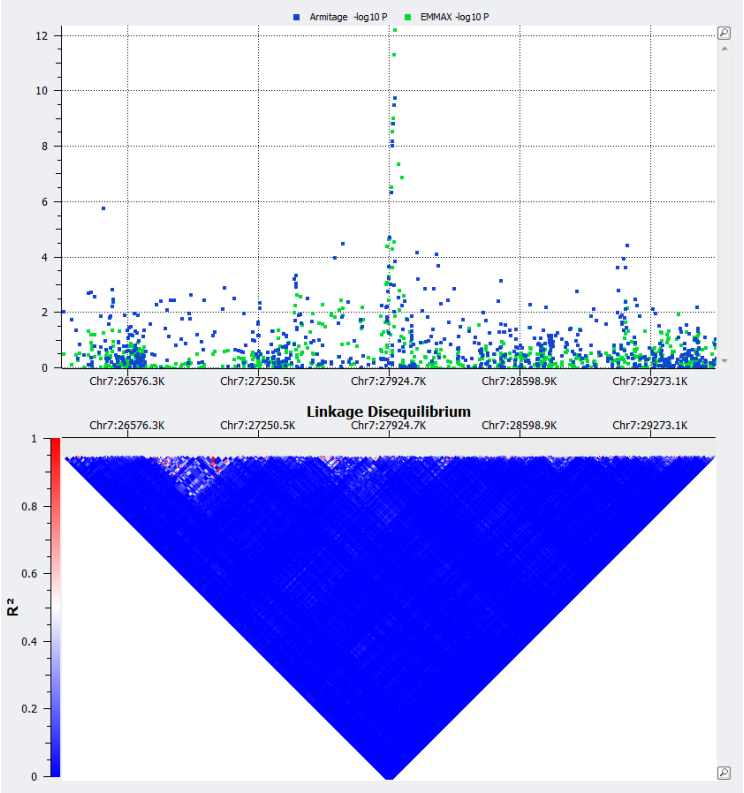
GOLDEN HELIX

Enabling Precision Medicine



New Enhancements: GWAS Workflows with SVS

August 9th, 2017

Gabe Rudy

VP Product & Engineering

CIOReview

20 most promising Biotech Technology Providers

pharma TECH OUTLOOK

Top 10 Analytics Solution Providers

Gartner

Hype Cycle for Life sciences

Golden Helix – Who We Are

Golden Helix is a global bioinformatics company founded in 1998.

GlaxoSmithKline

varSEQ®

- Variant Calling
- Filtering and Annotation
- Clinical Reports
- CNV Analysis
- Pipeline: Run Workflows

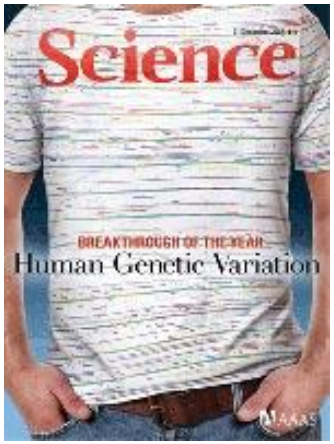
VARIANT WAREHOUSE

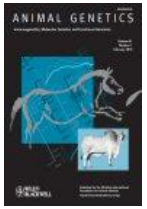
- Variant Warehouse
- Centralized Annotations
- Hosted Reports
- Sharing and Integration

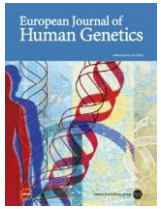
SNP & VARIATION SUITE

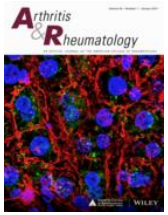
- GWAS
- Genomic Prediction
- Large-N Population Studies
- Large-N CNV-Analysis

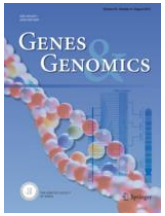
Cited in over 1100 peer-reviewed publications











GOLDEN HELIX
Enabling Precision Medicine

Over 350 customers globally

SickKids
THE HOSPITAL FOR SICK CHILDREN

Cincinnati Children's
Hospital Medical Center

The Children's Hospital of Philadelphia

Stanford University

UBC

Ucla

Yale

The Feinstein Institute for Medical Research
North Shore-Long Island Jewish Health System

North Shore LIJ

PREVENTION GENETICS
DISEASE PREVENTION THROUGH GENETIC TESTING

MAYO CLINIC

LINEAGEN

MOFFITT CANCER CENTER

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER

BAYER

Lilly

abbvie

TEVA

NATIONAL CANCER INSTITUTE

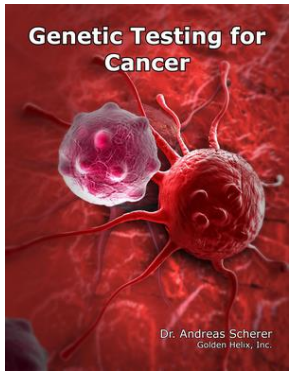
USDA
USDA Agricultural Research Service

FDA

Golden Helix – Who We Are

When you choose a Golden Helix solution, you get more than just software

- REPUTATION
- TRUST
- EXPERIENCE

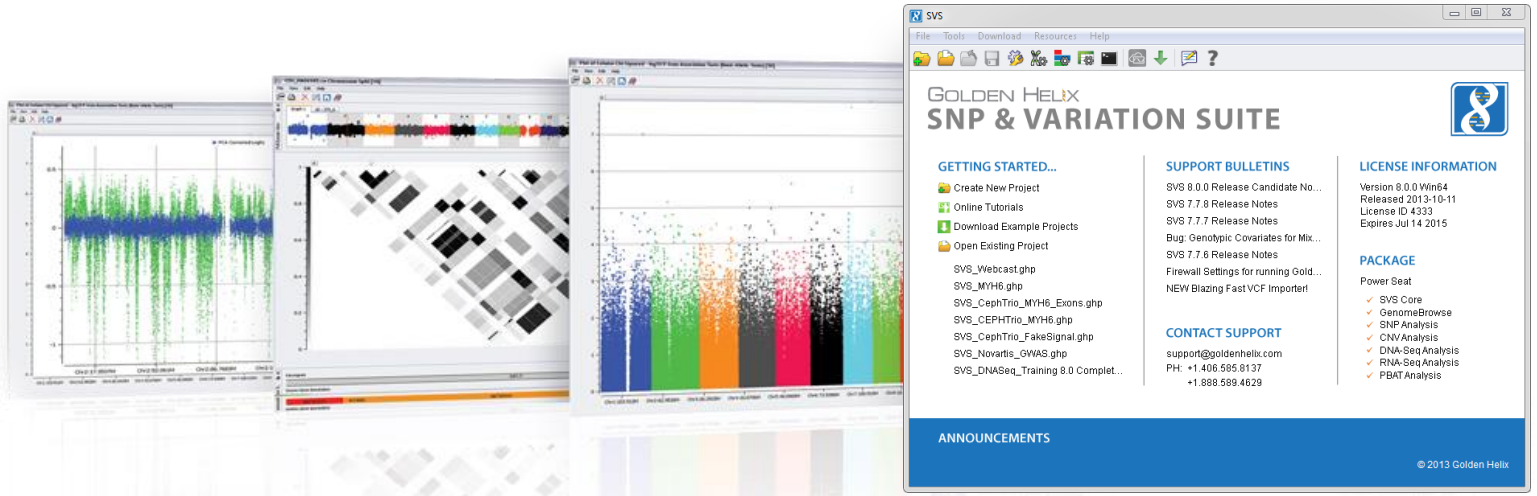


- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

SNP & Variation Suite (SVS)



Core Features

- Powerful Data Management
- Rich Visualizations (GenomeBrowse)
- Robust Statistics
- Flexible

Applications

- Genotype Analysis
- Agrigenomics Analysis
- DNA Sequence Analysis
- CNV Analysis

Agenda

1. GWAS Workflows in SVS
2. New and Enhanced Features
3. Questions

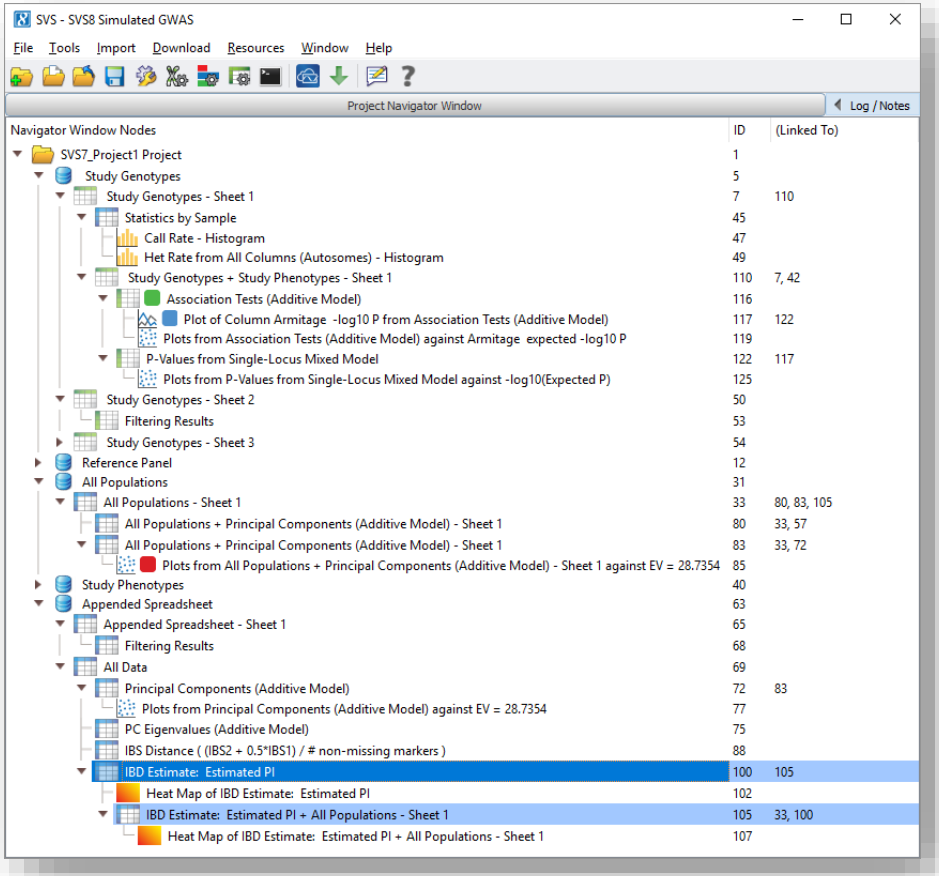
GWAS Workflow in SVS

Import
- SNP & NGS Data
- Phenotype Data
- Genotype Imputation

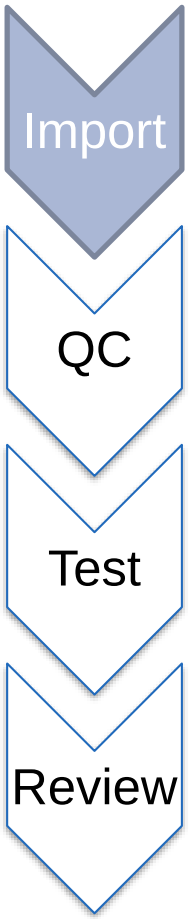
QC
- Sample QC
- Marker QC
- Population Structure

Test
- Association Test
- **Test Correction Techniques**
- Multi-Variable Analysis

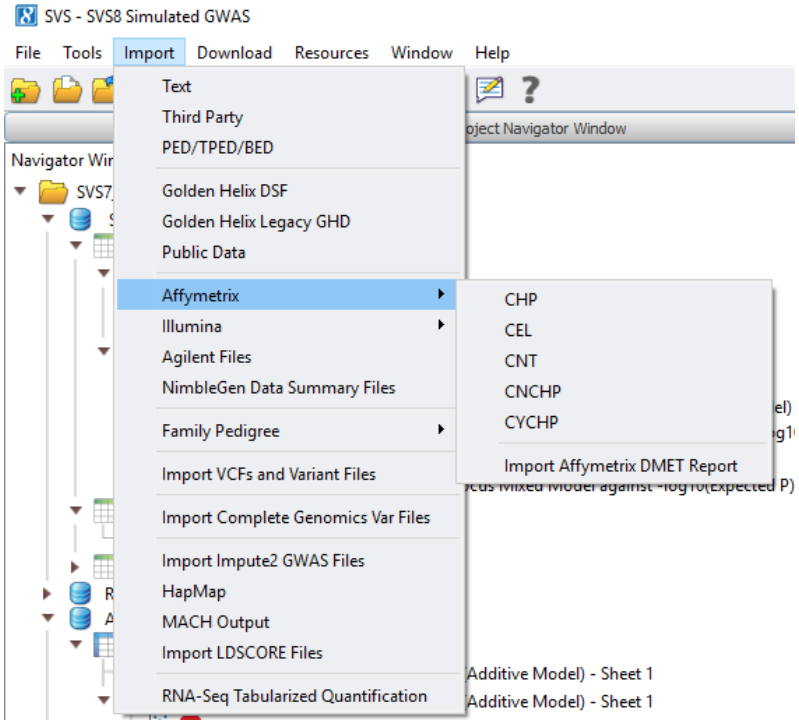
Review
- Visualization / Genomic Context
- Lambda / **LD Regression**
- Meta Analysis



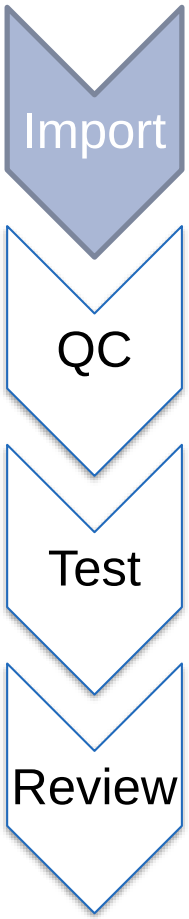
Import and Data Harmonizing in SVS

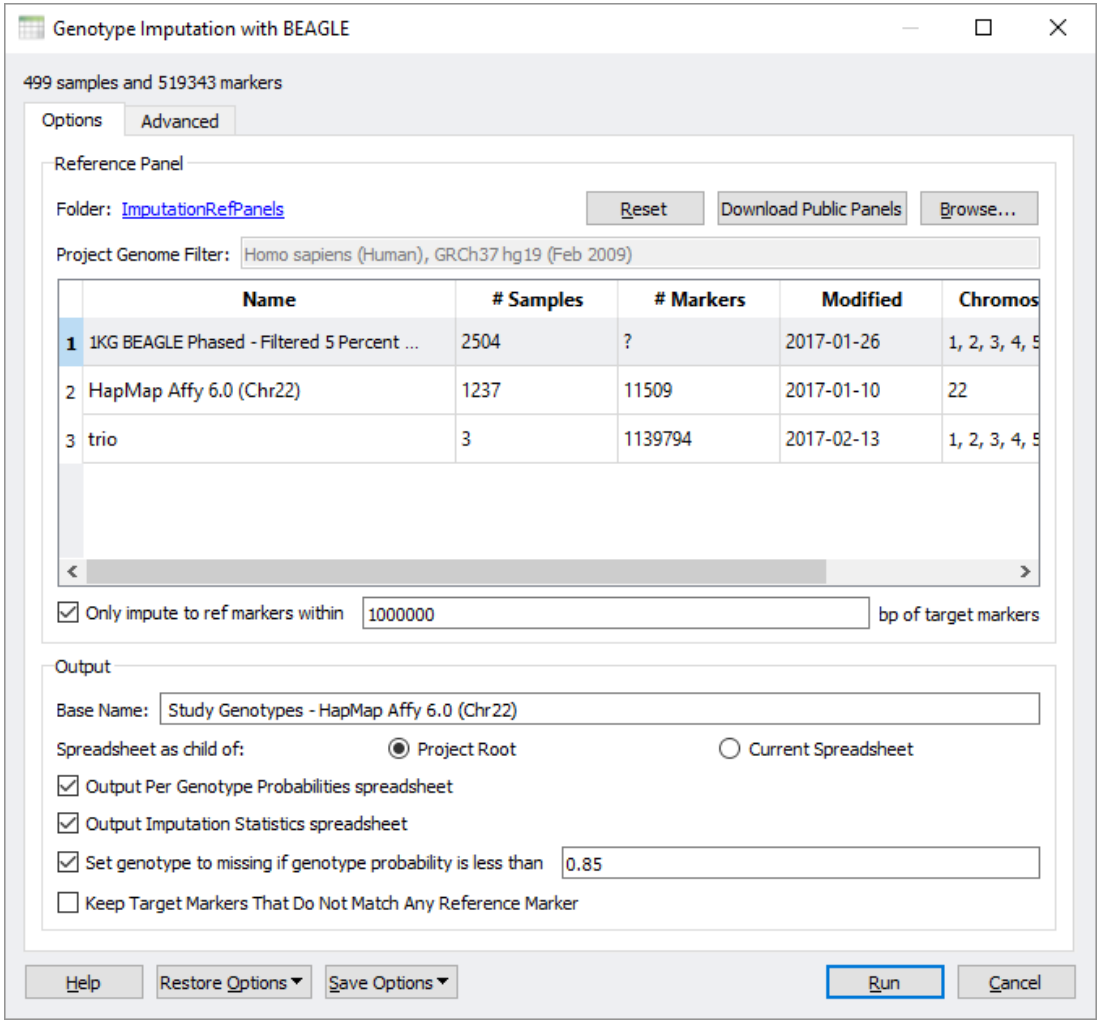


- **Prepare Analysis**
  - Genotype Data
  - Phenotype Data
  - Data Joins
  - Genomic Mapping
  - Remap using dbSNP
  - Genomic Annotations
- **Genotype Imputation**
  - Harmonize Multiple Platforms
  - Use Public Controls
  - Fill in Missing Genotypes
  - Increase Density

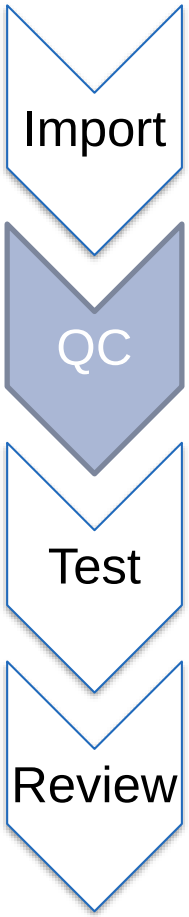


Genotype Imputation using BEAGLE in SVS

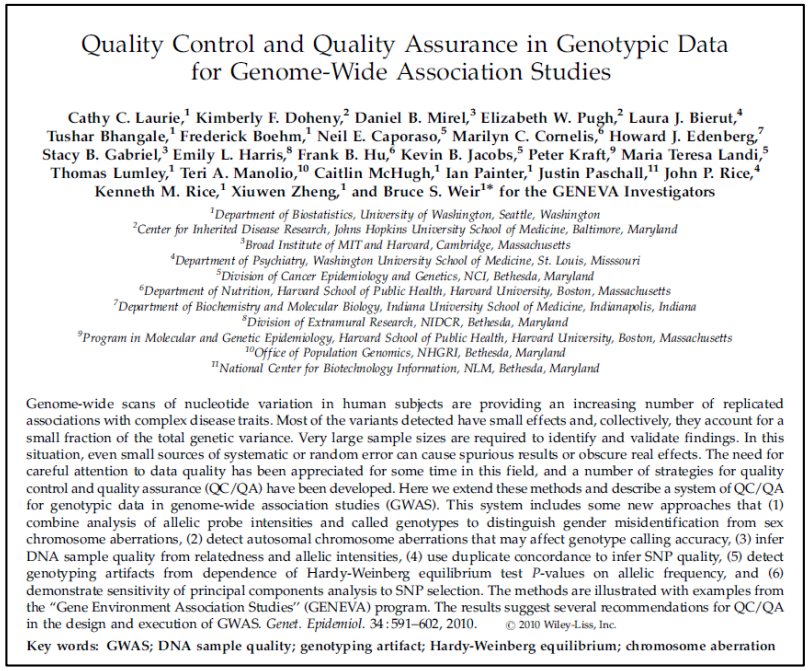




Quality Control & Quality Assurance



- **Sample QC:**
  - Call Rate / Het Rate
  - Gender Checks
  - Runs of Homozygosity
  - IBD Testing
  - Principle Component Analysis
  - Mendelian Errors (Pedigree)
- **Marker QC / Filtering**
  - Call Rate / HWE
  - Minor Allele Frequency
  - LD Pruning
  - Genomic Annotations
- **Further Recommendations**
  - QQ Plots after testing (covered later)



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

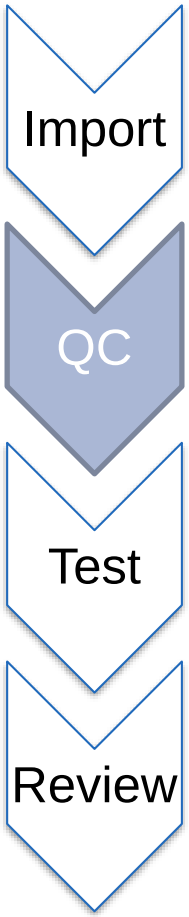
[1]*Department of Biostatistics, University of Washington, Seattle, Washington*
[2]*Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland*
[3]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts*
[4]*Department of Psychiatry, Washington University School of Medicine, St. Louis, Misssouri*
[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

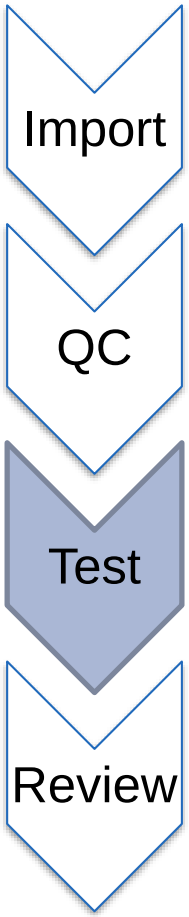
Genetic Epidemiology 34:591-602 (2010)

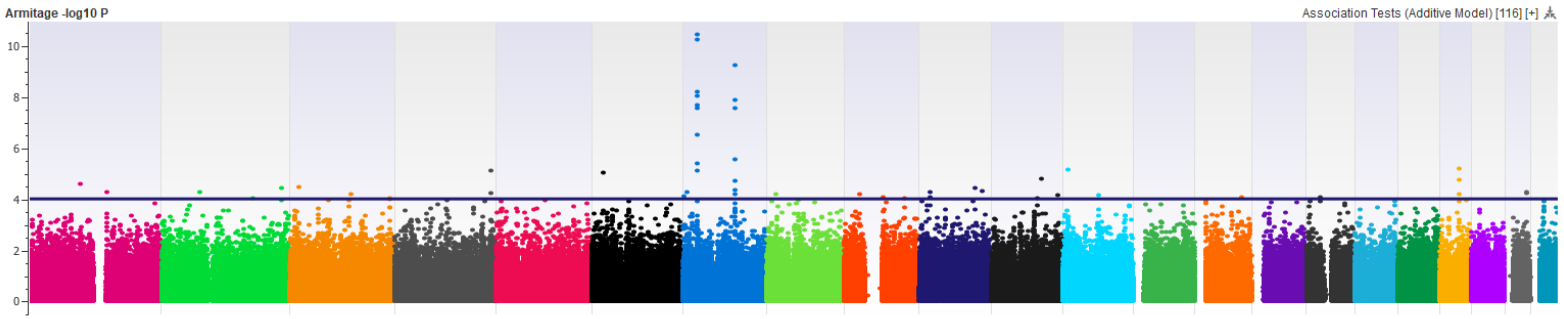
Quality Control & Quality Assurance



GOLDEN HELIX
SNP & VARIATION SUITE

Association Testing in SVS





- Genomic Model
  - Basic Allelic Tests
  - Genotypic Tests
  - **Additive Model**
  - Dominant Model
  - Recessive Model
- Multiple Test Correction
  - **Bonferroni Adjustment**
  - False Discovery Rate
  - Permutation Testing
- Test Statistic
  - Correlation/Trend Test
  - **Armitage Trend Test**
  - Exact Form of Armitage Test
  - (Pearson) Chi-Squared Test (+Yate's)
  - Fisher's Exact Test
  - Odds Ratio with Confidence Limits
  - Analysis of Deviance
  - F-Test
  - Logistic Regression
  - **Linear Regression**

Method for Specific Applications

Import

QC

Test

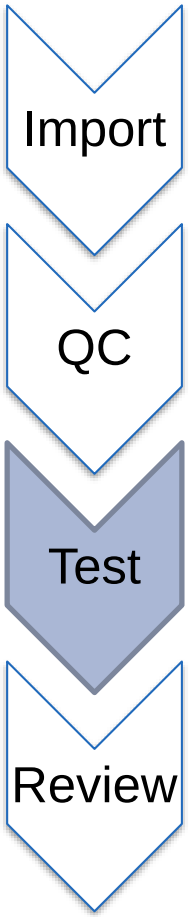
Review

- Haplotype Analysis
- PBAT Family Based Analysis
  - With sample level pedigree data
- Collapsing Methods
  - Complex trait analysis
  - Gene and region based tests
- **Agrigenomics**
  - Estimating Breeding Values
  - Genomic Prediction
  - Genetic Contribution of Traits
- Comparison of Multiple Traits
  - **Genetic Correlation**

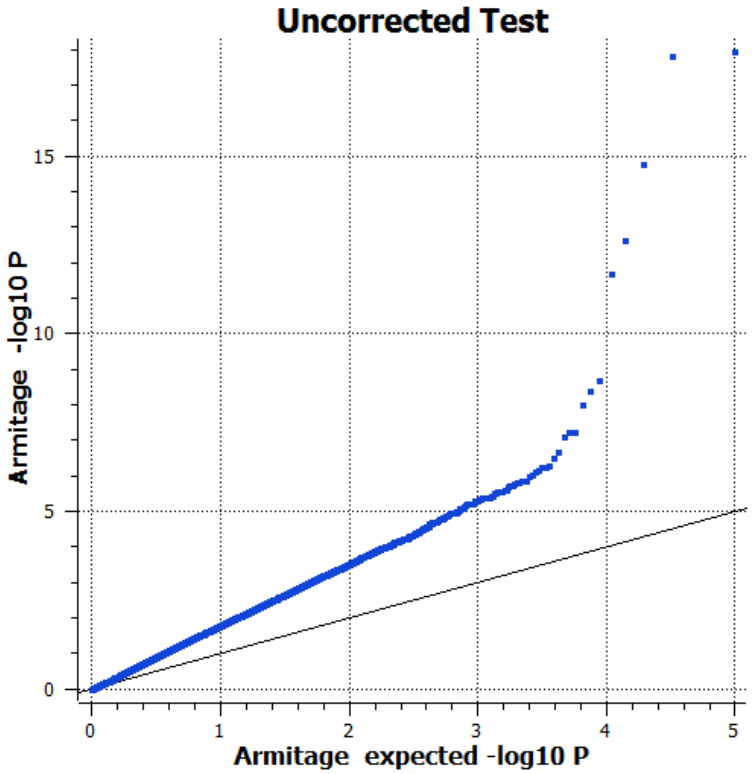





Test Correction Techniques

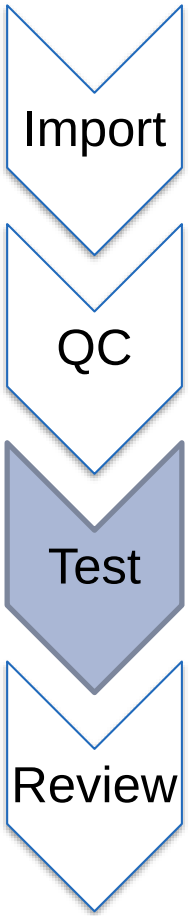


- The naïve approaches test a single marker with no correction
- Batch Effects, Population Structure and sharing of controls may violate assumptions of the naïve approaches and result in confounding of results.
- Stratification effects are more pronounced with larger sample sizes.
- Non-independence of samples is especially problematic in agrigenomic applications.



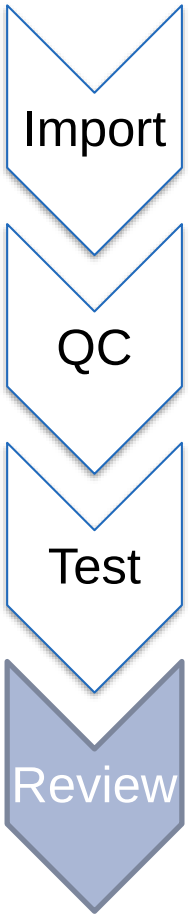
Example of uncorrected GWAS test
Lambda (λ) inflation factor of 2.48

Correcting for Population Stratification



- **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

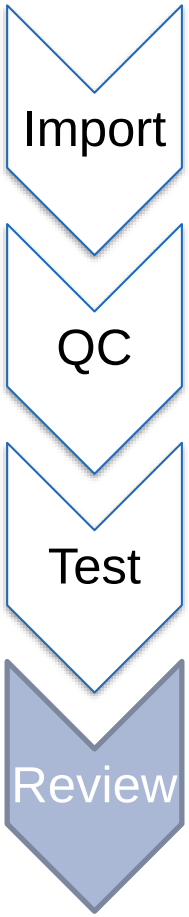
Review Test Results



- **Test Statistics Results**
  - Sort, Review Counts
  - Manhattan Plots
  - Annotate genes
  - Annotate phenotypes (PhoRank)
- **Test Validity:**
  - Lambda
  - Q-Q Plots
  - Meta-Analysis
- **Consider Other Statistics**
  - Correct for confounding trait (batch, population)
  - Permutation testing
  - LD Regression

```
Markers analyzed:
    519308 markers with two alleles
    (35 markers with one allele were found.)

Analysis parameters:
Genetic model/test: Additive model
Use missing values: No

Correct Input Data for Stratification using PCA: No

Test Statistic or Method:
* Armitage Trend Test

Multiple Testing Correction:
Bonferroni adjustment for 519308 markers: Yes
False Discovery Rate: Yes
Single Value Permutations: No
Full Scan Permutations: No

Genomic Control of Output Data for Stratification: Yes
Output data for P-P/Q-Q plots:  Yes
Output -log10 P:  Yes

Markerwise Genotype Statistics:
  Call Rate: No
  Number of Alleles: No
  Allele Frequencies: Yes
  Carrier Counts: No
  HWE P-Value: No
  Fisher's Exact Test for HWE P-Value: No
  Signed HWE R: No

  Also output -log10(Value): No
  Also output data for P-P/Q-Q plots: No

  Genotype Counts: Yes
  Allele Counts: Yes


Inflation Factor (Lambda) Found for Armitage : 2.66031
```

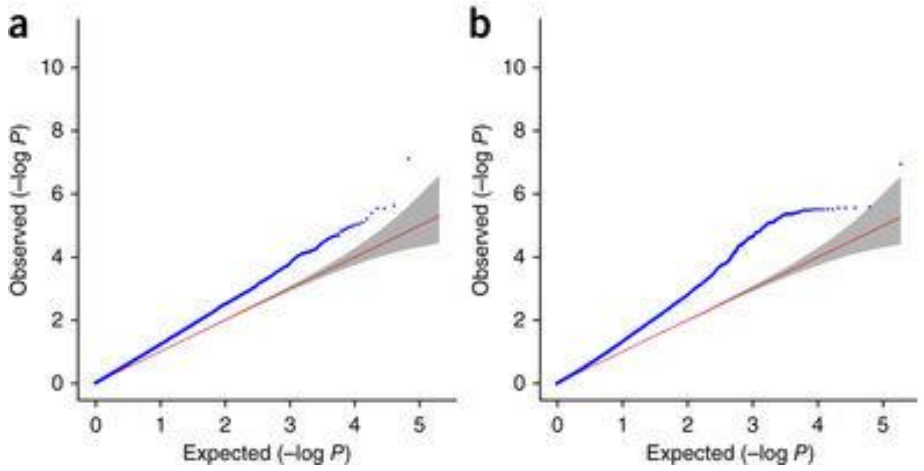
Review Test Results



GOLDEN HELIX
SNP & VARIATION SUITE

LD Score Regression

- **Precompute LD Scores for 1.2 million SNPs on two populations**
- **Join to your GWAS results**
  - Use RSID
- **Two modes:**
  - Compute Heritability estimate
  - Compute Genetic Correlation with additional traits



(a) Population stratification
(b) Polygenic genetic architecture

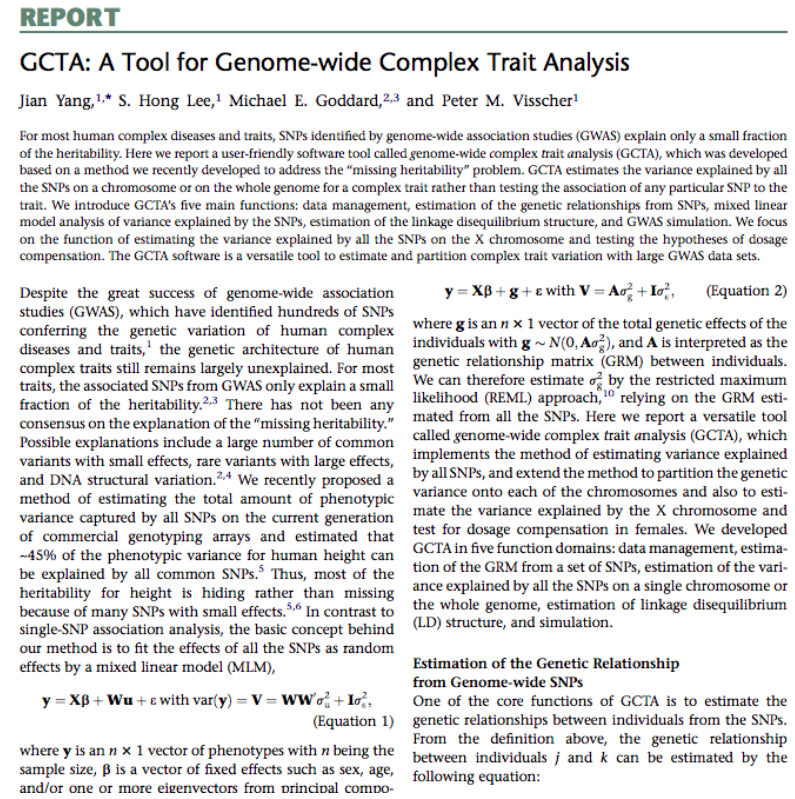
Bulik-Sullivan, et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nature Genetics, 2015.

GBLUP as Agrigenomic Workhorse

- **Computes a Genome Relationship Matrix (GRM) faster and more memory efficiently than IBD (which is N^2)**
- **Incorporates genomic relationship matrix (GRM) in mixed linear model framework to account for relatedness among samples**
- **Calculates allele substitution effect (ASE) for each SNP (directional)**
- **Computes estimated breeding values (GEBV) for samples**
- **Can predict phenotypes for samples with missing phenotype based on model trained on phenotyped samples**
- **Also calculates:**
  - Pseudo-heritability of trait
  - Genetic component of trait variance
  - Error component of trait variance

Enhanced GBLUP Capabilities for SVS

- GBLUP Enhancements adopted from GCTA paper
  - Alternative options for computing the genomic relationship matrix (GRM):
    - Compute different GRM for sex (X) and use in mixed model association test
  - The option to correct for gene by environment interactions based on an environment categorical variable
  - Inbreeding coefficient f now output from all algorithms
- New analysis to estimate genetic correlation of two traits:
  - Estimate the genetic variance of each trait and the genetic covariance between two traits that can be captured by all SNPs



REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

For most human complex diseases and traits, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability. Here we report a user-friendly software tool called genome-wide complex trait analysis (GCTA), which was developed based on a method we recently developed to address the "missing heritability" problem. GCTA estimates the variance explained by all the SNPs on a chromosome or on the whole genome for a complex trait rather than testing the association of any particular SNP to the trait. We introduce GCTA's five main functions: data management, estimation of the genetic relationships from SNPs, mixed linear model analysis of variance explained by the SNPs, estimation of the linkage disequilibrium structure, and GWAS simulation. We focus on the function of estimating the variance explained by all the SNPs on the X chromosome and testing the hypotheses of dosage compensation. The GCTA software is a versatile tool to estimate and partition complex trait variation with large GWAS data sets.

Despite the great success of genome-wide association studies (GWAS), which have identified hundreds of SNPs conferring the genetic variation of human complex diseases and traits,[1] the genetic architecture of human complex traits still remains largely unexplained. For most traits, the associated SNPs from GWAS only explain a small fraction of the heritability.[2,3] There has not been any consensus on the explanation of the "missing heritability." Possible explanations include a large number of common variants with small effects, rare variants with large effects, and DNA structural variation.[2,4] We recently proposed a method of estimating the total amount of phenotypic variance captured by all SNPs on the current generation of commercial genotyping arrays and estimated that ~45% of the phenotypic variance for human height can be explained by all common SNPs.[5] Thus, most of the heritability for height is hiding rather than missing because of many SNPs with small effects.[5,6] In contrast to single-SNP association analysis, the basic concept behind our method is to fit the effects of all the SNPs as random effects by a mixed linear model (MLM),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \text{ with } \mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbf{I}\sigma_\varepsilon^2, \quad \text{(Equation 1)}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypes with n being the sample size, $\boldsymbol{\beta}$ is a vector of fixed effects such as sex, age, and/or one or more eigenvectors from principal compo-

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \text{ with } \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2, \quad \text{(Equation 2)}$$

where $\mathbf{g}$ is an $n \times 1$ vector of the total genetic effects of the individuals with $\mathbf{g} \sim N(0, \mathbf{A}\sigma_g^2)$, and $\mathbf{A}$ is interpreted as the genetic relationship matrix (GRM) between individuals. We can therefore estimate σ_g^2 by the restricted maximum likelihood (REML) approach,[10] relying on the GRM estimated from all the SNPs. Here we report a versatile tool called genome-wide complex trait analysis (GCTA), which implements the method of estimating variance explained by all SNPs, and extend the method to partition the genetic variance onto each of the chromosomes and also to estimate the variance explained by the X chromosome and test for dosage compensation in females. We developed GCTA in five function domains: data management, estimation of the GRM from a set of SNPs, estimation of the variance explained by all the SNPs on a single chromosome or the whole genome, estimation of linkage disequilibrium (LD) structure, and simulation.

Estimation of the Genetic Relationship from Genome-wide SNPs

One of the core functions of GCTA is to estimate the genetic relationships between individuals from the SNPs. From the definition above, the genetic relationship between individuals j and k can be estimated by the following equation:

Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 88(1): 76-82.

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*. 2012;28(19):2540-2542.

New Methods

GOLDEN HELIX
SNP & VARIATION SUITE

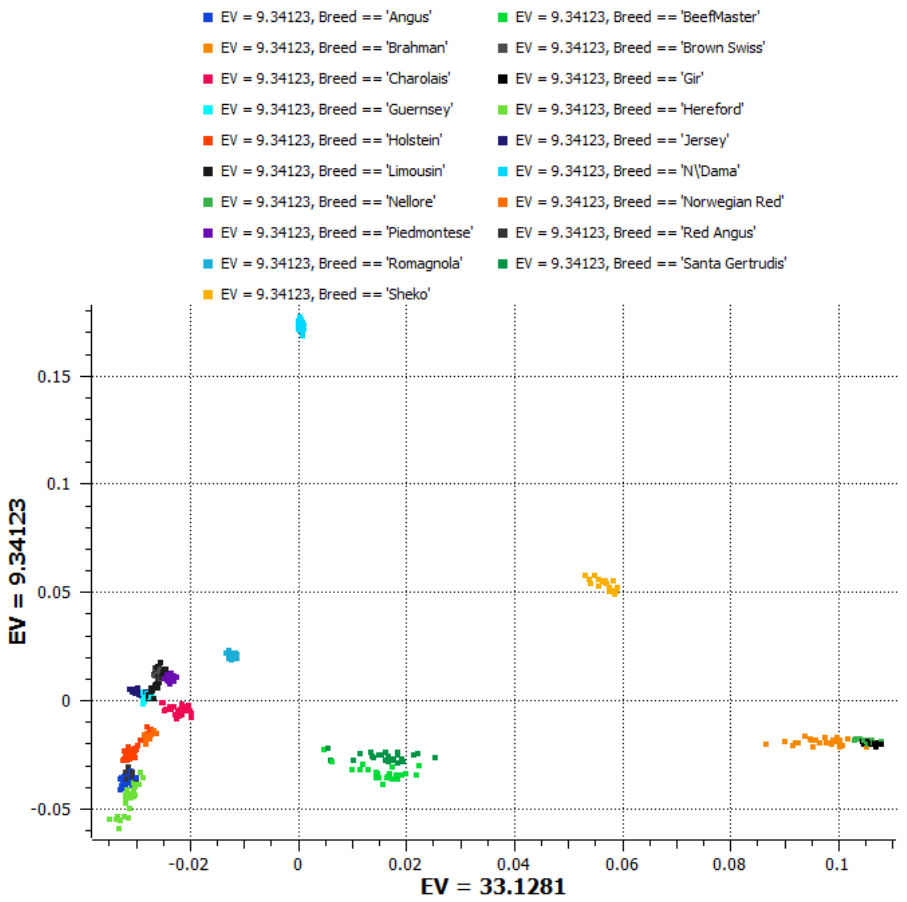
GWAS Workflow in SVS

Import
- SNP & NGS Data
- Phenotype Data
- Genotype Imputation

QC
- Sample QC
- Marker QC
- Population Structure

Test
- Association Test
- **Test Correction Techniques**
- Multi-Variable Analysis

Review
- Visualization / Genomic Context
- Lambda / **LD Regression**
- Meta Analysis



Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com