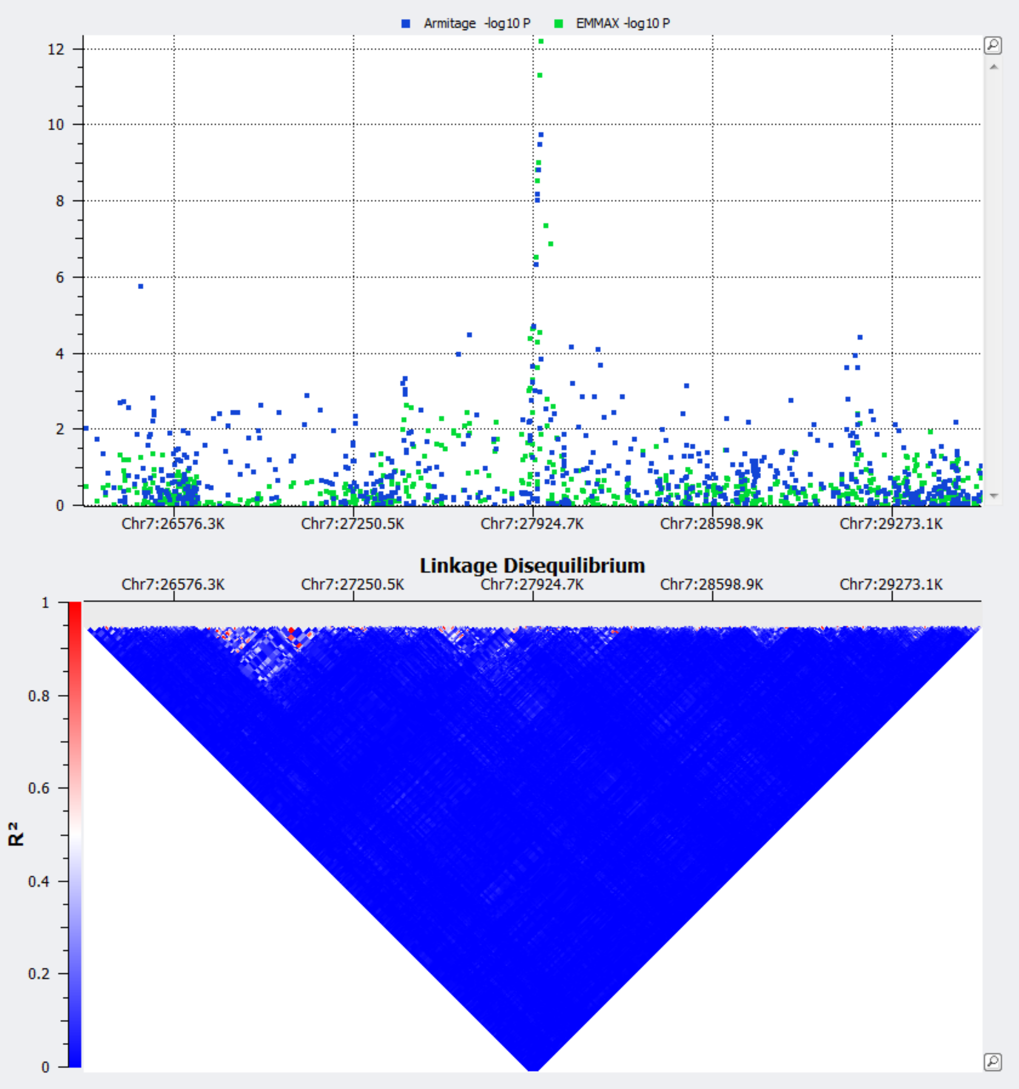


Back to Basics: Genome-Wide Association Studies

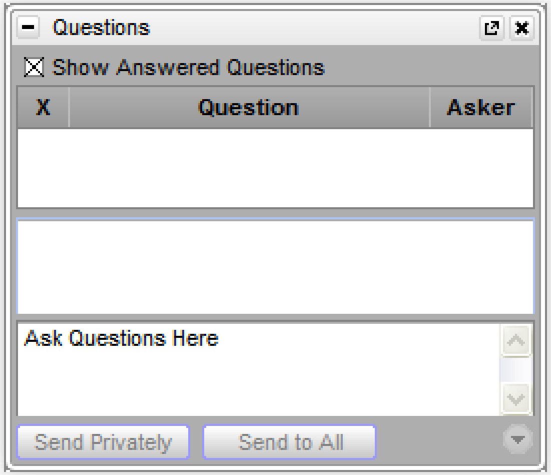
December 11, 2013

Bryce Christensen
Director of Services

Golden Helix
Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



Questions

☒ Show Answered Questions

X	Question	Asker

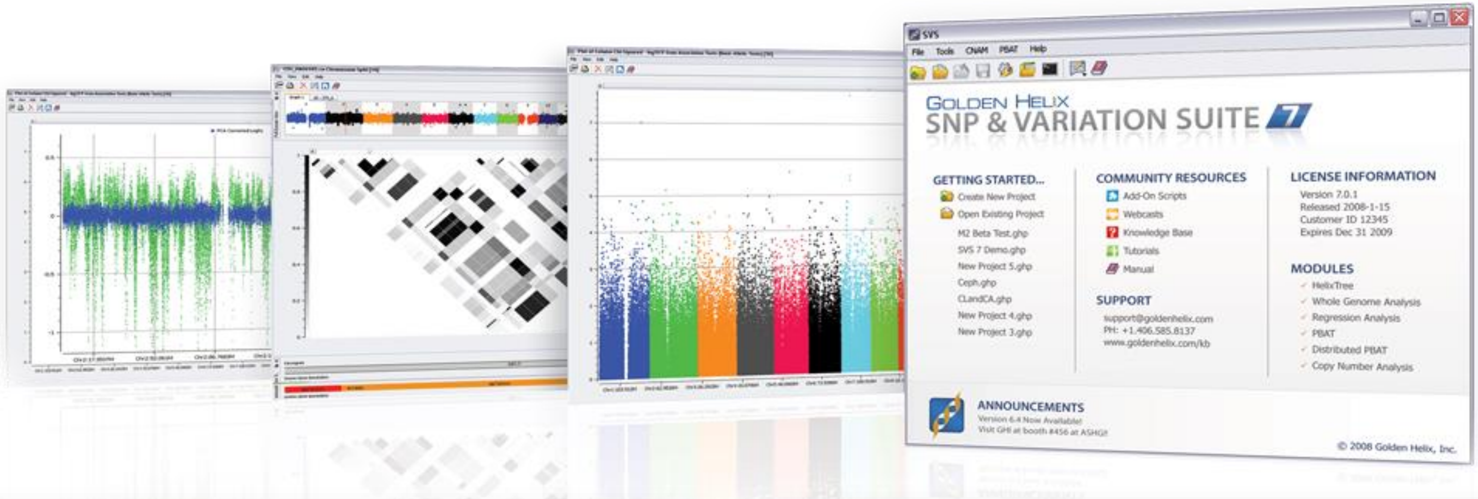
Ask Questions Here

About Golden Helix

Leaders in Genetic Analytics

- Founded in 1998
- Multi-disciplinary: computer science, bioinformatics, statistics, genetics
- Software and analytic services

SNP & Variation Suite (SVS)



GETTING STARTED...
- Create New Project
- Open Existing Project
- M2 Beta Test.ghp
- SVS 7 Demo.ghp
- New Project 5.ghp
- Ceph.ghp
- CLandCA.ghp
- New Project 4.ghp
- New Project 3.ghp

COMMUNITY RESOURCES
- Add-On Scripts
- Webcasts
- Knowledge Base
- Tutorials
- Manual

SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
www.goldenhelix.com/kb

LICENSE INFORMATION

Version 7.0.1
Released 2008-1-15
Customer ID 12345
Expires Dec 31 2009

MODULES
- ✓ HelixTree
- ✓ Whole Genome Analysis
- ✓ Regression Analysis
- ✓ PBAT
- ✓ Distributed PBAT
- ✓ Copy Number Analysis

ANNOUNCEMENTS

Version 6.4 Now Available!
Visit GHI at booth #456 at ASHG!

© 2008 Golden Helix, Inc.

Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible
- Easy-to-use

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq differential expression
- Family Based Association

Approximate Agenda

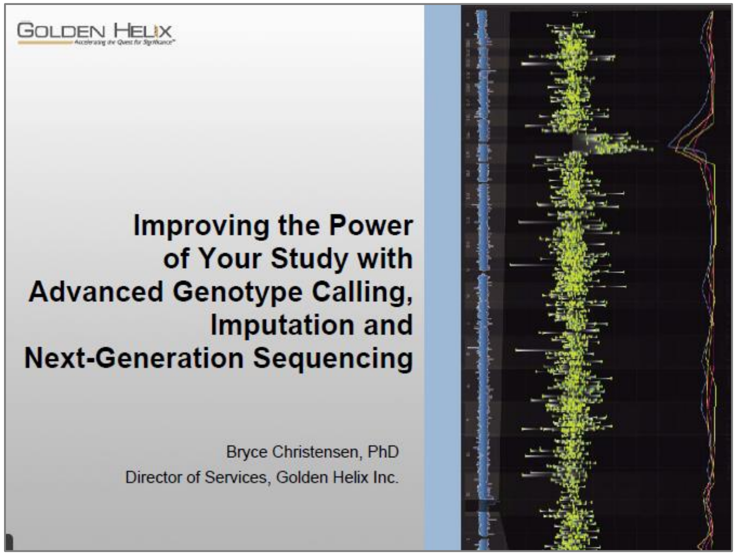
1 GWAS QC Considerations

2 Association Testing

3 What about Imputation?

4 Q&A

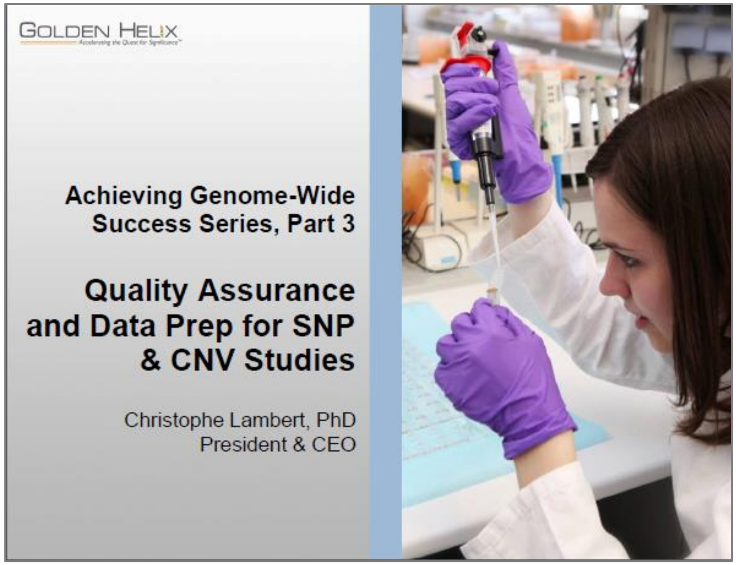
Previous GWAS Webcasts Online



Improving the Power of Your Study with Advanced Genotype Calling, Imputation and Next-Generation Sequencing

Bryce Christensen, PhD
Director of Services, Golden Helix Inc.

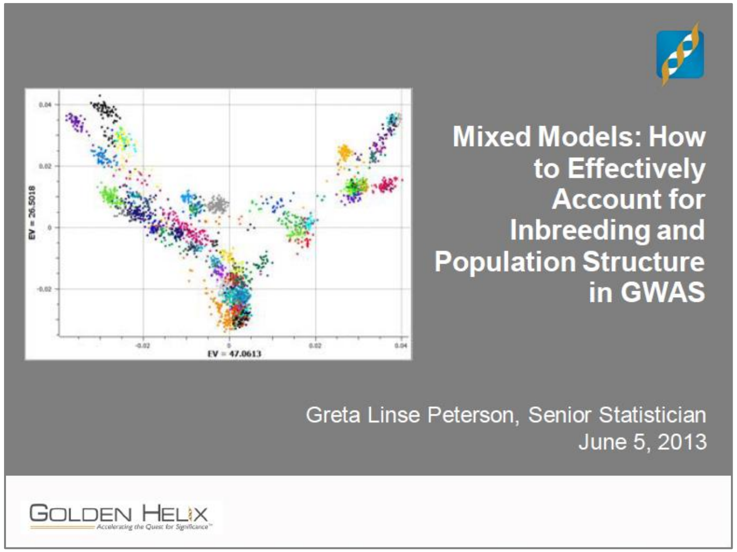
Raise Your Expectations of Agrigenomic Genetic Research Software: Introducing SNP & Variation Suite 7 (SVS)



Achieving Genome-Wide Success Series, Part 3

Quality Assurance and Data Prep for SNP & CNV Studies

Christophe Lambert, PhD
President & CEO



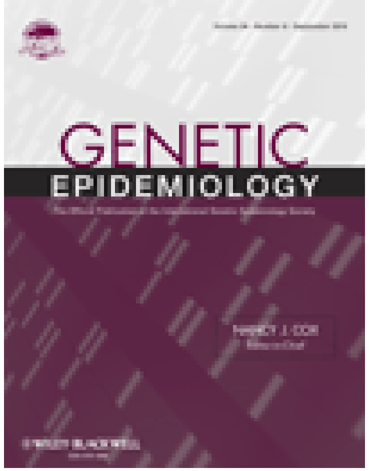
Mixed Models: How to Effectively Account for Inbreeding and Population Structure in GWAS

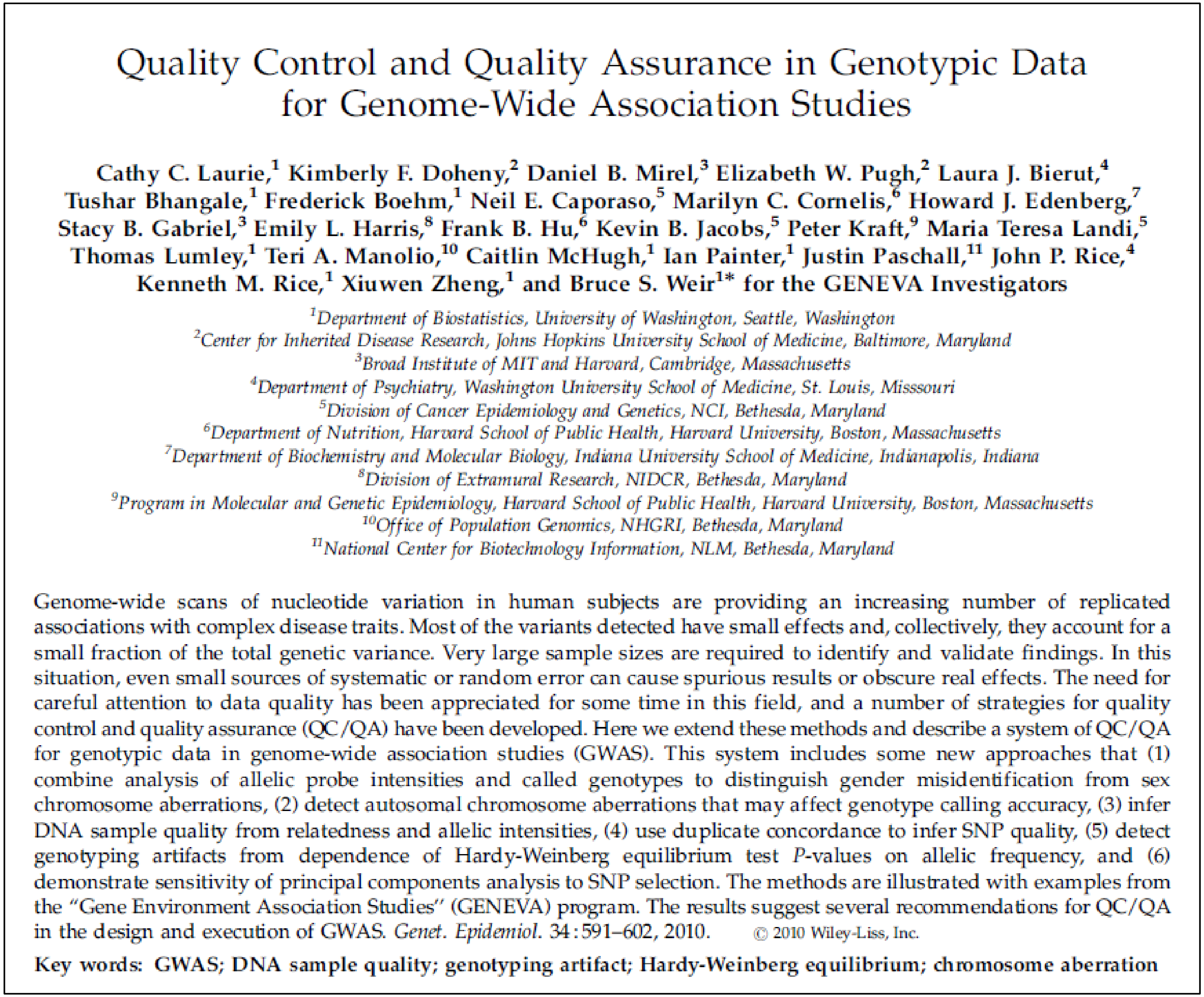
Greta Linse Peterson, Senior Statistician
June 5, 2013

GOLDEN HELIX
SNP & VARIATION SUITE 7

[Demonstration]

Genetic Epidemiology 34:591-602 (2010)





Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies

Cathy C. Laurie,[1] Kimberly F. Doheny,[2] Daniel B. Mirel,[3] Elizabeth W. Pugh,[2] Laura J. Bierut,[4] Tushar Bhangale,[1] Frederick Boehm,[1] Neil E. Caporaso,[5] Marilyn C. Cornelis,[6] Howard J. Edenberg,[7] Stacy B. Gabriel,[3] Emily L. Harris,[8] Frank B. Hu,[6] Kevin B. Jacobs,[5] Peter Kraft,[9] Maria Teresa Landi,[5] Thomas Lumley,[1] Teri A. Manolio,[10] Caitlin McHugh,[1] Ian Painter,[1] Justin Paschall,[11] John P. Rice,[4] Kenneth M. Rice,[1] Xiuwen Zheng,[1] and Bruce S. Weir[1]* for the GENEVA Investigators

[1]Department of Biostatistics, University of Washington, Seattle, Washington
[2]Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland
[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts
[4]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri
[5]Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland
[6]Department of Nutrition, Harvard School of Public Health, Harvard University, Boston, Massachusetts
[7]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana
[8]Division of Extramural Research, NIDCR, Bethesda, Maryland
[9]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Harvard University, Boston, Massachusetts
[10]Office of Population Genomics, NHGRI, Bethesda, Maryland
[11]National Center for Biotechnology Information, NLM, Bethesda, Maryland

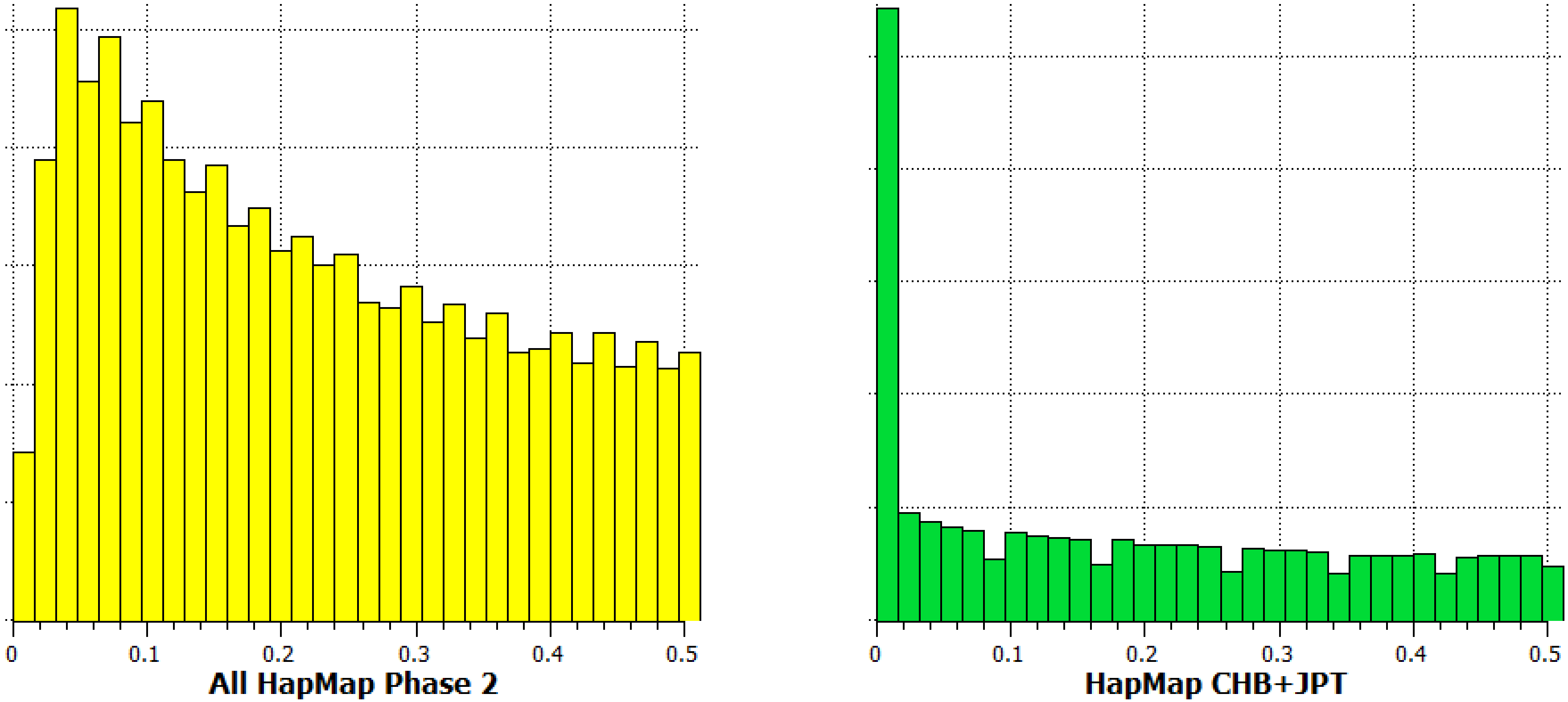
Genome-wide scans of nucleotide variation in human subjects are providing an increasing number of replicated associations with complex disease traits. Most of the variants detected have small effects and, collectively, they account for a small fraction of the total genetic variance. Very large sample sizes are required to identify and validate findings. In this situation, even small sources of systematic or random error can cause spurious results or obscure real effects. The need for careful attention to data quality has been appreciated for some time in this field, and a number of strategies for quality control and quality assurance (QC/QA) have been developed. Here we extend these methods and describe a system of QC/QA for genotypic data in genome-wide association studies (GWAS). This system includes some new approaches that (1) combine analysis of allelic probe intensities and called genotypes to distinguish gender misidentification from sex chromosome aberrations, (2) detect autosomal chromosome aberrations that may affect genotype calling accuracy, (3) infer DNA sample quality from relatedness and allelic intensities, (4) use duplicate concordance to infer SNP quality, (5) detect genotyping artifacts from dependence of Hardy-Weinberg equilibrium test *P*-values on allelic frequency, and (6) demonstrate sensitivity of principal components analysis to SNP selection. The methods are illustrated with examples from the "Gene Environment Association Studies" (GENEVA) program. The results suggest several recommendations for QC/QA in the design and execution of GWAS. *Genet. Epidemiol.* 34:591–602, 2010.

Key words: GWAS; DNA sample quality; genotyping artifact; Hardy-Weinberg equilibrium; chromosome aberration

GOLDEN HELIX SNP & VARIATION SUITE 7

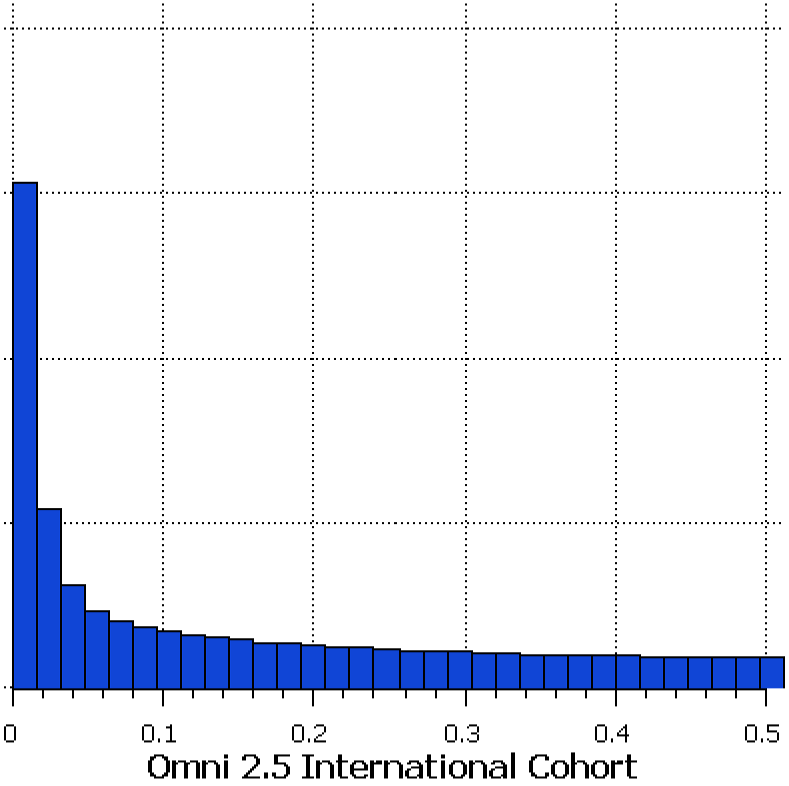
[Demonstration]

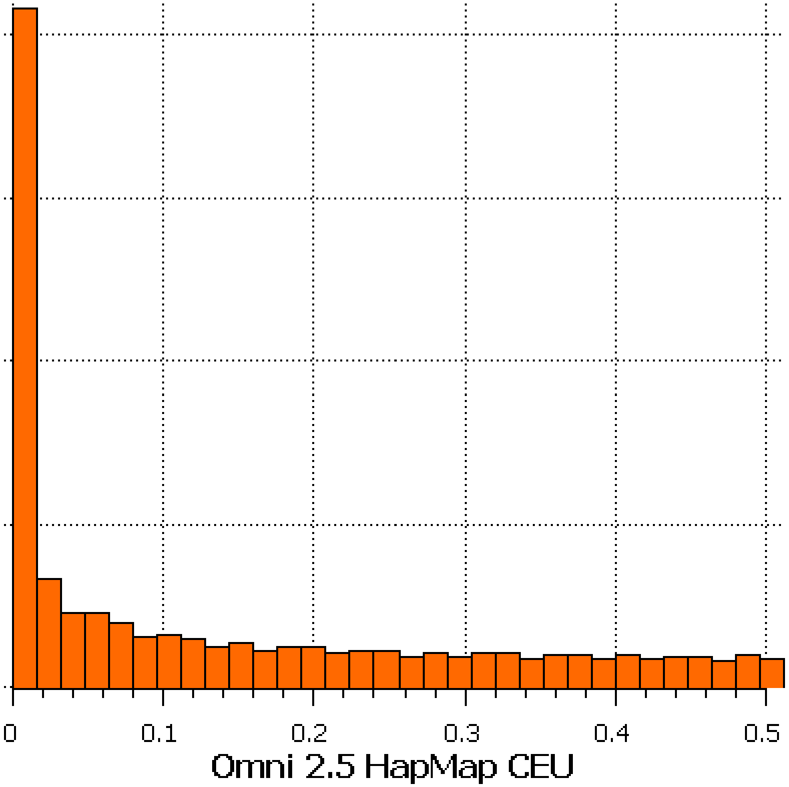
Autosomal MAF distribution on Affy 500k chip



- Most GWAS chips are designed to capture global variation
- Homogeneous cohorts will only be polymorphic for some subset of SNPs.

Autosomal MAF distribution for Illumina Omni-2.5

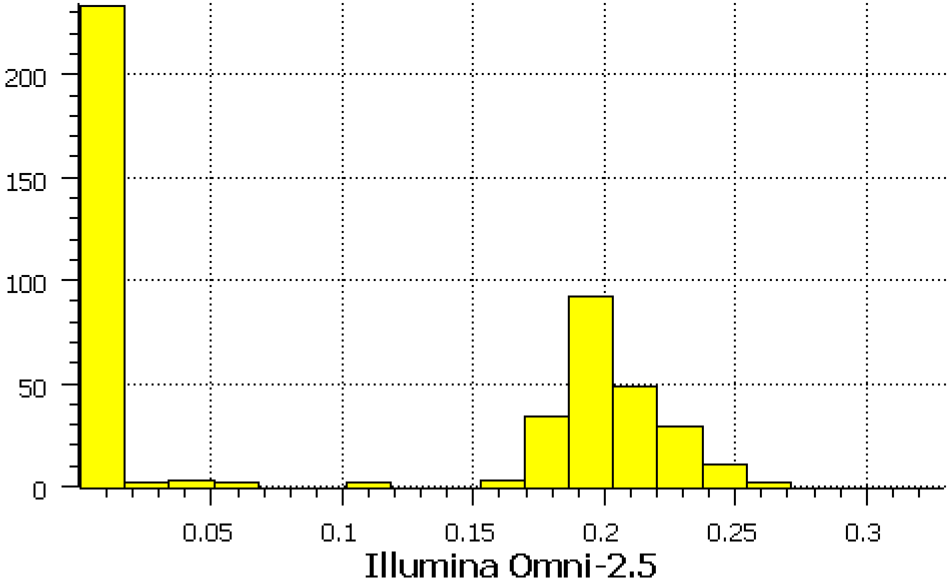


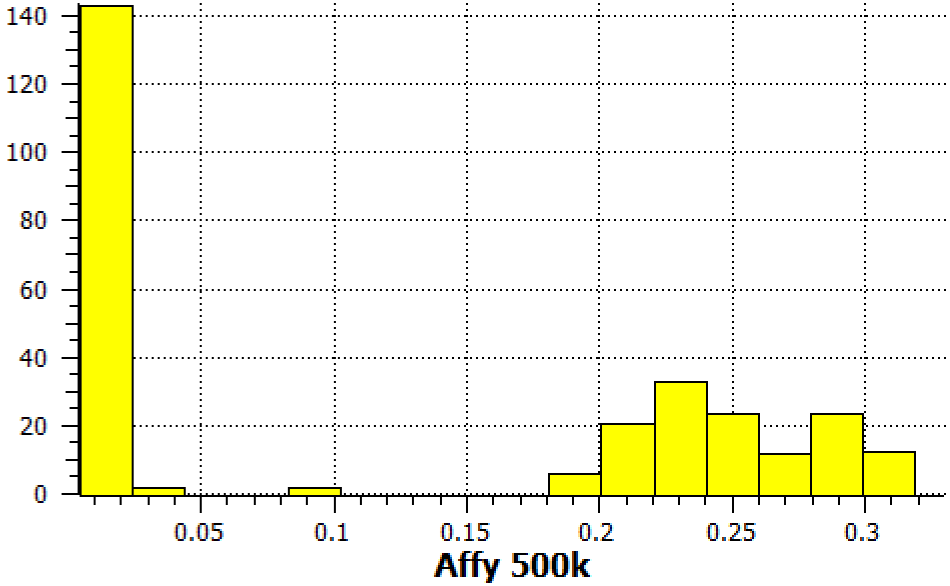


- **Higher-density chips have more rare content, so smaller relative proportion of SNPs will be polymorphic**
- **Diminishing returns with increased density**

X Heterozygosity Variability by Chip

- **Chip design can affect distribution of many statistics, including X heterozygosity**
- **Targeted chips may have minimal polymorphic content on X**
- **Adjust workflows accordingly**
  - Ex: Filter on MAF before running gender inference



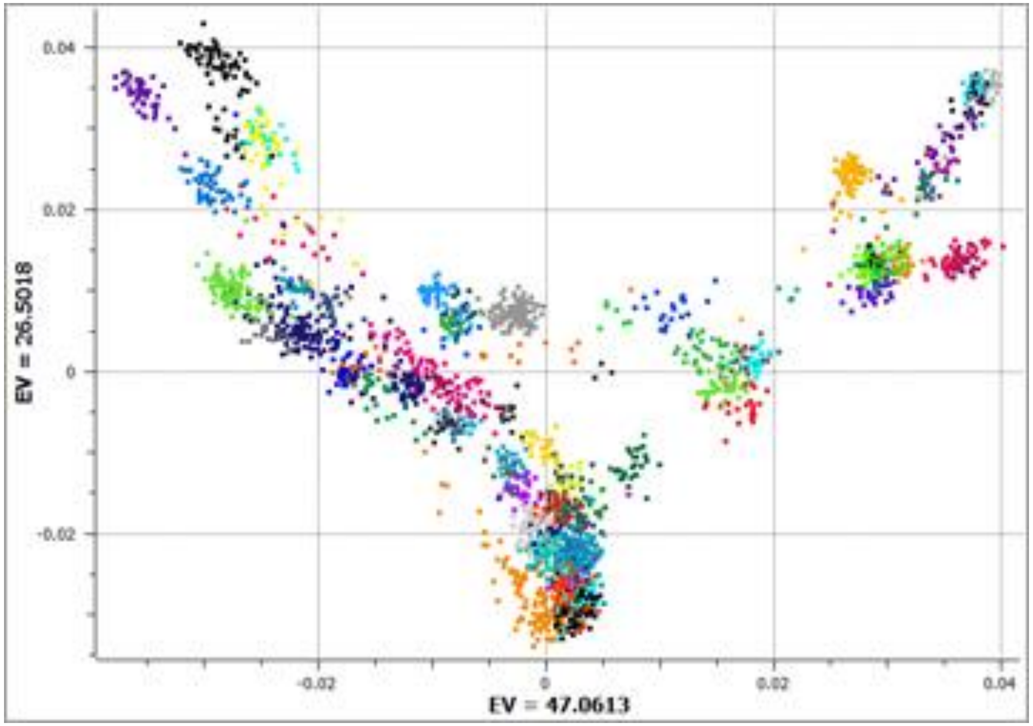


Why Care about Chip Design and Content?

- **Many sample statistics are based on allele frequencies, and behave differently from chip to chip**
  - IBD testing
  - Principal Components
  - Autosomal heterozygosity rates
  - Runs of homozygosity
- **Many of those statistics also assume that you are using a “GWAS” chip with uniform coverage, and may be confounded when using chips with targeted or non-uniform coverage content**
  - Exome chips
  - ImmunoChip
  - Cardio-MetaboChip
- **Adjust workflows accordingly!**
  - Use different MAF thresholds with targeted chips
  - Filter to polymorphic SNPs and prune for LD before running IBD or PCA

GOLDEN HELIX
SNP & VARIATION SUITE 7

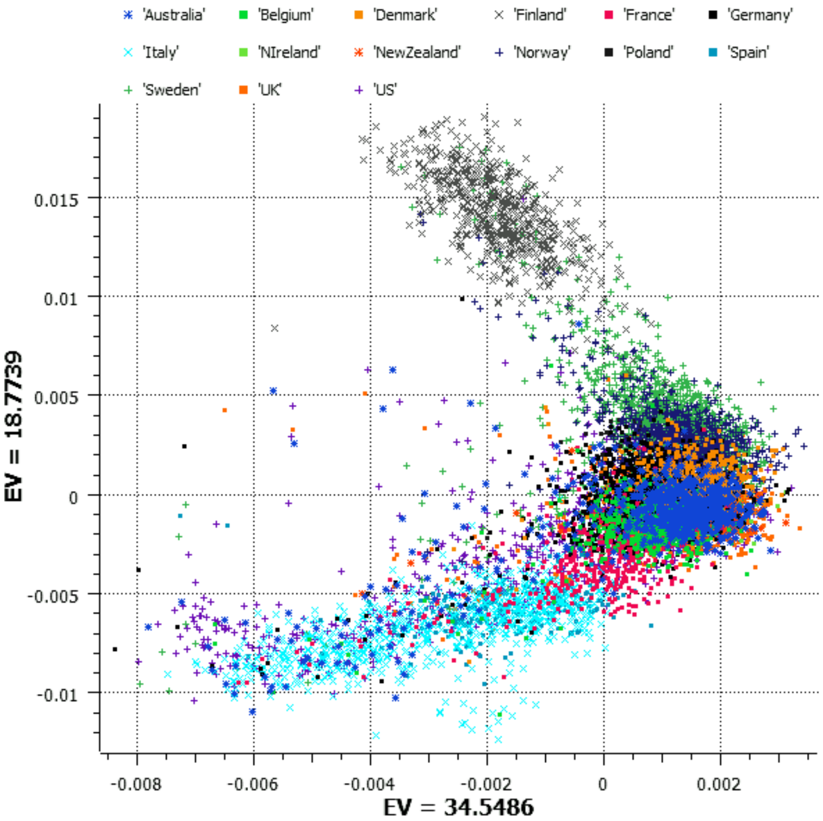
[Demonstration]



Mixed Models: How to Effectively Account for Inbreeding and Population Structure in GWAS

Original Slides by
Greta Linse Peterson, Senior Statistician

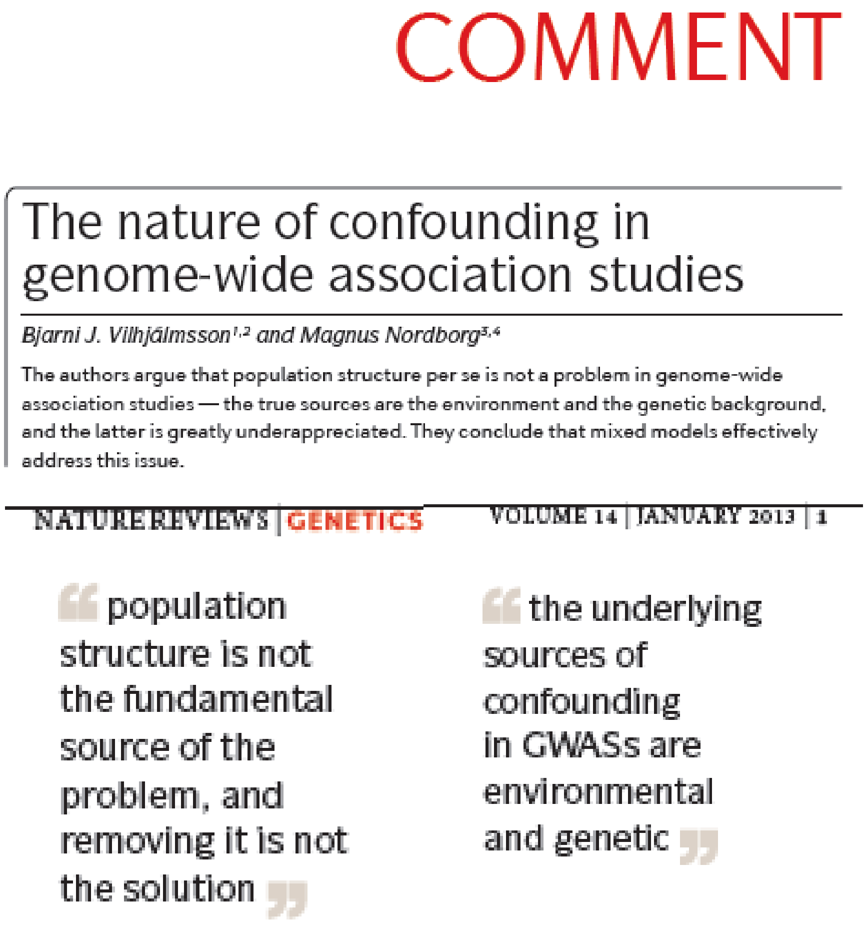
A brief background of GWAS



- First the naïve approaches: Trend Tests, Contingency Tables, Linear/Logistic Regression
- Batch Effects, Population Structure and sharing of controls may violate assumptions of the naïve approaches and result in confounding of results.
- Stratification effects are more pronounced with larger sample sizes.
- Non-independence of samples is especially problematic in agrigenomic applications.

The Real Problem

- **Vilhjalmsson and Nordborg (2013) argue that "population structure" itself is not a problem for GWAS.**
- **The real problems are the environment and the genetic background of a population.**
  - PCA can serve as a proxy for both, but doesn't entirely explain either.
- **The solution is to account for the relatedness between all pairs of samples in a mixed linear model.**



COMMENT

The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] and Magnus Nordborg[3,4]

The authors argue that population structure per se is not a problem in genome-wide association studies — the true sources are the environment and the genetic background, and the latter is greatly underappreciated. They conclude that mixed models effectively address this issue.

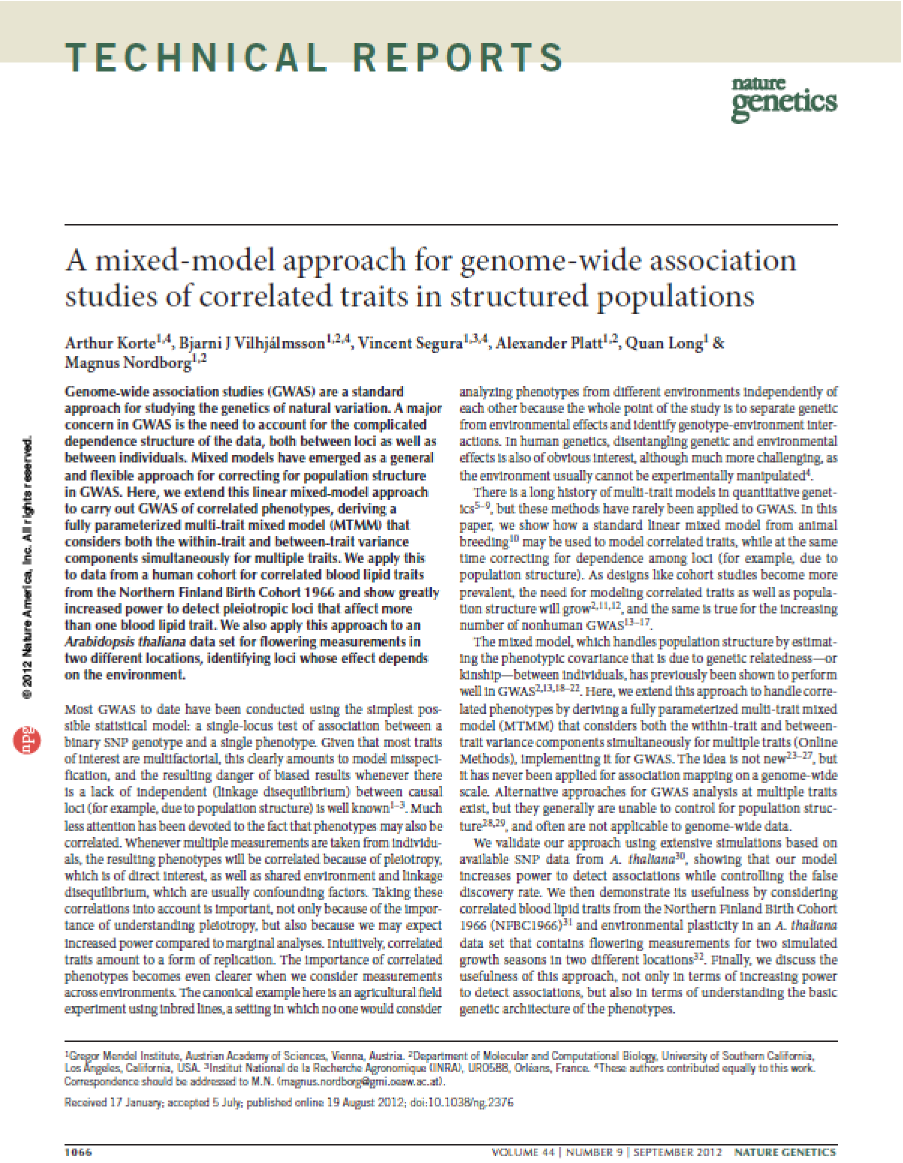
NATURE REVIEWS | GENETICS VOLUME 14 | JANUARY 2013 | 1

"population structure is not the fundamental source of the problem, and removing it is not the solution"

"the underlying sources of confounding in GWASs are environmental and genetic"

Mixed Model Method Overview

- Calculate kinship matrix defining pairwise relationships between all sample pairs.
- Include kinship matrix as random effect in MLM regression.
- May also include PCs and other factors as fixed effects.
- Allows for population-based and family-based cohorts to be analyzed together.



TECHNICAL REPORTS

nature genetics

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte[1,4], Bjarni J Vilhjálmsson[1,2,4], Vincent Segura[1,3,4], Alexander Platt[1,2], Quan Long[1] & Magnus Nordborg[1,2]

Genome-wide association studies (GWAS) are a standard approach for studying the genetics of natural variation. A major concern in GWAS is the need to account for the complicated dependence structure of the data, both between loci as well as between individuals. Mixed models have emerged as a general and flexible approach for correcting for population structure in GWAS. Here, we extend this linear mixed-model approach to carry out GWAS of correlated phenotypes, deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits. We apply this to data from a human cohort for correlated blood lipid traits from the Northern Finland Birth Cohort 1966 and show greatly increased power to detect pleiotropic loci that affect more than one blood lipid trait. We also apply this approach to an *Arabidopsis thaliana* data set for flowering measurements in two different locations, identifying loci whose effect depends on the environment.

Most GWAS to date have been conducted using the simplest possible statistical model: a single-locus test of association between a binary SNP genotype and a single phenotype. Given that most traits of interest are multifactorial, this clearly amounts to model misspecification, and the resulting danger of biased results whenever there is a lack of independent (linkage disequilibrium) between causal loci (for example, due to population structure) is well known[1–3]. Much less attention has been devoted to the fact that phenotypes may also be correlated. Whenever multiple measurements are taken from individuals, the resulting phenotypes will be correlated because of pleiotropy, which is of direct interest, as well as shared environment and linkage disequilibrium, which are usually confounding factors. Taking these correlations into account is important, not only because of the importance of understanding pleiotropy, but also because we may expect increased power compared to marginal analyses. Intuitively, correlated traits amount to a form of replication. The importance of correlated phenotypes becomes even clearer when we consider measurements across environments. The canonical example here is an agricultural field experiment using inbred lines, a setting in which no one would consider analyzing phenotypes from different environments independently of each other because the whole point of the study is to separate genetic from environmental effects and identify genotype-environment interactions. In human genetics, disentangling genetic and environmental effects is also of obvious interest, although much more challenging, as the environment usually cannot be experimentally manipulated[4].

There is a long history of multi-trait models in quantitative genetics[5–9], but these methods have rarely been applied to GWAS. In this paper, we show how a standard linear mixed model from animal breeding[10] may be used to model correlated traits, while at the same time correcting for dependence among loci (for example, due to population structure). As designs like cohort studies become more prevalent, the need for modeling correlated traits as well as population structure will grow[2,11,12], and the same is true for the increasing number of nonhuman GWAS[13–17].

The mixed model, which handles population structure by estimating the phenotypic covariance that is due to genetic relatedness—or kinship—between individuals, has previously been shown to perform well in GWAS[2,13,18–22]. Here, we extend this approach to handle correlated phenotypes by deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits (Online Methods), implementing it for GWAS. The idea is not new[23–27], but it has never been applied for association mapping on a genome-wide scale. Alternative approaches for GWAS analysis at multiple traits exist, but they generally are unable to control for population structure[28,29], and often are not applicable to genome-wide data.

We validate our approach using extensive simulations based on available SNP data from *A. thaliana*[30], showing that our model increases power to detect associations while controlling the false discovery rate. We then demonstrate its usefulness by considering correlated blood lipid traits from the Northern Finland Birth Cohort 1966 (NFBC1966)[31] and environmental plasticity in an *A. thaliana* data set that contains flowering measurements for two simulated growth seasons in two different locations[32]. Finally, we discuss the usefulness of this approach, not only in terms of increasing power to detect associations, but also in terms of understanding the basic genetic architecture of the phenotypes.

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. [2]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [3]Institut National de la Recherche Agronomique (INRA), UR0588, Orléans, France. [4]These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).

Received 17 January; accepted 5 July; published online 19 August 2012; doi:10.1038/ng.2376

Applied Example: WTCCC-2 Multiple Sclerosis

- About 10k cases and 17k controls from world-wide Caucasian populations
- Naïve GWAS: λ=2.48
- PCA adjusted: λ=1.21
- Stratified analysis in ancestry-matched subgroups, with results combined in meta analysis: λ=1.44
- MLM approach: λ=1.04!

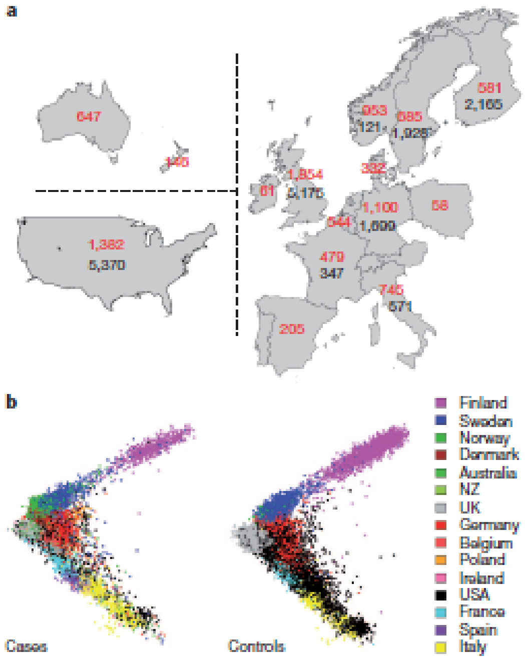


Figure 1 | Distribution of cases and controls. a, b, All cases and controls were drawn from populations with European ancestry; cases from 15 countries and controls from 8. **a,** Numbers of case (red) and control (black) samples from each country. **b,** The projection of samples onto the first two principal components of genetic variation, with cases shown on the left and controls on the right. The axes are orientated to approximate the geography, and samples are colour coded as indicated in the legend. NZ, New Zealand. We genotyped the cases (9,772) and some Swedish controls (527) using the Illumina Human 660-Quad platform, and the UK controls (5,175, the WTCCC2 common control set[12,13]) using the Illumina 1.2M platform. All other controls were genotyped externally using various Illumina genotyping systems (see Supplementary Information).

11 AUGUST 2011 | VOL 476 | NATURE

Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis

The International Multiple Sclerosis Genetics Consortium* & the Wellcome Trust Case Control Consortium 2*

Methods Implemented in SVS

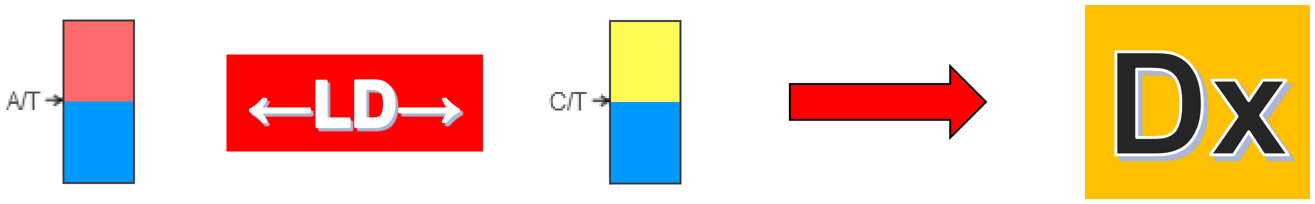
- **Regression with PCA Correction**
  - Accounts for the relationship between samples with Principal Components
  - Need to know how many components to correct for
- **EMMAX**
  - Adjusts for the pair-wise relationship between all samples using a kinship matrix
  - Approximates the variance components and uses the same variance for all probes
  - Tests a single locus at a time
- **MLMM**
  - Adjusts for the pair-wise relationship between all samples using a kinship matrix
  - Approximates the variance components and uses the same variance for all probes, but re-computes at every step
  - Stepwise EMMAX, assumes multiple loci are associated with the phenotype
- **GBLUP**
  - Adjusts for the pair-wise relationship between all samples using a kinship matrix
  - Computes allele substitution effects to determine best genomic predictors of the phenotype

GOLDEN HELIX

SNP & VARIATION SUITE 7

[Demonstration]

Standard GWAS is based on tag-SNPs



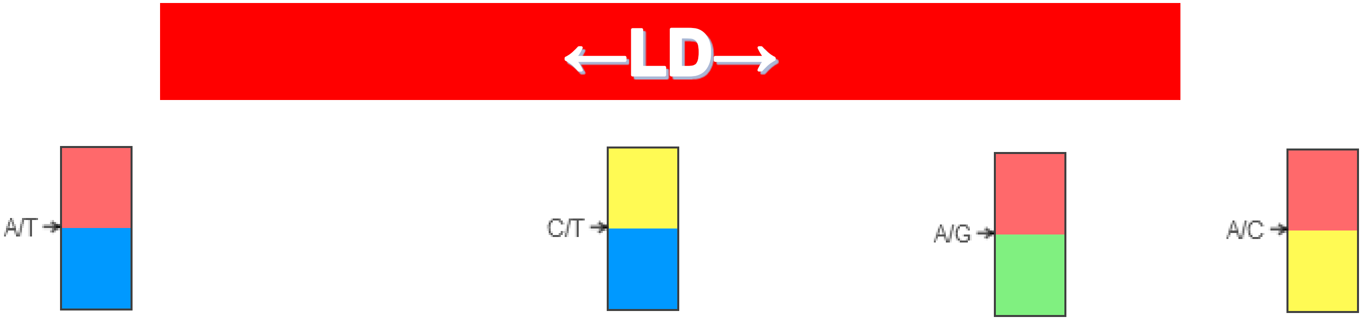
A. Observed GWAS SNP

B. Untyped Causal SNP

C. Disease Outcome

- **We typically test for the relationship between A and C, assuming that B probably won’t be on the array.**
- **BUT: Correlation is not transitive.**
  - If A is correlated with B, and B with C, A is not necessarily correlated with C.
- **Is that a problem?**
- **What does it mean for imputation?**

Imputation Implications



A. Observed GWAS SNP

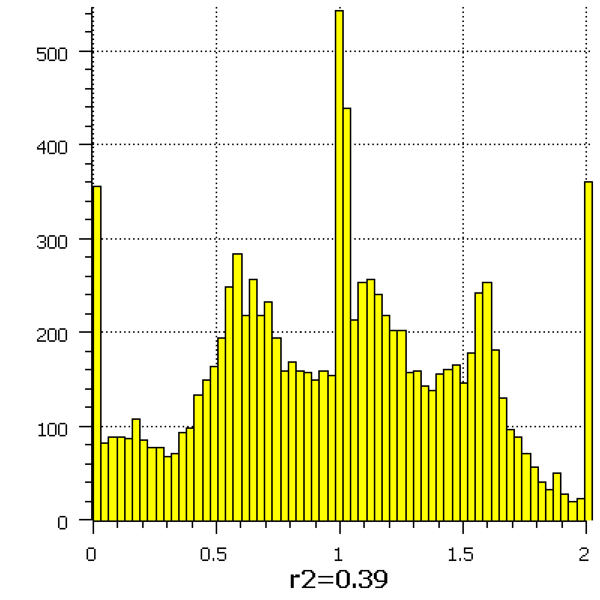
B. Untyped Causal SNP

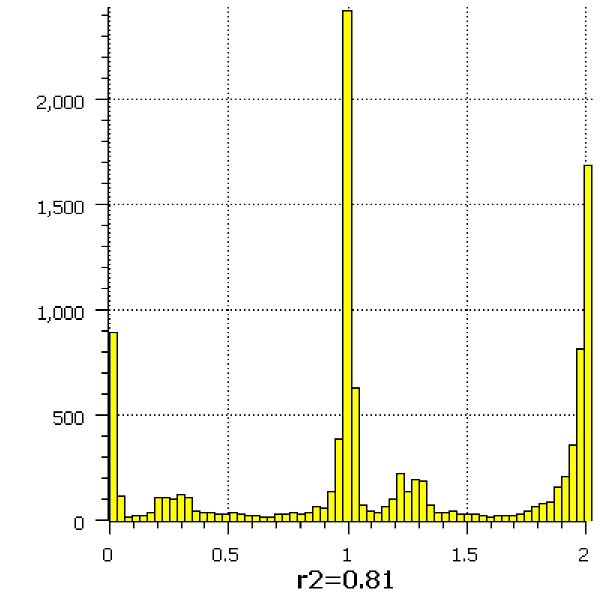
D. Additional GWAS SNPs in LD with **B.**

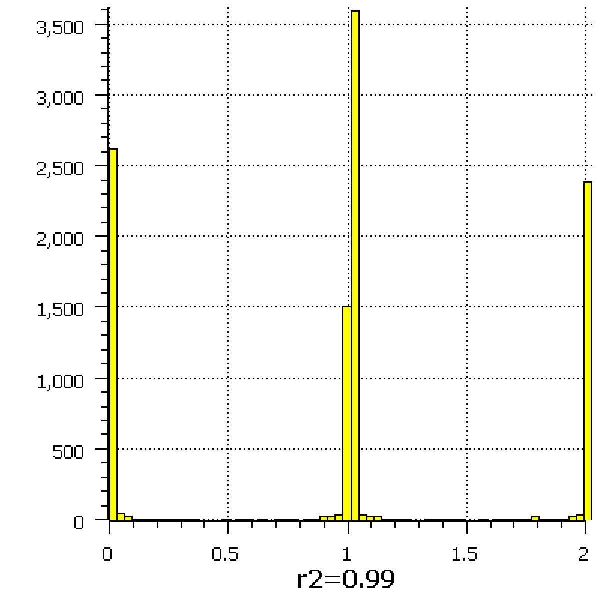
- **Imputation accuracy is usually improved when several GWAS SNPs contribute to the imputed genotype of a given variant.**
- **Testing disease association with accurately imputed variants is the best available alternative to sequencing, and much cheaper.**
- **As always: Carefully follow up on any significant results!**

Imputation: What to Watch For

- Accuracy metrics from imputation software
- Always look for inter-cohort differences
- Example: Beagle's Allelic R^2 stat. Look at the allelic dosage histograms:

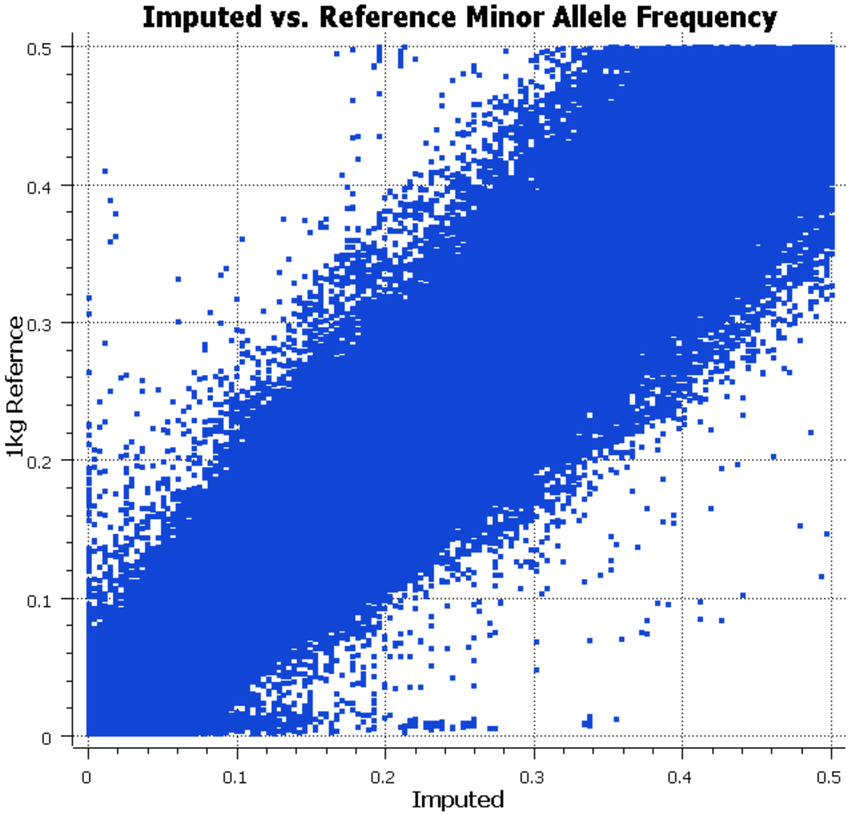






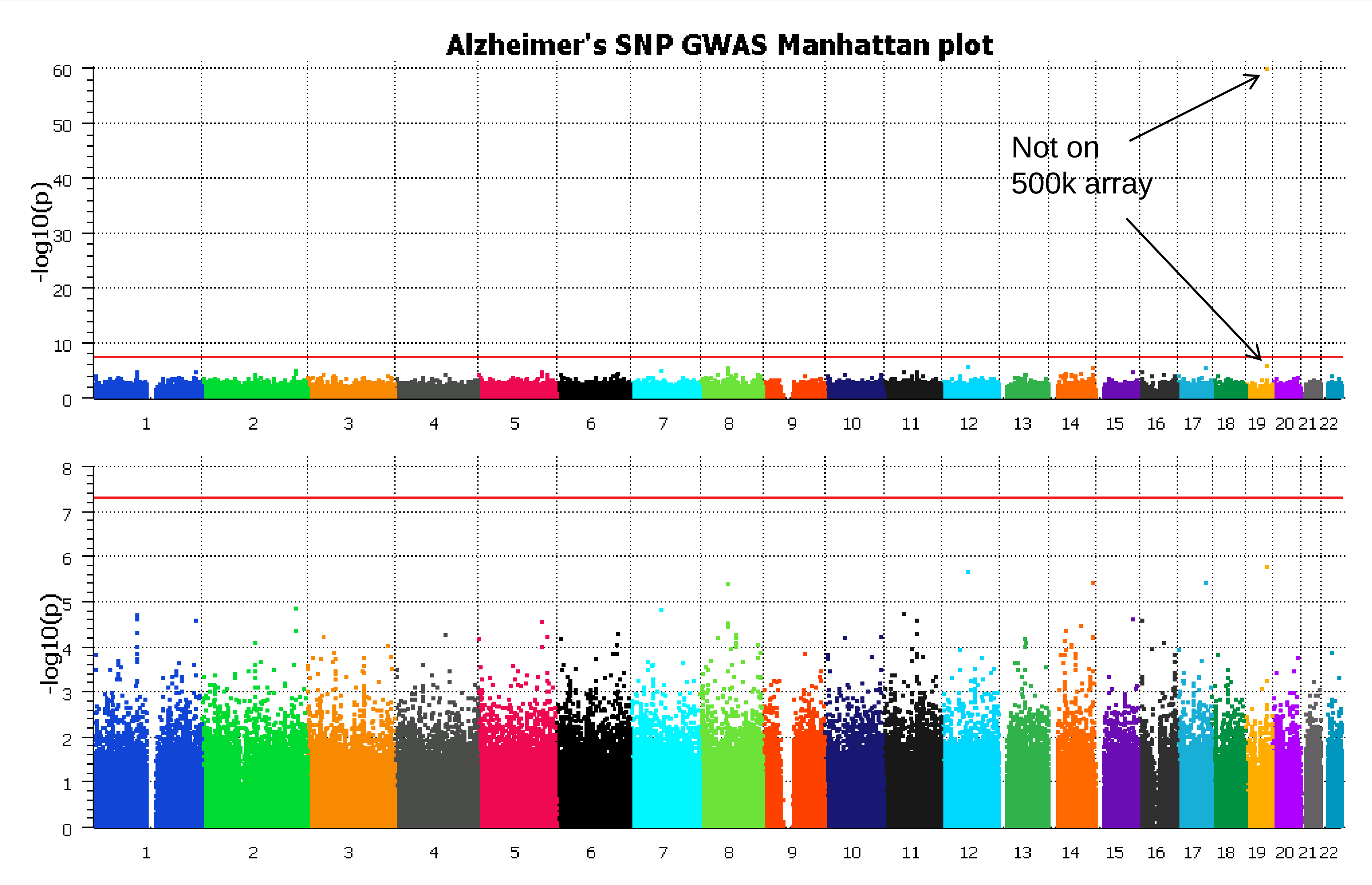
Imputation: What to Watch For

- **Imputed allele frequencies different from reference panel frequencies**
  - Especially when common alleles are imputed with 0 frequency.
- **Watch for inter-cohort differences.**



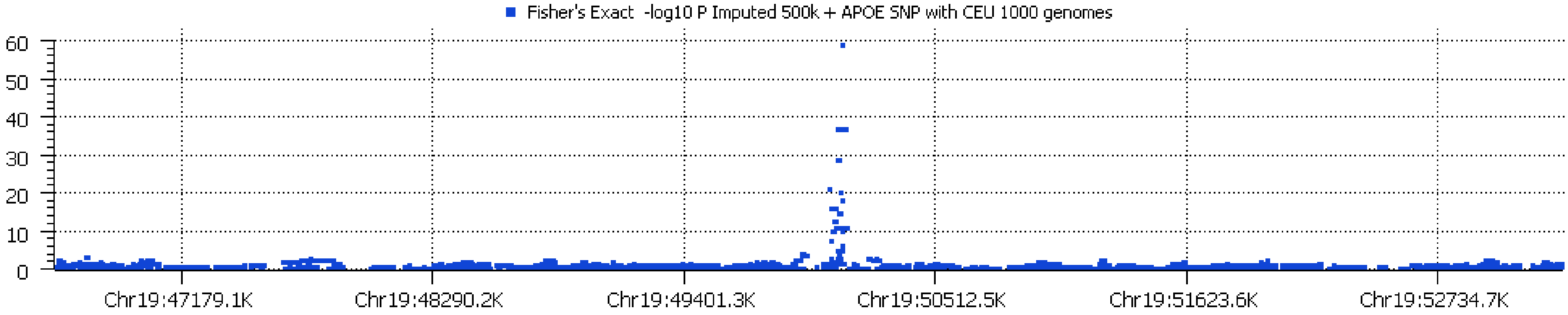
rs429358 Significant But Not Assayed By 500k

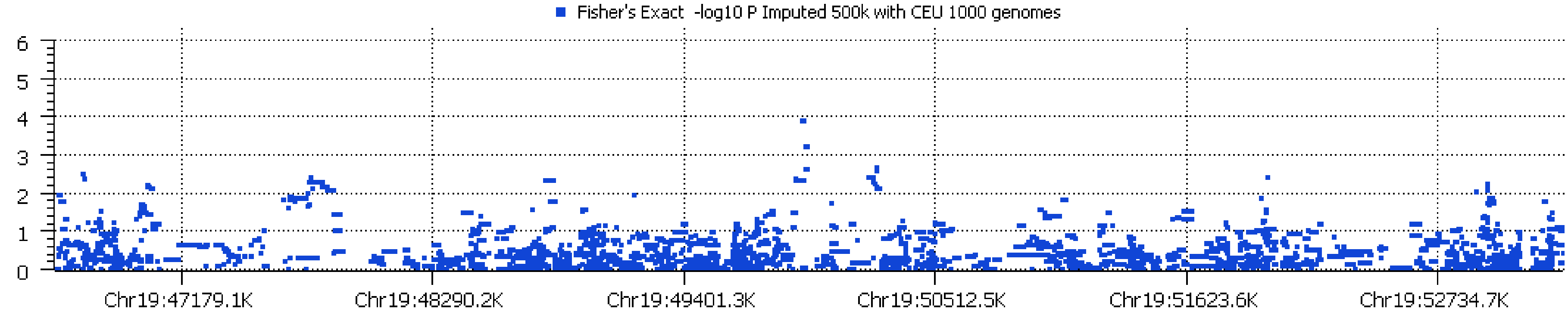
Would 1000 Genomes imputation have found it?

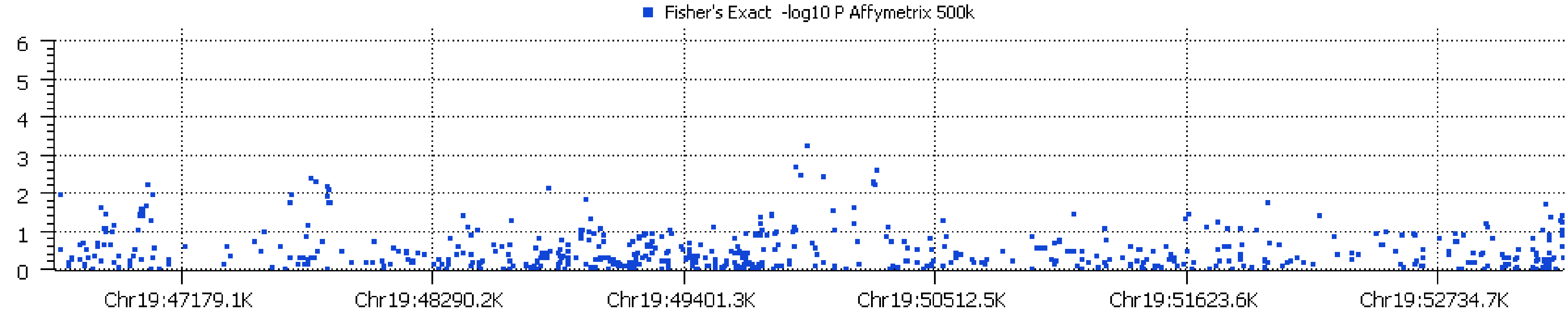


Not on 500k array

Alzheimer’s 1kG Imputation Results





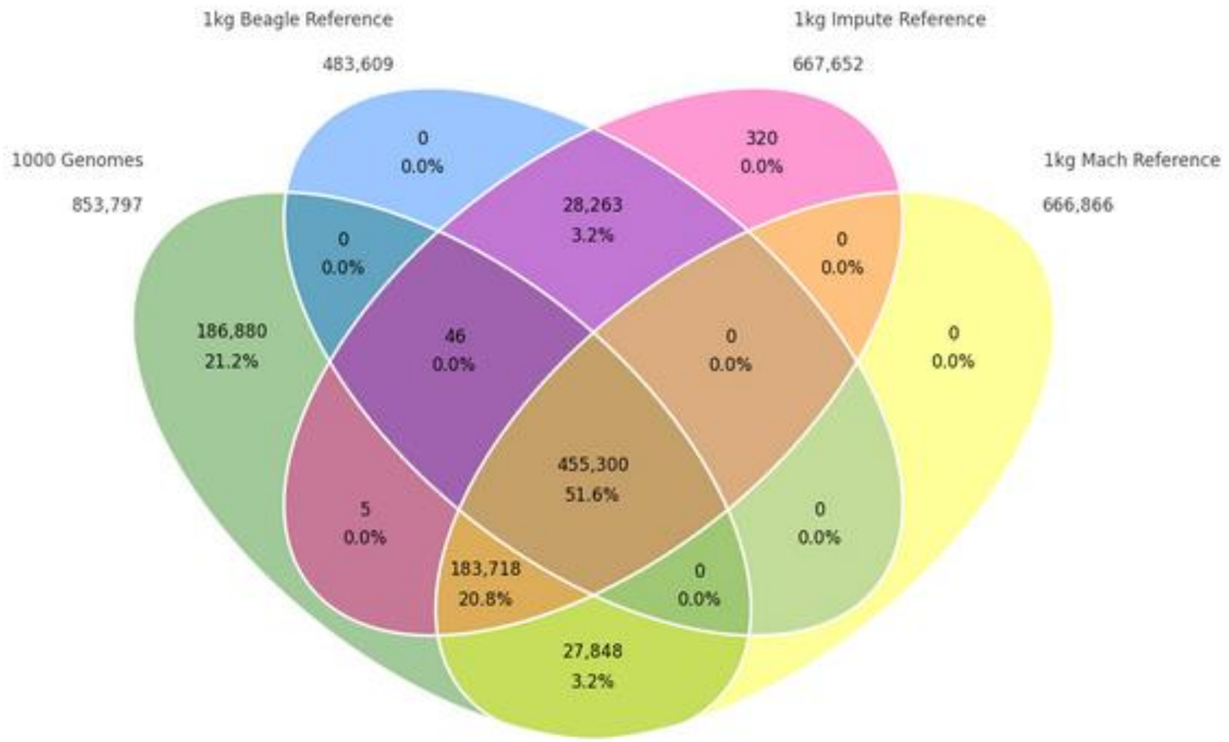


Genotype Imputation: Why?

- **Fill in the blanks—improve SNP call rate in GWAS**
  - This is where imputation started
- **Increase density of genotype calls**
  - Define and/or refine the search space for identifying candidate causal variants around GWAS signals
- **Harmonize different array platforms for mega-analysis or meta-analysis**
  - Additional power to be gained from increased sample sizes
  - Very common in disease consortia
- **Identify new associations not observed in GWAS?**
  - Rare, but possible to identify a new locus
  - Remember: our reference panels are usually made up of healthy people…

SVS and Imputation

- SVS does not have an imputation algorithm
- Add-on functions available to read and write file formats used by Beagle, MACH/Minimac and Impute2.
- SVS supports analysis of imputed genotypes, including allelic dosage formats



Review

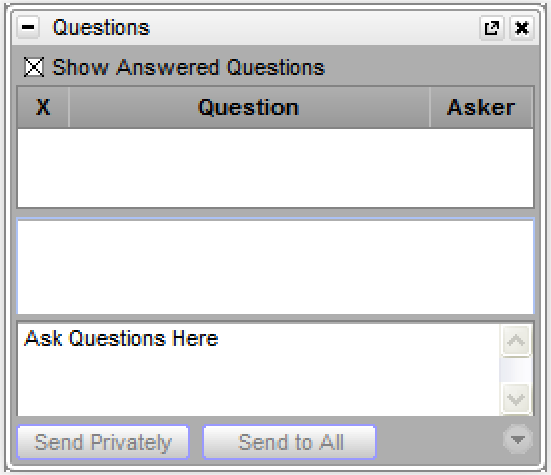
- **GWAS is not dead**
- **Golden Helix SVS is a powerful platform for GWAS analysis**
  - Data management
  - Quality Assurance
  - Visualization
  - Association Testing
  - LD & Haplotype Analysis
- **New analysis methods like mixed model regression continue to improve GWAS quality**
- **Imputation is very powerful, but has limitations**
- **Look for new GWAS features in SVS 8.1!**

Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com
- Check out our abstract competition!

Questions?

Use the Questions pane in your GoToWebinar window



Questions
☒ Show Answered Questions
X
Question
Asker
Ask Questions Here
Send Privately
Send to All