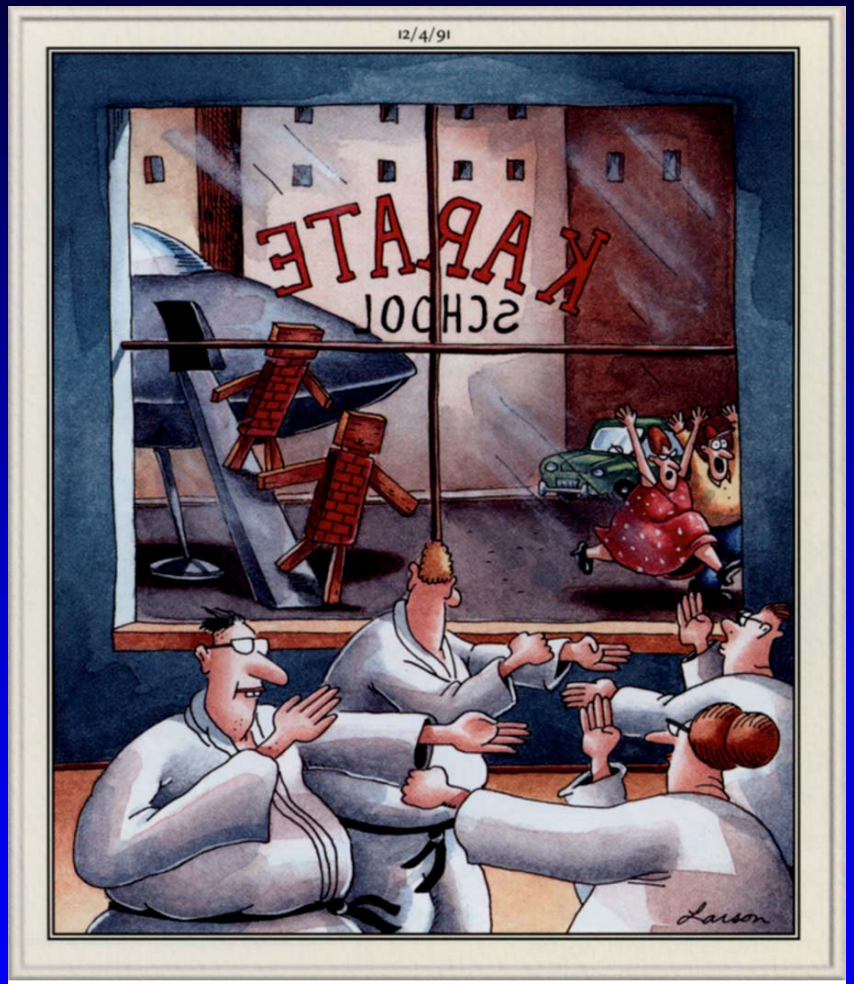


Courtesy, G Abecasis and J Paschall

Summary Points: Genotyping Quality Control

- Sample checks for identity, gender error, heterozygosity, cryptic relatedness
- Association analysis is often quickest way to find genotyping errors
- Correction for genotyping errors often wipes out most or even all associations
- Low MAF SNPs are most difficult to call
- Inspection of genotyping cluster plots is crucial!

Class Participation Exercise!



The class abruptly stopped practicing. Here was a chance to not only employ their stills, but also to save the entire town.

Larson, G. *The Complete Far Side*. 2003.

Dr. X conducts a 100,000-SNP GWA study of rheumatoid arthritis in 1,000 cases and 1,000 controls and finds 1,121 SNPs significant at $p < 5 \times 10^{-7}$.

What would you advise Dr. X to do next?

A. Genotype all these SNPs in a suitably-sized replication sample
B. Genotype the top 5,000 SNPs in a S-SRS
C. Examine cluster plots for these 1,121 SNPs
D. Review genotyping quality control metrics, particularly call rate, HWE, and concordance
E. Compare sources and characteristics of cases and controls

Dr. X discovers that her cases were selected from a private referral rheumatology hospital in Baltimore, and her controls from a general medical clinic at a nearby urban teaching hospital.

What should be her next step?

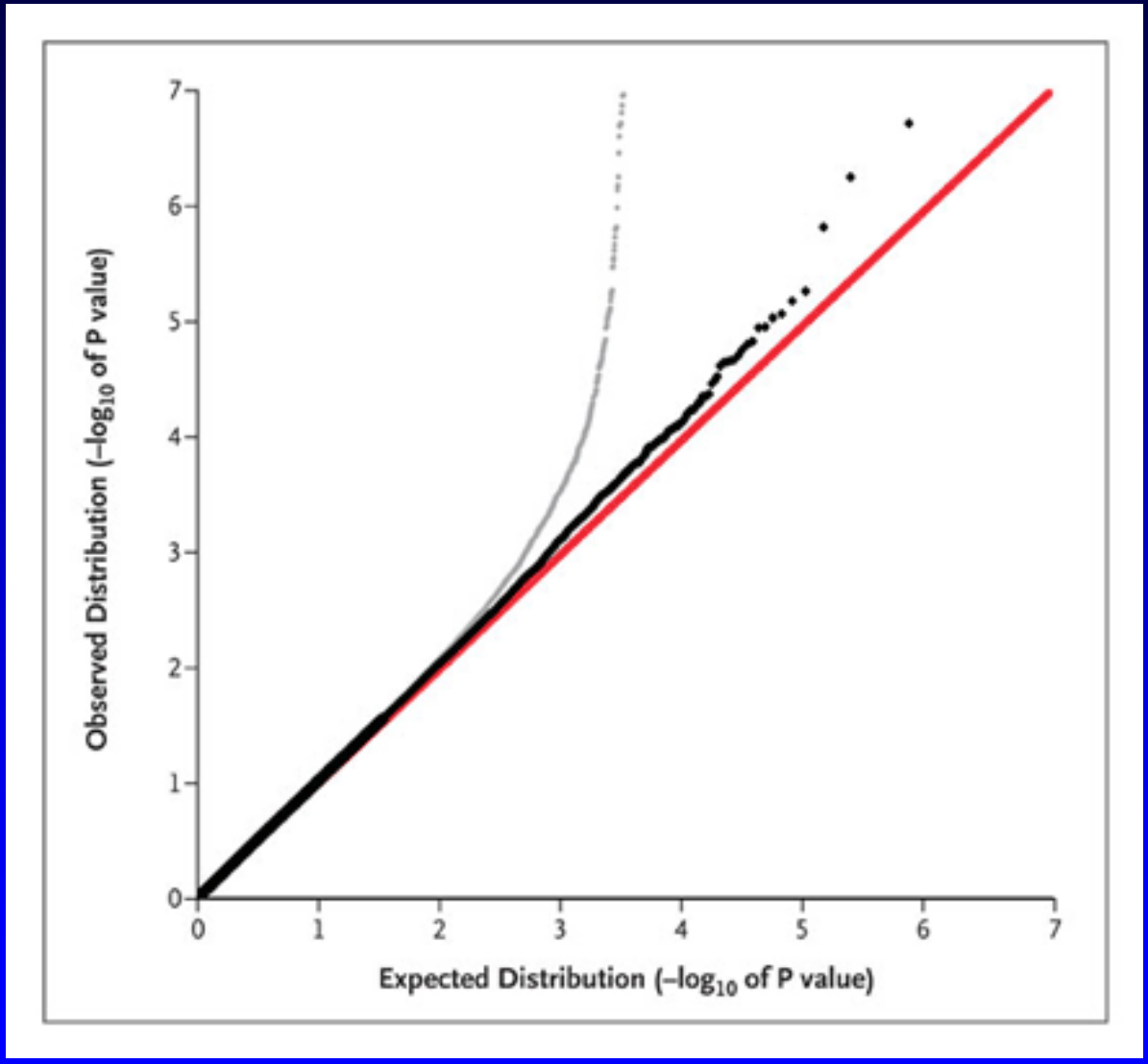
A. Select a group of RA-free controls from the same rheum. hospital and repeat her study

B. Examine a Q-Q plot for inflation of test statistics

C. Compare exposures and other characteristics of cases and controls and adjust results for those that differ

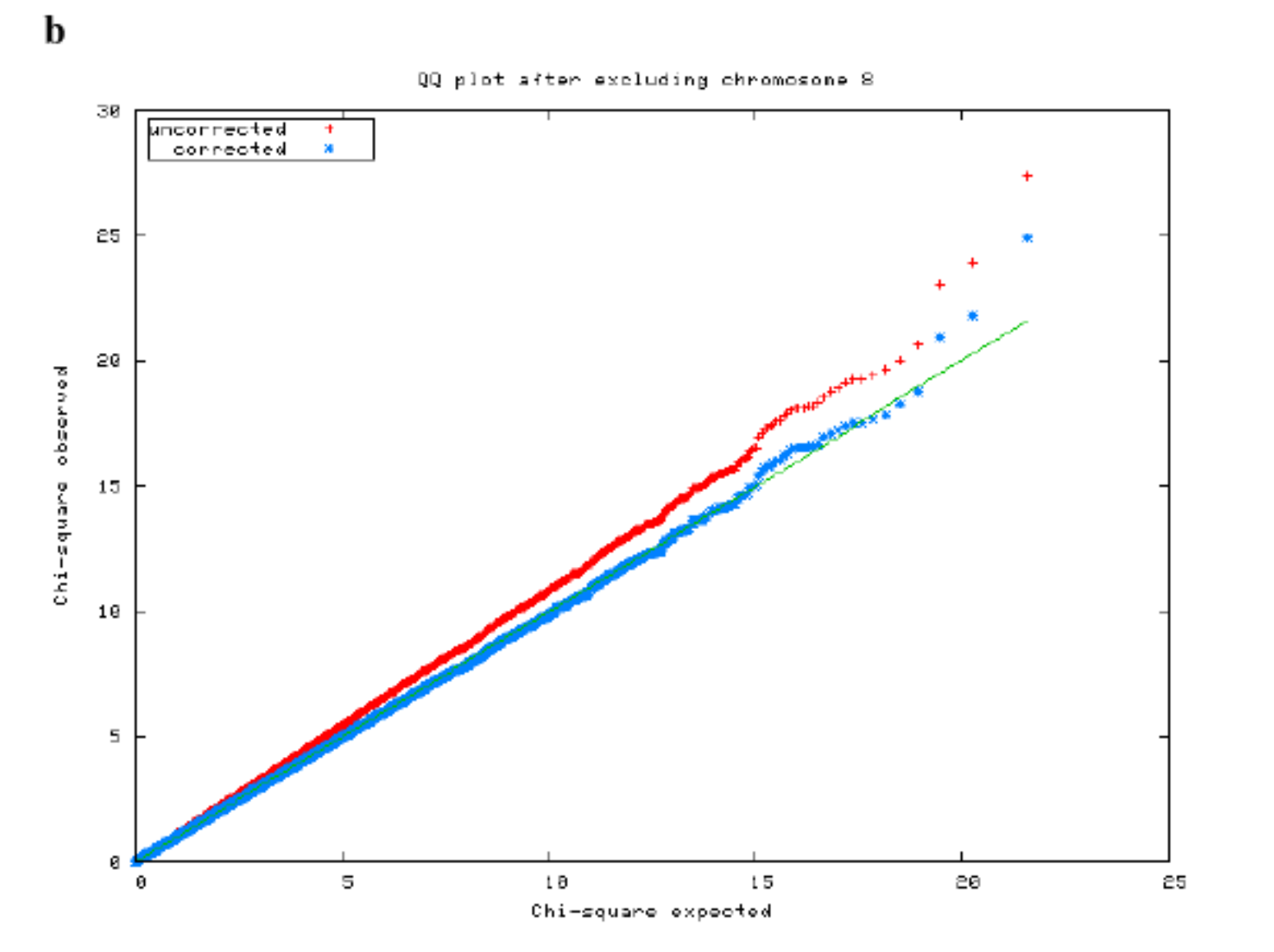
D. Compare frequencies of AIMs for evidence of population stratification

Q-Q Plot for Multiple Sclerosis



Hafler D et al, *N Engl J Med* 2007; 357:851-62.

Q-Q Plot for Prostate Cancer (excl Chr 8)



Gudmundsson J et al, *Nat Genet* 2007; 39:977-83.

Dr. X constructs a Q-Q plot and finds marked departure from the expected distribution. Her estimated λ is 1.11 and correction for it leaves 873 SNPs significant at $p < 5 \times 10^{-7}$.

What should she do now?

A. Test these 873 SNPs in a S-SRS

B. Adjust association statistics for other differences between cases and controls

C. Review genotype quality control metrics and compare association statistics at varying QC thresholds

Which ones?

Minor allele frequency

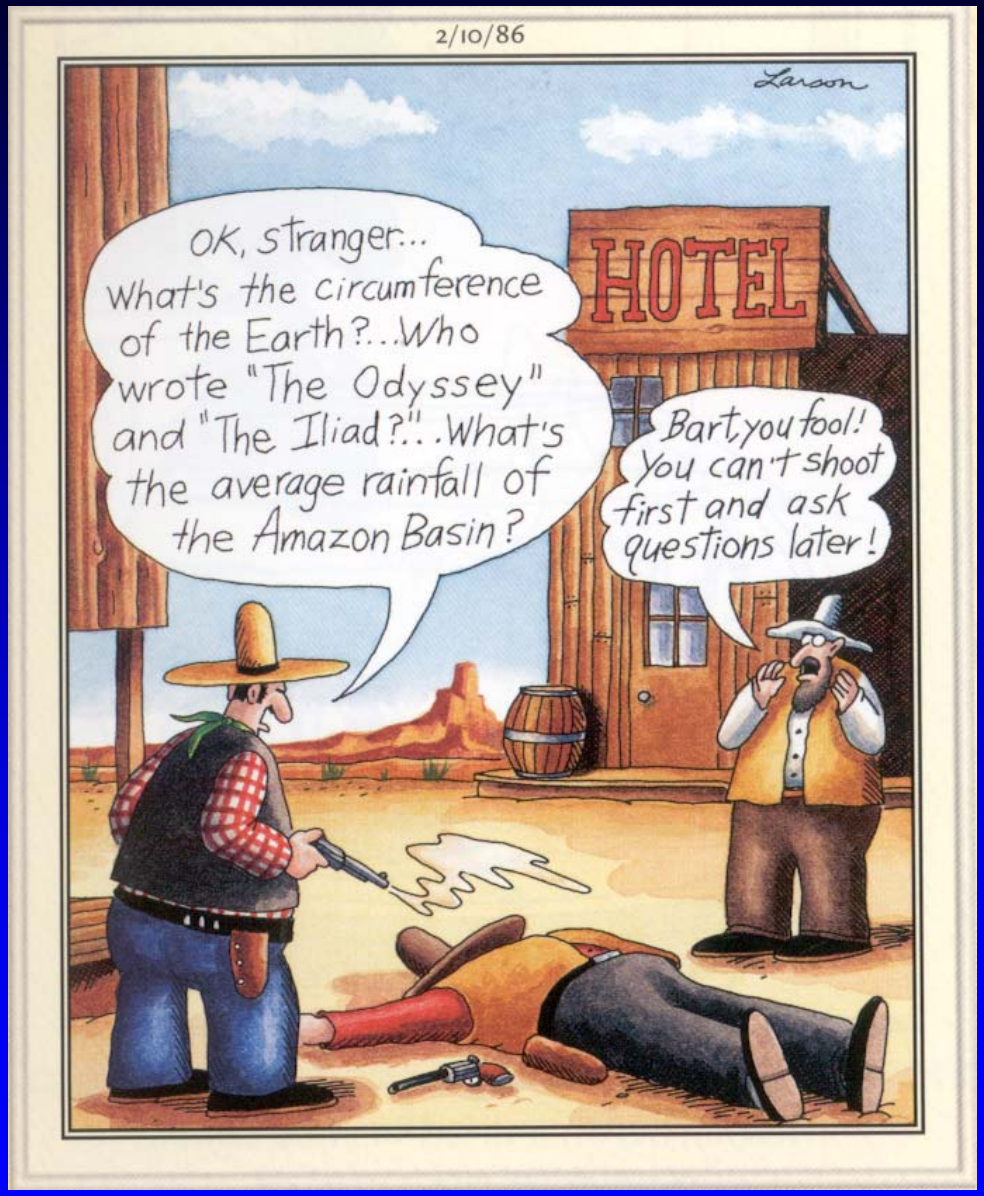
Call rate (SNPs and samples)

Heterozygosity

She discovers that nearly all the associated SNPs had low call rates and MAFs, and she filters all such SNPs out of her analysis. Among the remaining 80,000 SNPs passing these filters, there are now no SNPs significant at $p < 5 \times 10^{-7}$ though two are significant at $p < 6 \times 10^{-7}$.

What should she do now?

A. Combine her study with another GWA to increase sample size and power
B. Lower her significance threshold to allow for the reduced number of tests (to $p < 6.3 \times 10^{-7}$)
C. Increase the density of her scan to 500,000 or 1M markers
D. Examine cluster plots of these 2 SNPs
E. Test the top 4,000 SNPs in a S-SRS



Larson, G. *The Complete Far Side*. 2003.