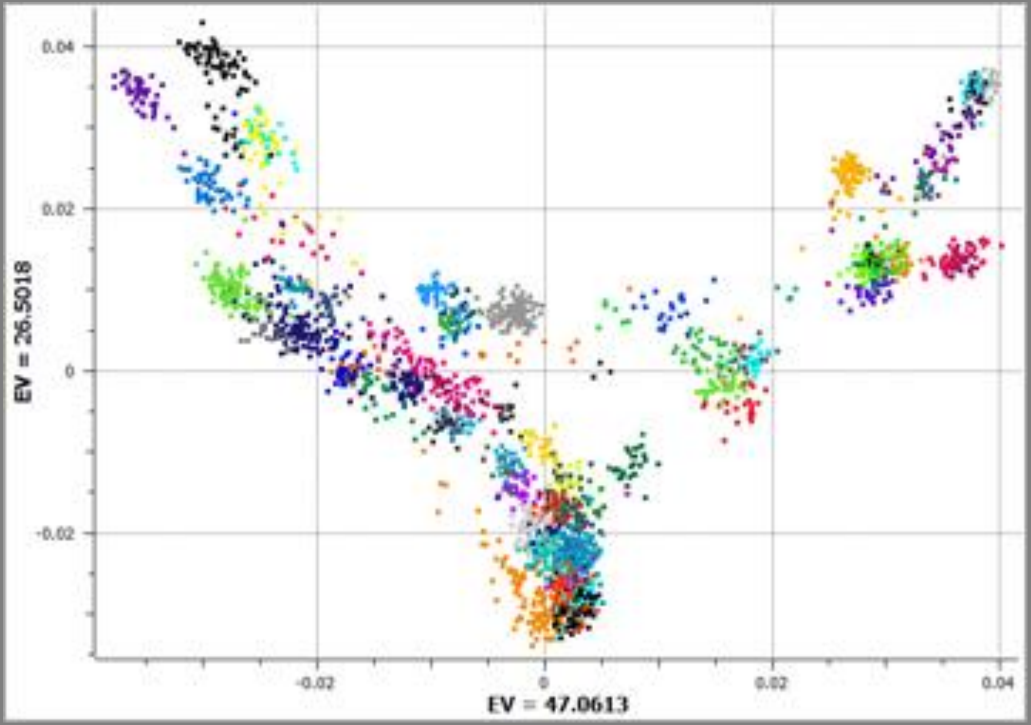


Mixed Models: How to Effectively Account for Inbreeding and Population Structure in GWAS

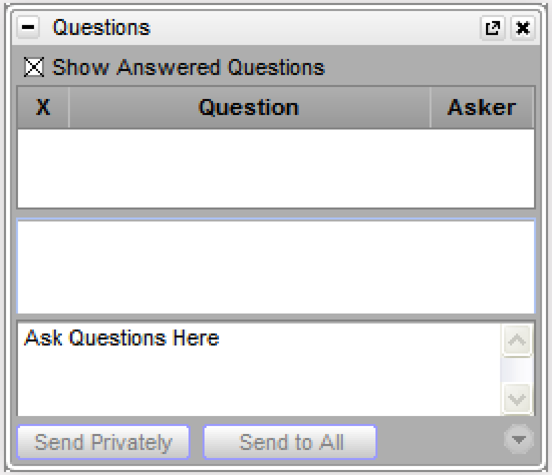
Greta Linse Peterson, Senior Statistician

June 5, 2013

Golden Helix
Accelerating the Quest for Significance™

Questions During the Presentation

Use the Questions pane in your GoToWebinar window



Agenda

1 Background of GWAS Approaches

2 Review of Select Mixed Model Methods

3 Mixed Models in SVS

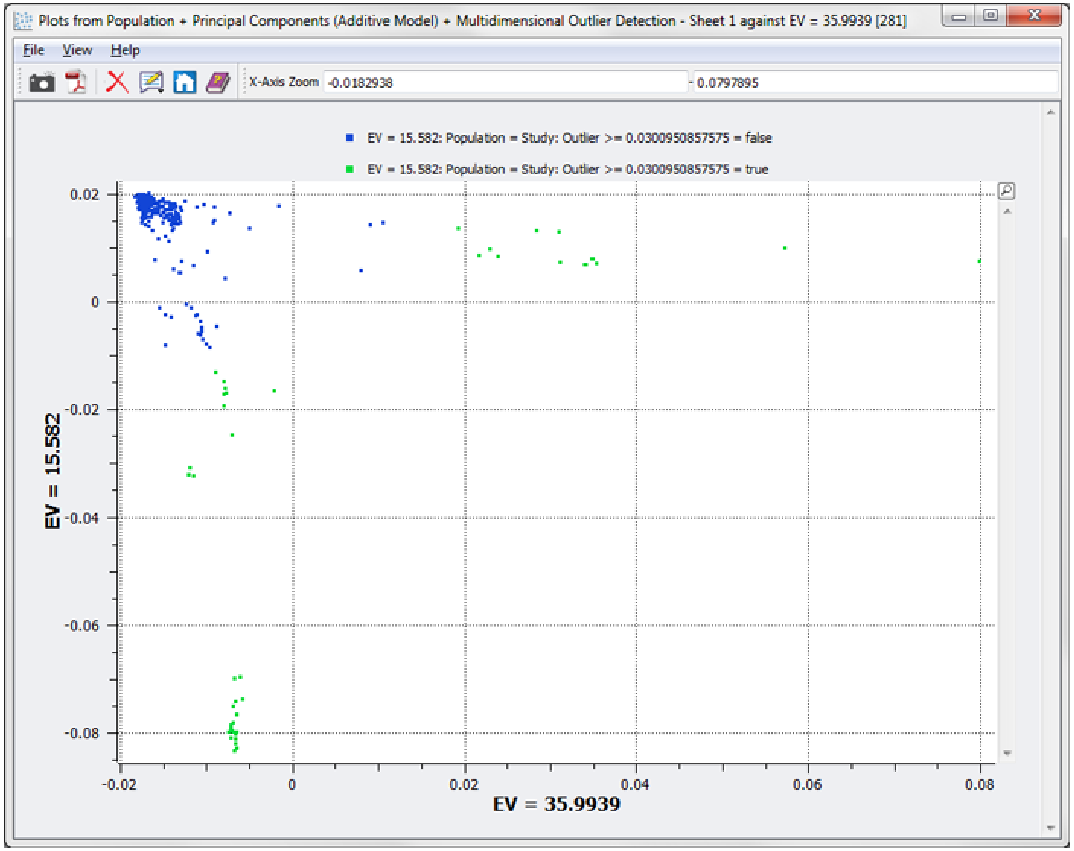
4 Demo

5 Compare Results

[Poll: What category of species are you studying?]

A brief background of GWAS

- First the naïve approaches: Correlation/Trend Test, Linear/Logistic Regression
- Batch Effects, Population Structure and sharing of controls violated assumptions of the naïve approaches.



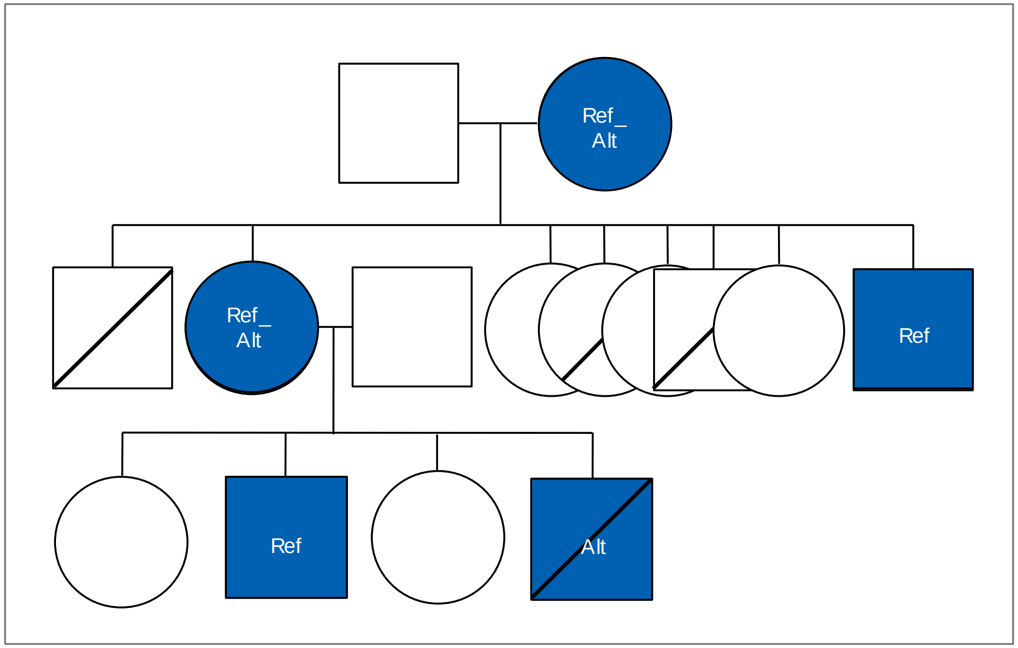
Plots from Population + Principal Components (Additive Model) + Multidimensional Outlier Detection - Sheet 1 against EV = 35.9939 [281]

X-Axis Zoom -0.0182938 - 0.0797895

File
View
Help
■ EV = 15.582: Population = Study: Outlier >= 0.0300950857575 = false
■ EV = 15.582: Population = Study: Outlier >= 0.0300950857575 = true
0.02
0
-0.02
-0.04
-0.06
-0.08
EV = 15.582
-0.02
0
0.02
0.04
0.06
0.08
EV = 35.9939

Goal of better GWAS approaches

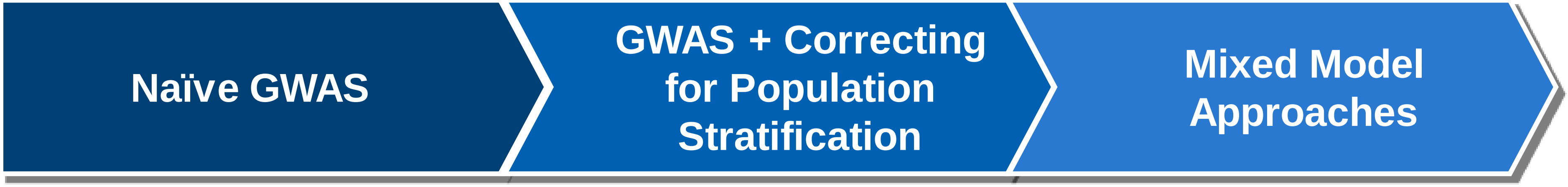
- Minimize false positives, obtain cleaner results, don’t over correct the data to miss out on interesting results
- Handle population, family-based or mixed study designs.



Essential Definitions

- **Mixed Linear Model:**
  - $Y = X\beta + Z\mu + \epsilon$ where $\mu \sim N(0, \sigma_G^2 K^*)$, $\epsilon \sim N(0, \sigma_e^2 I)$ and $Cov(\mu, \epsilon) = 0$
- **Fixed Factors:**
  - Sex, age, known loci
- **Random Effects:**
  - Family or Population Structure, batch effects
- **Kinship Matrix:**
  - Any N x N matrix that describes the pairwise relationships between N samples
- **Null Hypothesis (generally):** $\sigma_G^2 = 0$

Approximate Timeline



Naïve GWAS

GWAS + Correcting for Population Stratification

Mixed Model Approaches

- Corr/Trend Test
- Regression Analysis

- Genomic Control
- Structured Association (STRUCTURE)
- PCA Correction (Eigenstrat Price 2006)

- EMMA (Kang 2008)
- BLUP/GBLUP Approaches for GWAS (Zhang 2008)
- EMMAX (Kang 2010)
- MLMM (Segura 2012)

Methods for MLMs in GWAS

Agenda

1 Background of GWAS Approaches

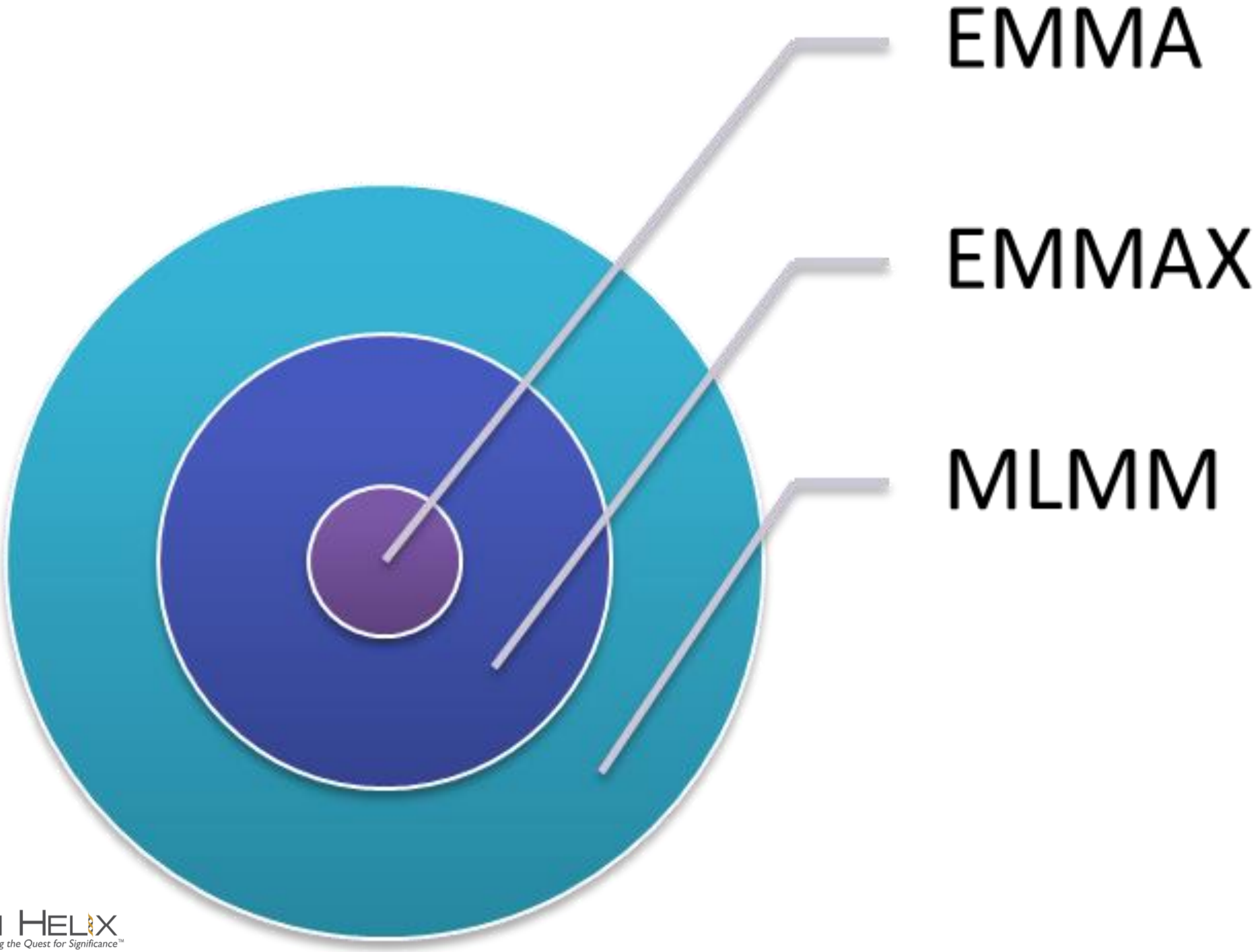
2 **Review of Select Mixed Model Methods**

3 Mixed Models in SVS

4 Demo

5 Compare Results

EMMA/EMMAX/MLMM Relationship



Methods Overview

- **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 relationship between samples using a kinship matrix
  - Approximates the variance components and uses the same variance for all probes
  - Tests a single loci at a time
- **MLMM**
  - Adjusts for the relationship between 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 relationship between samples using a kinship matrix
  - Computes allele substitution effects to determine best genomic predictors of the phenotpye

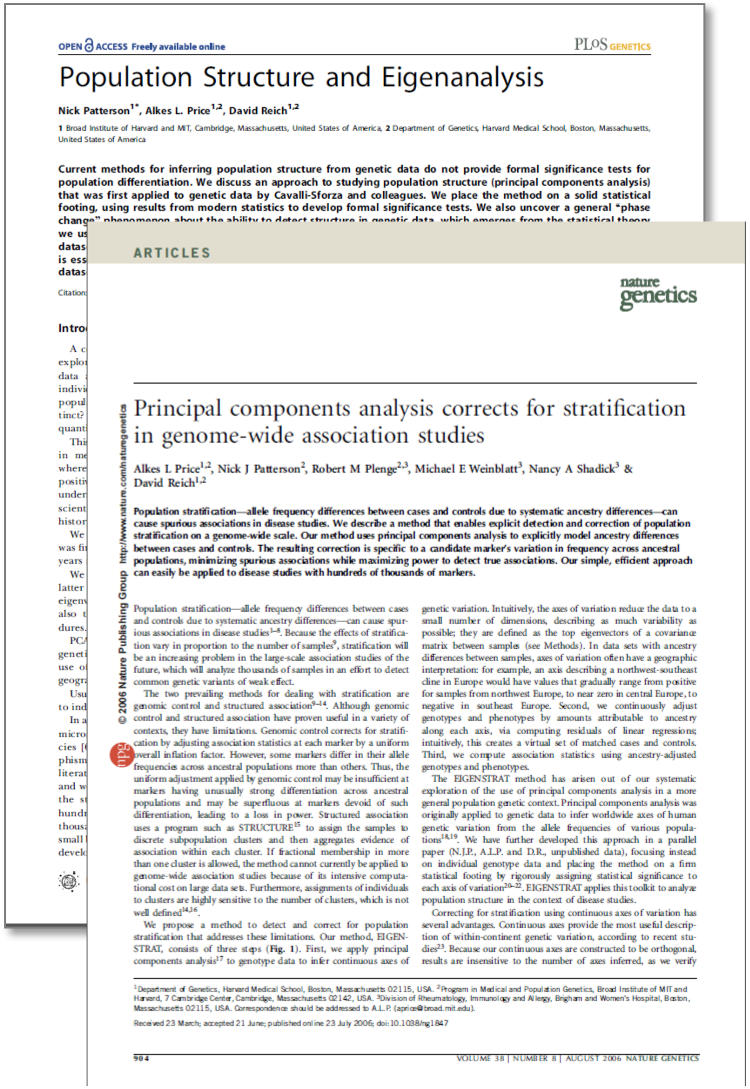
Method Comparison

	Population Structure as Fixed Effect	Multiple Loci	EMMA used	Uses Kinship as Random Effect	Output Random Effect Component	Compute Allele Substitution Effects	Compute P-Value
Regression with PCA	Yes	No	No	No	No	No	Yes
EMMAX	Yes	No	Yes	Yes	No	No	Yes
MLMM	Yes	Yes	Yes	Yes	No	No	Yes
GBLUP	No	No	Yes*	Yes	Yes	Yes	No

* Uses EMMA for REML estimates

Regression with PCA method overview

- First compute the principal components
  - Assumes the first few components are associated with the largest batch effects including population structure, plate effects, etc.
- Decide how many components to correct for
- Either run regression on PCA corrected data or on genotype data including top principal components as fixed factors



OPEN ACCESS Freely available online PLoS GENETICS

Population Structure and Eigenanalysis

Nick Patterson[1*], Alkes L. Price[1,2], David Reich[1,2]

1 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, 2 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

Current methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation. We discuss an approach to studying population structure (principal components analysis) that was first applied to genetic data by Cavalli-Sforza and colleagues. We place the method on a solid statistical footing, using results from modern statistics to develop fomal significance tests. We also uncover a general "phase change" phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory

ARTICLES

nature genetics

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

© 2006 Nature Publishing Group http://www.nature.com/naturegenetics

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies[1–8]. Because the effects of stratification vary in proportion to the number of samples[9], stratification will be an increasing problem in the large-scale association studies of the future, which will analyze thousands of samples in an effort to detect common genetic variants of weak effect.

The two prevailing methods for dealing with stratification are genomic control and structured association[9–14]. Although genomic control and structured association have proven useful in a variety of contexts, they have limitations. Genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor. However, some markers differ in their allele frequencies across ancestral populations more than others. Thus, the uniform adjustment applied by genomic control may be insufficient at markers having unusually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss in power. Structured association uses a program such as STRUCTURE[15] to assign the samples to discrete subpopulation clusters and then aggregates evidence of association within each cluster. If fractional membership in more than one cluster is allowed, the method cannot currently be applied to genome-wide association studies because of its intensive computational cost on large data sets. Furthermore, assignments of individuals to clusters are highly sensitive to the number of clusters, which is not well defined[14,16].

We propose a method to detect and correct for population stratification that addresses these limitations. Our method, EIGENSTRAT, consists of three steps (**Fig. 1**). First, we apply principal components analysis[17] to genotype data to infer continuous axes of genetic variation. Intuitively, the axes of variation reduce the data to a small number of dimensions, describing as much variability as possible; they are defined as the top eigenvectors of a covariance matrix between samples (see Methods). In data sets with ancestry differences between samples, axes of variation often have a geographic interpretation: for example, an axis describing a northwest-southeast cline in Europe would have values that gradually range from positive for samples from northwest Europe, to near zero in central Europe, to negative in southeast Europe. Second, we continuously adjust genotypes and phenotypes by amounts attributable to ancestry along each axis, via computing residuals of linear regressions; intuitively, this creates a virtual set of matched cases and controls. Third, we compute association statistics using ancestry-adjusted genotypes and phenotypes.

The EIGENSTRAT method has arisen out of our systematic exploration of the use of principal components analysis in a more general population genetic context. Principal components analysis was originally applied to genetic data to infer worldwide axes of human genetic variation from the allele frequencies of various populations[18,19]. We have further developed this approach in a parallel paper (N.J.P., A.L.P. and D.R., unpublished data), focusing instead on individual genotype data and placing the method on a firm statistical footing by rigorously assigning statistical significance to each axis of variation[20–22]. EIGENSTRAT applies this toolkit to analyze population structure in the context of disease studies.

Correcting for stratification using continuous axes of variation has several advantages. Continuous axes provide the most useful description of within-continent genetic variation, according to recent studies[23]. Because our continuous axes are constructed to be orthogonal, results are insensitive to the number of axes inferred, as we verify

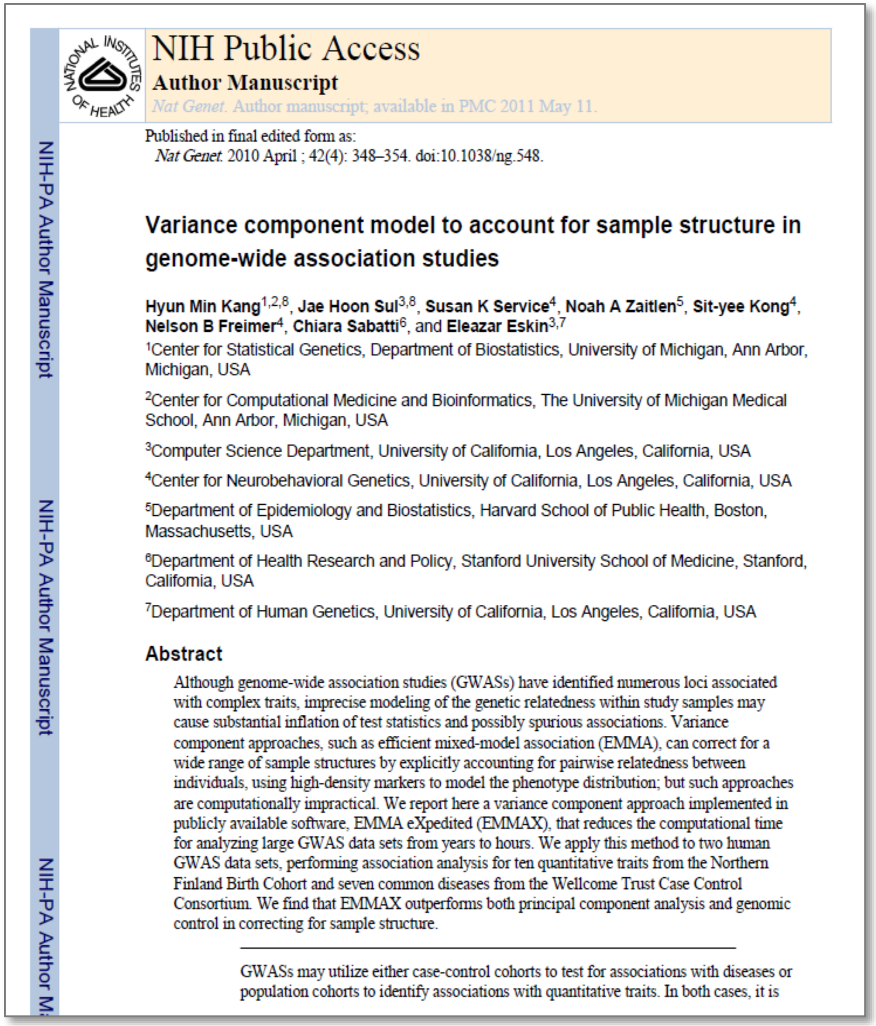
[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. Correspondence should be addressed to A.L.P. (aprice@broad.mit.edu).

Received 23 March; accepted 21 June; published online 23 July 2006; doi:10.1038/ng1847

904 VOLUME 38 | NUMBER 8 | AUGUST 2006 NATURE GENETICS

EMMAX method overview

- Published in 2010 by the authors of EMMA
- Assumes a complex disease and that all SNP loci have a small effect on the phenotypic trait by themselves
- Instead of re-computing the variance components for every SNP (under the Alternative Hypothesis) computes it once under the Null Hypothesis
- Null Hypothesis: $\sigma_G^2 = 0$;



Published in final edited form as:
Nat Genet. 2010 April ; 42(4): 348–354. doi:10.1038/ng.548.

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], **Jae Hoon Sul**[3,8], **Susan K Service**[4], **Noah A Zaitlen**[5], **Sit-yee Kong**[4], **Nelson B Freimer**[4], **Chiara Sabatti**[6], and **Eleazar Eskin**[3,7]

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Center for Computational Medicine and Bioinformatics, The University of Michigan Medical School, Ann Arbor, Michigan, USA

[3]Computer Science Department, University of California, Los Angeles, California, USA

[4]Center for Neurobehavioral Genetics, University of California, Los Angeles, California, USA

[5]Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

[6]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA

[7]Department of Human Genetics, University of California, Los Angeles, California, USA

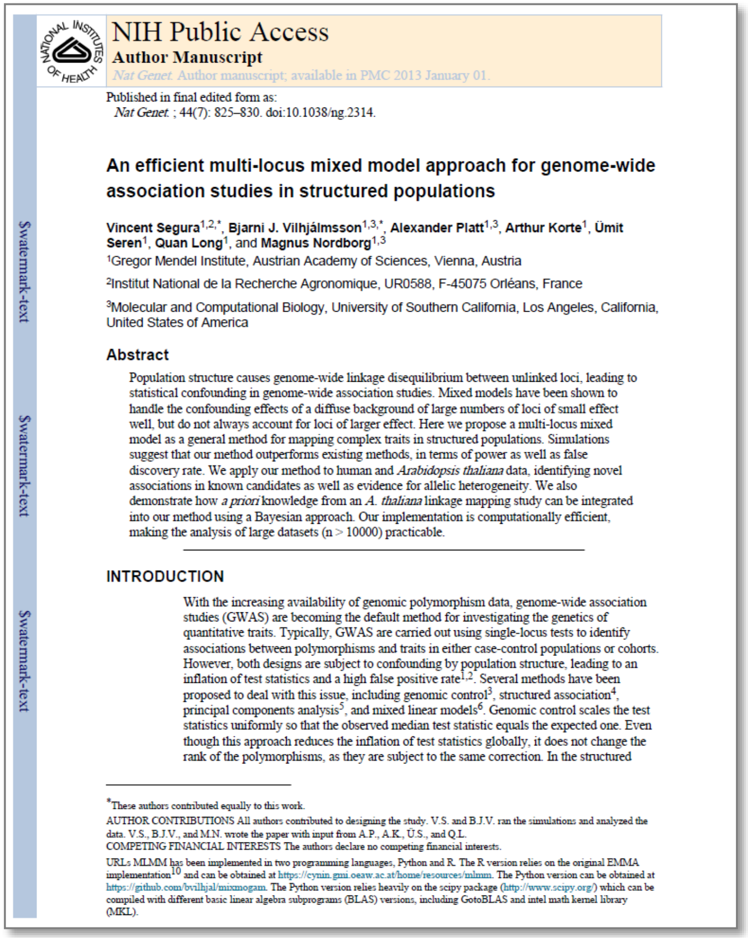
Abstract

Although genome-wide association studies (GWASs) have identified numerous loci associated with complex traits, imprecise modeling of the genetic relatedness within study samples may cause substantial inflation of test statistics and possibly spurious associations. Variance component approaches, such as efficient mixed-model association (EMMA), can correct for a wide range of sample structures by explicitly accounting for pairwise relatedness between individuals, using high-density markers to model the phenotype distribution; but such approaches are computationally impractical. We report here a variance component approach implemented in publicly available software, EMMA eXpedited (EMMAX), that reduces the computational time for analyzing large GWAS data sets from years to hours. We apply this method to two human GWAS data sets, performing association analysis for ten quantitative traits from the Northern Finland Birth Cohort and seven common diseases from the Wellcome Trust Case Control Consortium. We find that EMMAX outperforms both principal component analysis and genomic control in correcting for sample structure.

GWASs may utilize either case-control cohorts to test for associations with diseases or population cohorts to identify associations with quantitative traits. In both cases, it is

MLMM method overview

- “Multiple-Loci Mixed Models”; stepwise EMMAX
- Assumes complex diseases where multiple loci are associated with the phenotype
- Cofactors are selected in a stepwise fashion by choosing the probe with the smallest p-value
- Since EMMAX is used, genetic and error are computed once for each step.
- Genetic and error variances are then re-estimated at for every step



Published in final edited form as:
Nat Genet. ; 44(7): 825–830. doi:10.1038/ng.2314.

An efficient multi-locus mixed model approach for genome-wide association studies in structured populations

Vincent Segura[1,2,*], **Bjarni J. Vilhjálmsson**[1,3,*], **Alexander Platt**[1,3], **Arthur Korte**[1], **Ümit Seren**[1], **Quan Long**[1], and **Magnus Nordborg**[1,3]

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria

[2]Institut National de la Recherche Agronomique, UR0588, F-45075 Orléans, France

[3]Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America

Abstract

Population structure causes genome-wide linkage disequilibrium between unlinked loci, leading to statistical confounding in genome-wide association studies. Mixed models have been shown to handle the confounding effects of a diffuse background of large numbers of loci of small effect well, but do not always account for loci of larger effect. Here we propose a multi-locus mixed model as a general method for mapping complex traits in structured populations. Simulations suggest that our method outperforms existing methods, in terms of power as well as false discovery rate. We apply our method to human and *Arabidopsis thaliana* data, identifying novel associations in known candidates as well as evidence for allelic heterogeneity. We also demonstrate how *a priori* knowledge from an *A. thaliana* linkage mapping study can be integrated into our method using a Bayesian approach. Our implementation is computationally efficient, making the analysis of large datasets (n > 10000) practicable.

INTRODUCTION

With the increasing availability of genomic polymorphism data, genome-wide association studies (GWAS) are becoming the default method for investigating the genetics of quantitative traits. Typically, GWAS are carried out using single-locus tests to identify associations between polymorphisms and traits in either case-control populations or cohorts. However, both designs are subject to confounding by population structure, leading to an inflation of test statistics and a high false positive rate[1,2]. Several methods have been proposed to deal with this issue, including genomic control[3], structured association[4], principal components analysis[5], and mixed linear models[6]. Genomic control scales the test statistics uniformly so that the observed median test statistic equals the expected one. Even though this approach reduces the inflation of test statistics globally, it does not change the rank of the polymorphisms, as they are subject to the same correction. In the structured

[*]These authors contributed equally to this work.

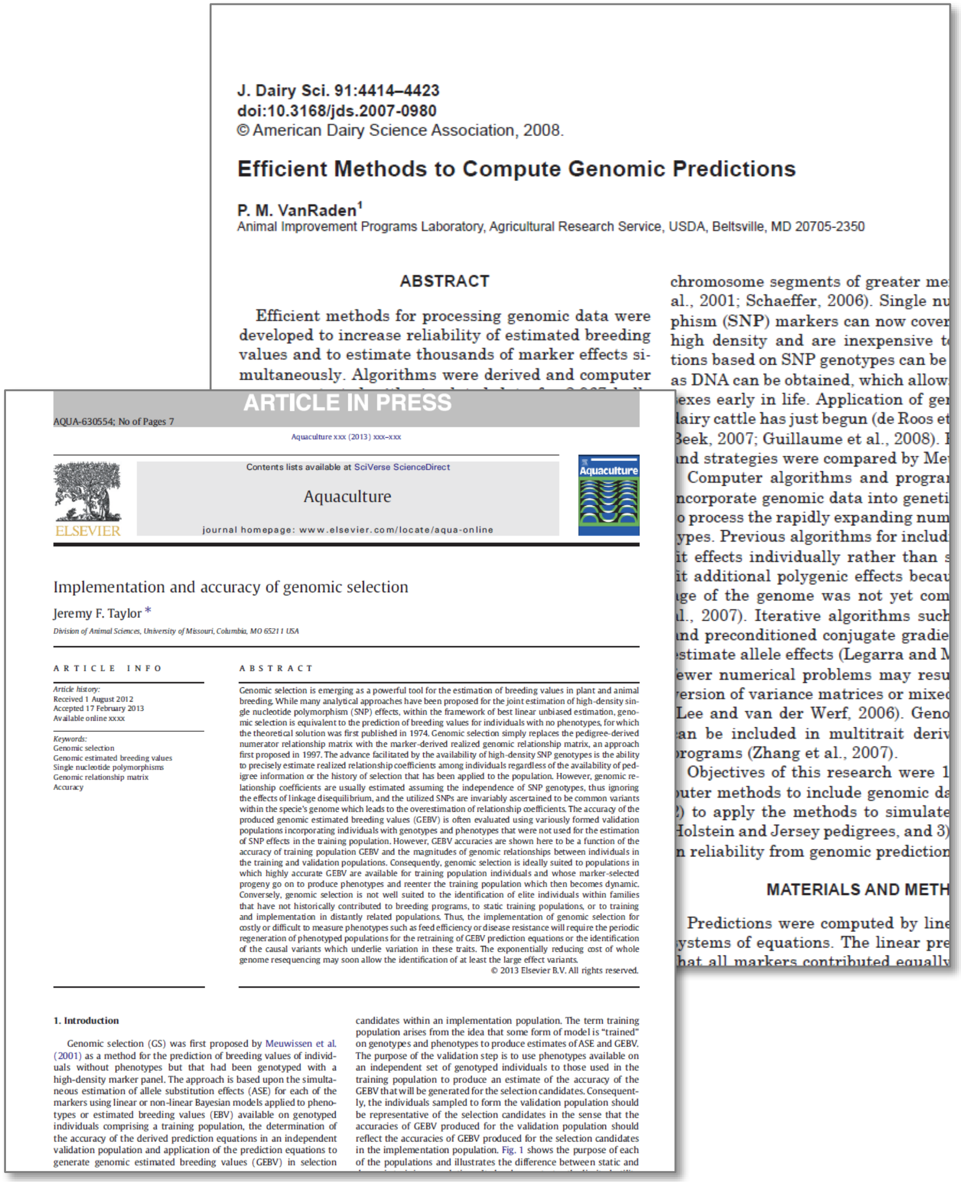
AUTHOR CONTRIBUTIONS All authors contributed to designing the study. V.S. and B.J.V. ran the simulations and analyzed the data. V.S., B.J.V., and M.N. wrote the paper with input from A.P., A.K., Ü.S., and Q.L.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

URLs MLMM has been implemented in two programming languages, Python and R. The R version relies on the original EMMA implementation[10] and can be obtained at https://cynin.gmi.oeaw.ac.at/home/resources/mlmm. The Python version can be obtained at https://github.com/bvilhjal/mixmogam. The Python version relies heavily on the scipy package (http://www.scipy.org/) which can be compiled with different basic linear algebra subprograms (BLAS) versions, including GotoBLAS and intel math kernel library (MKL).

GBLUP method overview

- Best Linear Unbiased Predictor (BLUP) provides residual errors
  - Residual Breeding Values for plant/animal studies
- Estimates of allele substitution effects
- Pseudo-heritability estimate can be used to compare the method with other methods
- Uses a genomic relationship matrix which computes faster than IBS



J. Dairy Sci. 91:4414–4423
doi:10.3168/jds.2007-0980
© American Dairy Science Association, 2008.

Efficient Methods to Compute Genomic Predictions

P. M. VanRaden[1]
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

Efficient methods for processing genomic data were developed to increase reliability of estimated breeding values and to estimate thousands of marker effects simultaneously. Algorithms were derived and computer

chromosome segments of greater me
al., 2001; Schaeffer, 2006). Single nu
phism (SNP) markers can now cover
high density and are inexpensive t
tions based on SNP genotypes can be
as DNA can be obtained, which allow
exes early in life. Application of ge
lairy cattle has just begun (de Roos et
Beek, 2007; Guillaume et al., 2008). I
nd strategies were compared by Me
Computer algorithms and progra
ncorporate genomic data into geneti
o process the rapidly expanding num
ypes. Previous algorithms for includ
it effects individually rather than s
it additional polygenic effects becau
ge of the genome was not yet com
l., 2007). Iterative algorithms such
nd preconditioned conjugate gradie
stimate allele effects (Legarra and M
ewer numerical problems may resu
ersion of variance matrices or mixed
Lee and van der Werf, 2006). Geno
an be included in multitrait deriv
rograms (Zhang et al., 2007).
Objectives of this research were 1
uter methods to include genomic da
) to apply the methods to simulate
Holstein and Jersey pedigrees, and 3)
n reliability from genomic prediction

MATERIALS AND METH

Predictions were computed by line
ystems of equations. The linear pre
hat all markers contributed equally

AQUA-630554; No of Pages 7

ARTICLE IN PRESS

Aquaculture xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Aquaculture

journal homepage: www.elsevier.com/locate/aqua-online

Implementation and accuracy of genomic selection

Jeremy F. Taylor *
Division of Animal Sciences, University of Missouri, Columbia, MO 65211 USA

ARTICLE INFO

Article history:
Received 1 August 2012
Accepted 17 February 2013
Available online xxxx

Keywords:
Genomic selection
Genomic estimated breeding values
Single nucleotide polymorphisms
Genomic relationship matrix
Accuracy

ABSTRACT

Genomic selection is emerging as a powerful tool for the estimation of breeding values in plant and animal breeding. While many analytical approaches have been proposed for the joint estimation of high-density single nucleotide polymorphism (SNP) effects, within the framework of best linear unbiased estimation, genomic selection is equivalent to the prediction of breeding values for individuals with no phenotypes, for which the theoretical solution was first published in 1974. Genomic selection simply replaces the pedigree-derived numerator relationship matrix with the marker-derived realized genomic relationship matrix, an approach first proposed in 1997. The advance facilitated by the availability of high-density SNP genotypes is the ability to precisely estimate realized relationship coefficients among individuals regardless of the availability of pedigree information or the history of selection that has been applied to the population. However, genomic relationship coefficients are usually estimated assuming the independence of SNP genotypes, thus ignoring the effects of linkage disequilibrium, and the utilized SNPs are invariably ascertained to be common variants within the specie's genome which leads to the overestimation of relationship coefficients. The accuracy of the produced genomic estimated breeding values (GEBV) is often evaluated using variously formed validation populations incorporating individuals with genotypes and phenotypes that were not used for the estimation of SNP effects in the training population. However, GEBV accuracies are shown here to be a function of the accuracy of training population GEBV and the magnitudes of genomic relationships between individuals in the training and validation populations. Consequently, genomic selection is ideally suited to populations in which highly accurate GEBV are available for training population individuals and whose marker-selected progeny go on to produce phenotypes and reenter the training population which then becomes dynamic. Conversely, genomic selection is not well suited to the identification of elite individuals within families that have not historically contributed to breeding programs, to static training populations, or to training and implementation in distantly related populations. Thus, the implementation of genomic selection for costly or difficult to measure phenotypes such as feed efficiency or disease resistance will require the periodic regeneration of phenotyped populations for the retraining of GEBV prediction equations or the identification of the causal variants which underlie variation in these traits. The exponentially reducing cost of whole genome resequencing may soon allow the identification of at least the large effect variants.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Genomic selection (GS) was first proposed by Meuwissen et al. (2001) as a method for the prediction of breeding values of individuals without phenotypes but that had been genotyped with a high-density marker panel. The approach is based upon the simultaneous estimation of allele substitution effects (ASE) for each of the markers using linear or non-linear Bayesian models applied to phenotypes or estimated breeding values (EBV) available on genotyped individuals comprising a training population, the determination of the accuracy of the derived prediction equations in an independent validation population and application of the prediction equations to generate genomic estimated breeding values (GEBV) in selection candidates within an implementation population. The term training population arises from the idea that some form of model is "trained" on genotypes and phenotypes to produce estimates of ASE and GEBV. The purpose of the validation step is to use phenotypes available on an independent set of genotyped individuals to those used in the training population to produce an estimate of the accuracy of the GEBV that will be generated for the selection candidates. Consequently, the individuals sampled to form the validation population should be representative of the selection candidates in the sense that the accuracies of GEBV produced for the validation population should reflect the accuracies of GEBV produced for the selection candidates in the implementation population. Fig. 1 shows the purpose of each of the populations and illustrates the difference between static and

When are Mixed Models Good to Use

- Have a dataset with inbreeding or some population structure
- Dataset is filtered down to samples and SNPs with:
  - “Good” Call Rate
  - SNP MAF > 0.05 (common variants)
- Whole Genome Sequencing data is fine if looking for common variants
- NOT for RARE VARIANT ANALYSIS!!!!

Which Model to Use?

Regression with PCA	• Homogeneous datasets or datasets with simple structure
EMMAX	•Complex diseases on a structured population, assuming all loci have a small effect on the trait
MLMM	• Complex diseases on a structured population, assuming there are several loci that have a large effect on the trait and the rest have small effects on the trait
GBLUP	• Obtain estimated breeding values, rank allele substitution effects to find QTL or find genomic relationship matrix in structured populations

Agenda

1 Background of GWAS Approaches

2 Review of Select Mixed Model Methods

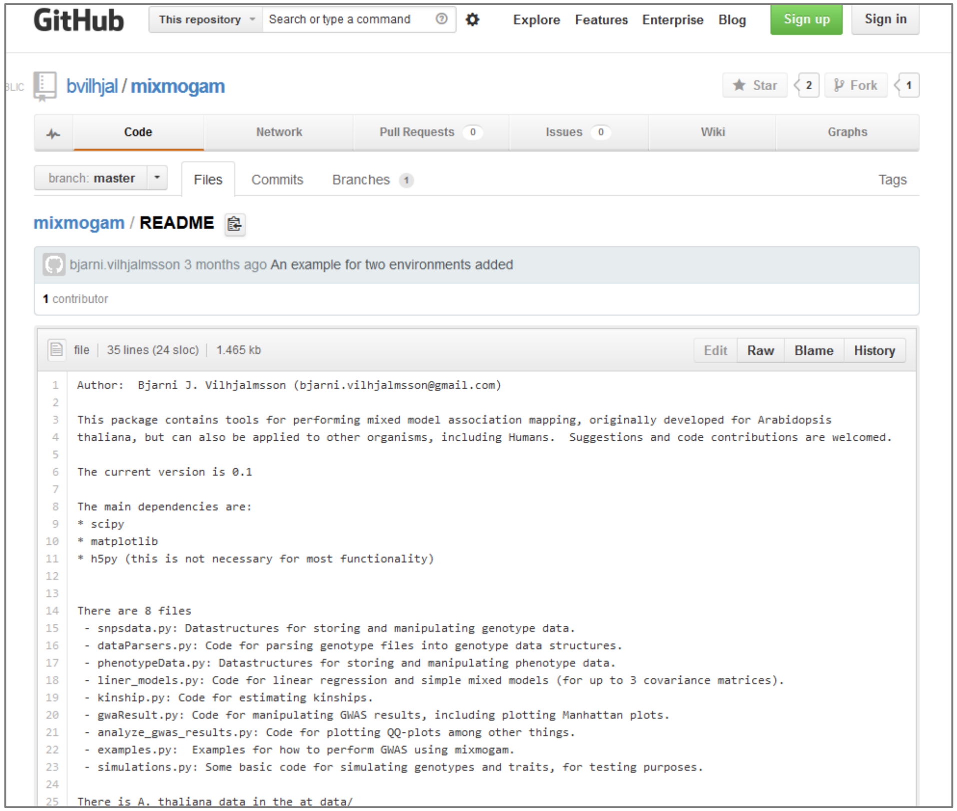
3 Mixed Models in SVS

4 Demo

5 Compare Results

Mixed Models in SVS

- Derived from the mixmogam python package
- By B. Vilhjalmsson, coauthor of MLMM paper*
- Note, GBLUP also uses utilities from mixmogam



bjarni.vilhjalmsson 3 months ago An example for two environments added

1 contributor

file | 35 lines (24 sloc) | 1.465 kb

```
Author:  Bjarni J. Vilhjalmsson (bjarni.vilhjalmsson@gmail.com)

This package contains tools for performing mixed model association mapping, originally developed for Arabidopsis
thaliana, but can also be applied to other organisms, including Humans.  Suggestions and code contributions are welcomed.

The current version is 0.1

The main dependencies are:
* scipy
* matplotlib
* h5py (this is not necessary for most functionality)


There are 8 files
 - snpsdata.py: Datastructures for storing and manipulating genotype data.
 - dataParsers.py: Code for parsing genotype files into genotype data structures.
 - phenotypeData.py: Datastructures for storing and manipulating phenotype data.
 - liner_models.py: Code for linear regression and simple mixed models (for up to 3 covariance matrices).
 - kinship.py: Code for estimating kinships.
 - gwaResult.py: Code for manipulating GWAS results, including plotting Manhattan plots.
 - analyze_gwas_results.py: Code for plotting QQ-plots among other things.
 - examples.py:  Examples for how to perform GWAS using mixmogam.
 - simulations.py: Some basic code for simulating genotypes and traits, for testing purposes.

There is A. thaliana data in the at data/
```

* V. Segura et al. “An efficient multi-locus mixed model approach for genome-wide association studies in structured populations” (Nat Genetics, 2012)

SVS Implementation

- Provides user friendly interface for:
  - GBLUP
  - Mixed Linear Models Analysis
- Runs directly from a spreadsheet and has an options dialog where you can select your fixed factors and other parameters
- Visualization of results in SVS' Genome Browser is quick and easy
- Optimized so that analyses run fast

Agenda

1 Background of GWAS Approaches

2 Review of Select Mixed Model Methods

3 Mixed Models in SVS

4 **Demo**

5 Compare Results

Why Sheep? What about Humans or…?

- The Sheep HapMap dataset was chosen because of
  - the large number of samples and
  - the large number of breeds
- The dataset was available for public use on request from the ISGC
- The dataset was sufficiently structured enough to demonstrate all of the four methods

- Other datasets used by Mixed Model papers include:
  - WTCCC (all diseases including RA and T1D)
  - NFBC66
  - Arabidopsis thaliana dataset
  - Zea maize dataset
  - Various cattle datasets
- Mixed models used on datasets not expected to have family structure (WTCCC and NFBC66)

First a little about the dataset…

- Sheep HapMap SNP50_Breedv1 dataset
- Illumina 50k SNP array
- 49,034 markers were left after filtering by the consortium
- 110 unmapped markers
- Only 1 marker in Chr Y

International Sheep Genomics Consortium

The International Sheep Genomics Consortium is a partnership of scientists and funding agencies from Australia, Austria, Brazil, China, Finland, France, Germany, Greece, India, Iran, Israel, Italy, Kenya, New Zealand, Norway, Spain, Switzerland, Turkey, United Kingdom and United States to develop public genomic resources that will help researchers find genes associated with production, quality and disease traits in sheep.

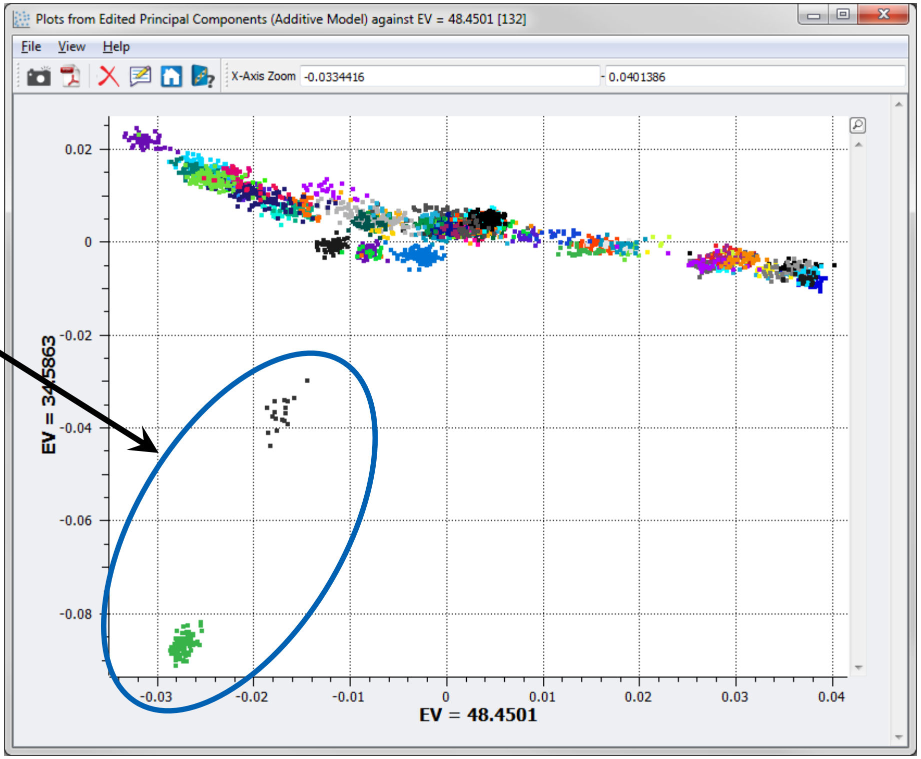
The project commenced informally in 2002 with the creation of a high quality ovine BAC library, and was built on an existing collaboration for the International Mapping Flock that was created nearly a decade earlier.

What's New

- Access Details for the new High Density (HD) SNP Chip
- ISGC SNP Loci For Parentage
- ISGC Releases Sheep Genome Assembly OARv2.0
- ISGC Presentations from PAG conference 2011
- Sheep Community submits White Paper to USDA

Sample Statistics/Filtering

- Removed samples from Boreray & Soay breeds
  - 72 Breeds & Cross-Breeds left
- Imputed gender from heterozygosity rates in the X chromosome
  - Males: 1611
  - Females: 1081

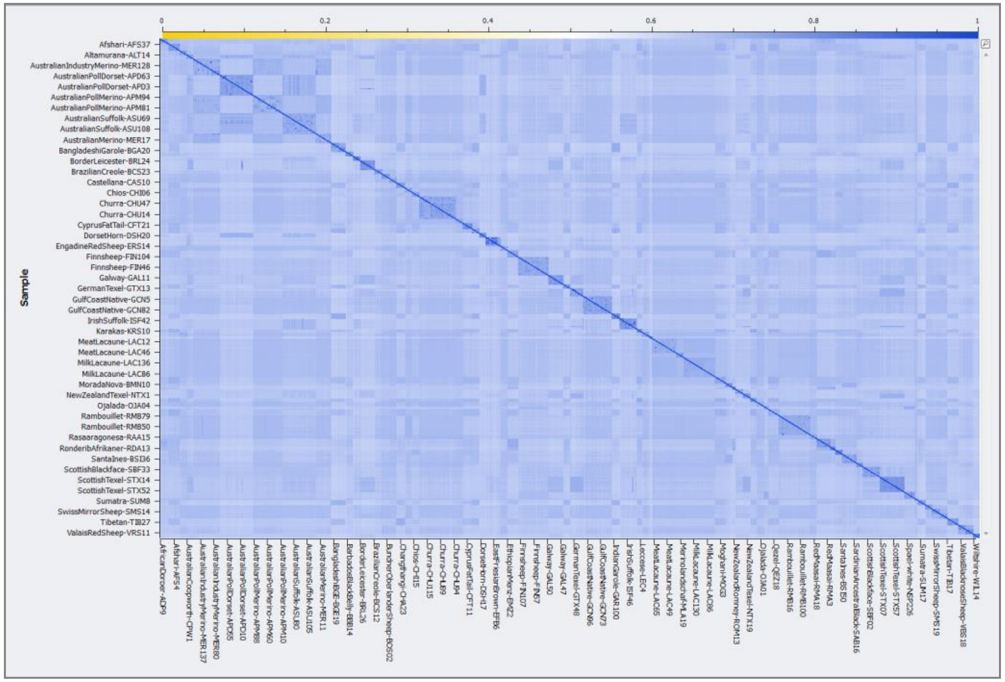


Plots from Edited Principal Components (Additive Model) against EV = 48.4501 [132]

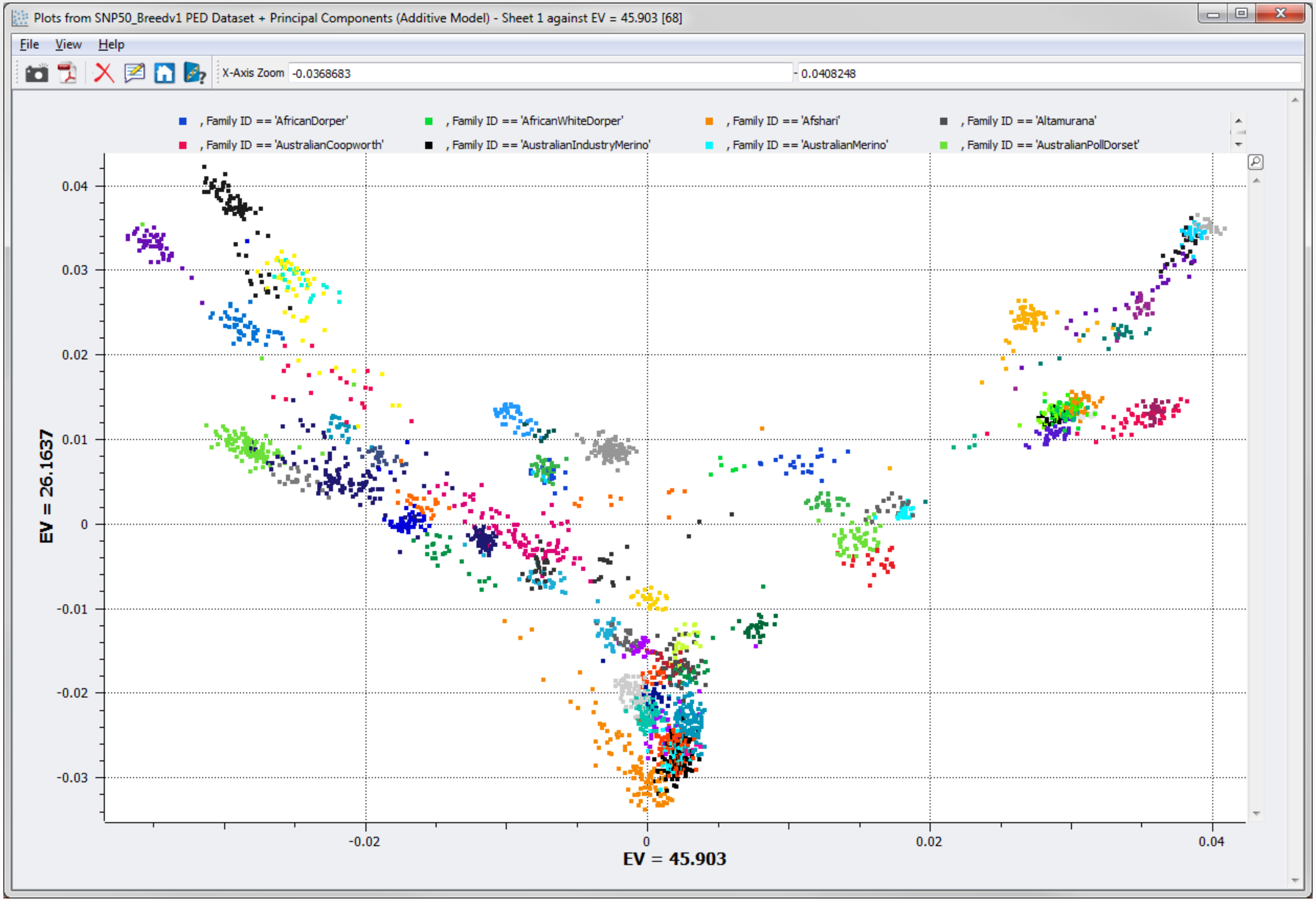
X-Axis Zoom -0.0334416 - 0.0401386

IBS and PCA on Marker Subset

- Filtered down to $MAF \geq\ 0.05$
- LD pruned
  - $R^2 \geq\ 0.5$ (CHM method)
  - Window of 50 markers
  - Step size of 5 markers
- Left 45,117 total markers (44,057 autosomal markers)
- Performed IBS & PCA analysis on remaining samples and markers



Sheep HapMap PCA Plot



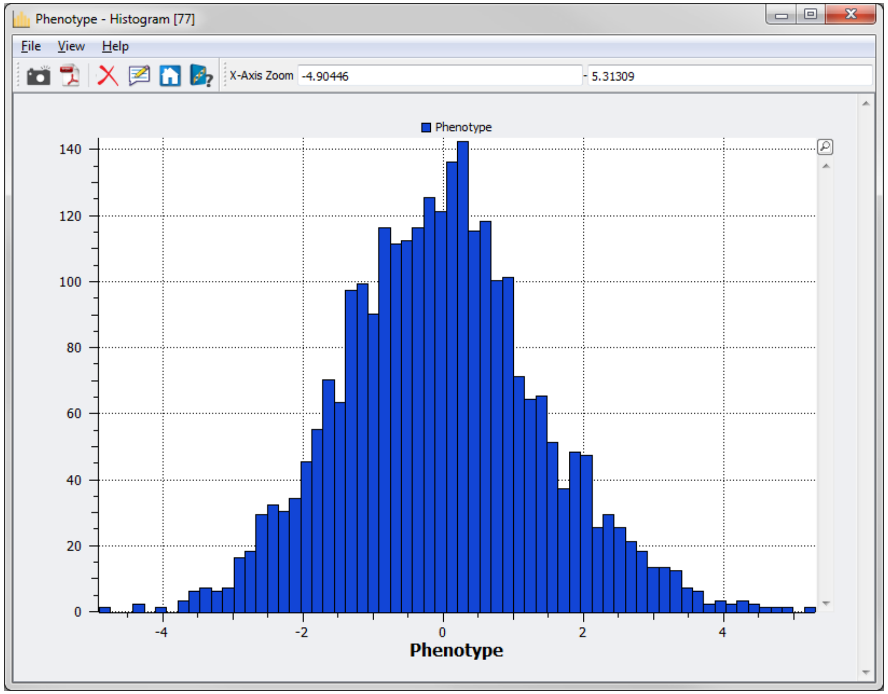
Plots from SNP50_Breedv1 PED Dataset + Principal Components (Additive Model) - Sheet 1 against EV = 45.903 [68]

X-Axis Zoom -0.0368683 - 0.0408248

File
View
Help
🟦 , Family ID == 'AfricanDorper'
🟩 , Family ID == 'AfricanWhiteDorper'
🟧 , Family ID == 'Afshari'
⬛ , Family ID == 'Altamurana'
🟥 , Family ID == 'AustralianCoopworth'
⬛ , Family ID == 'AustralianIndustryMerino'
🟦 , Family ID == 'AustralianMerino'
🟩 , Family ID == 'AustralianPollDorset'
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
EV = 26.1637
-0.02
0
0.02
0.04
EV = 45.903

Simulated Phenotype

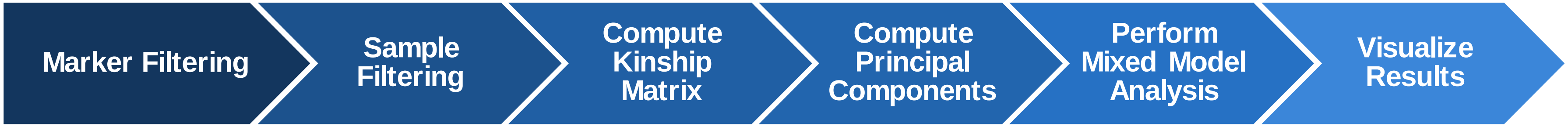
- Filtered markers down to those within predicted gene regions
- Randomly selected 25 causal markers
- Set $h^2 = 0.4$
- Used a χ^2 distribution for the effect sizes
- Added an error adjustment from a skewed normal distribution



Phenotype - Histogram [77]

X-Axis Zoom -4.90446 - 5.31309

Analysis steps



Marker Filtering

- Call Rate
- MAF
- LD Prune for Kinship, PCA

Sample Filtering

- Call Rate

Compute Kinship Matrix

- IBS, or
- IBD, or
- Gen Rel Matrix

Compute Principal Components

- On filtered, LD pruned Dataset
- X Chr filtered out

Perform Mixed Model Analysis

- EMMAX
- MLMM
- GBLUP

Visualize Results

- P-value plot
- Venn Diagram

GOLDEN HELIX

SNP & VARIATION SUITE 7

[Demo]

Agenda

1 Background of GWAS Approaches

2 Review of Select Mixed Model Methods

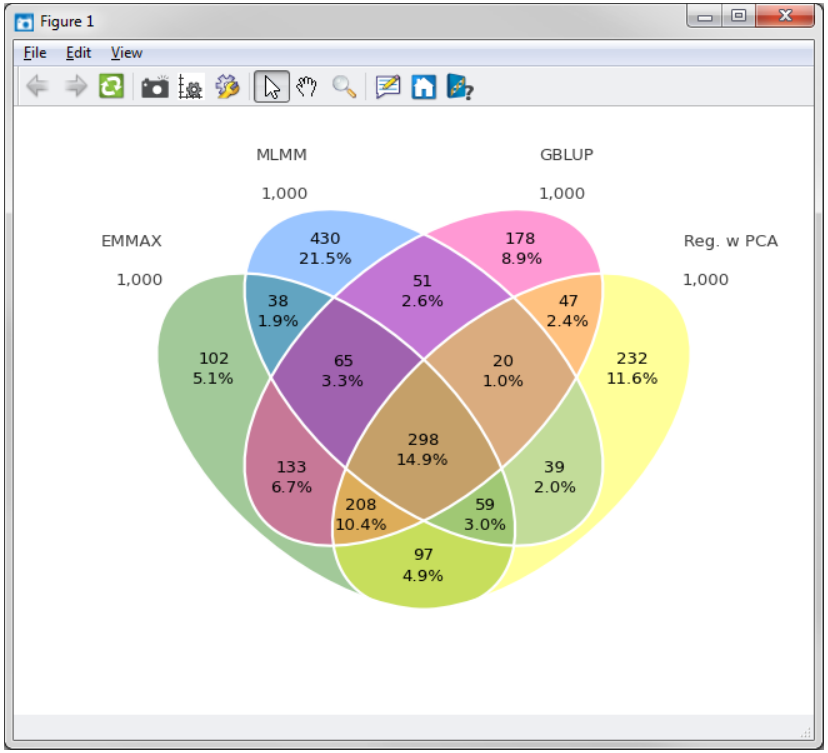
3 Mixed Models in SVS

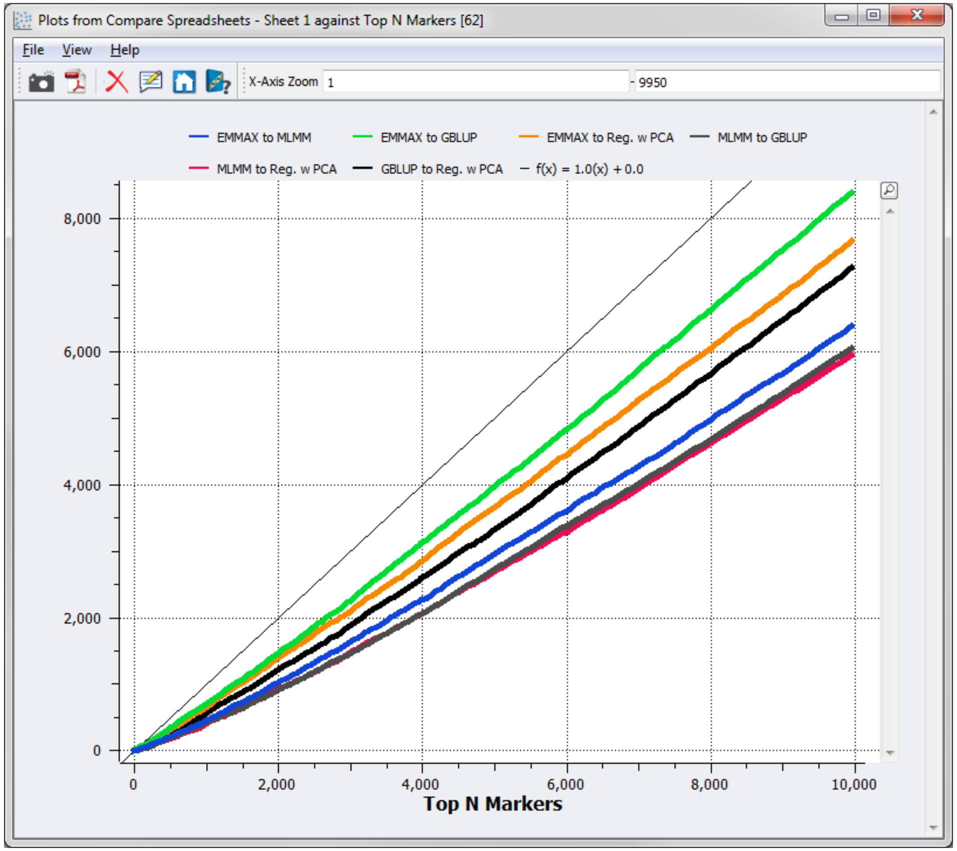
4 Demo

5 **Compare Results**

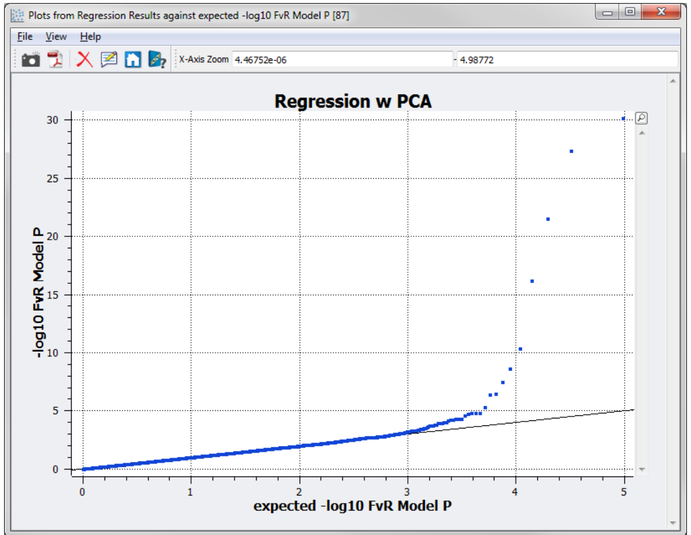
Compare the methods

- Top 1000 markers show some overlap in results



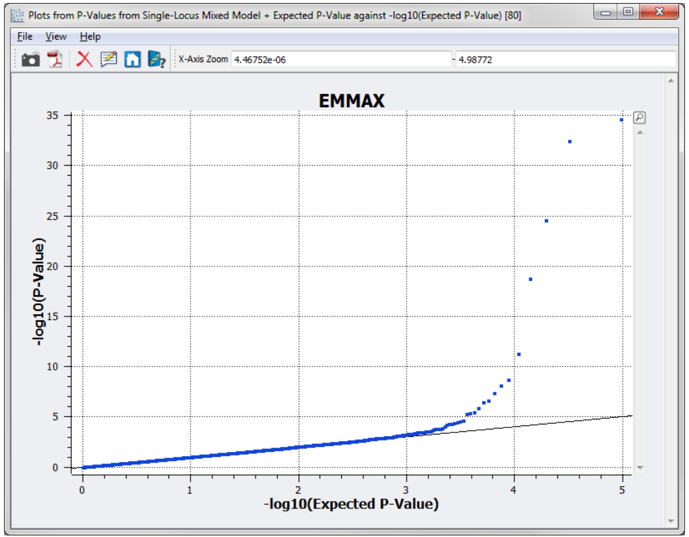


QQ Plots of methods



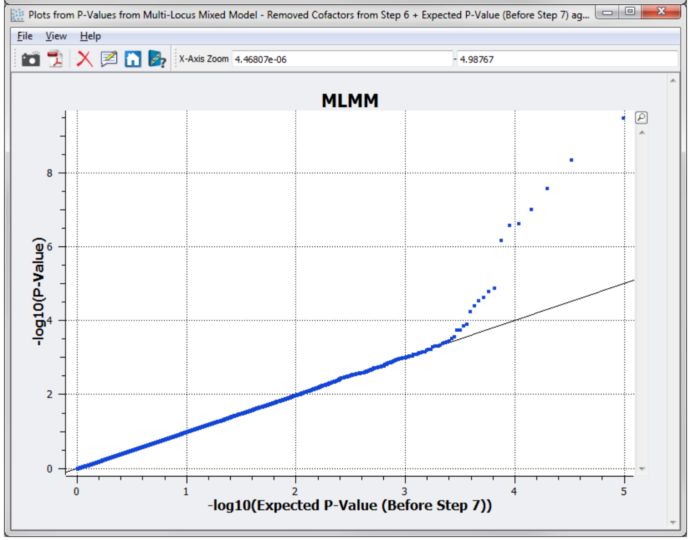
Plots from Regression Results against expected -log10 FvR Model P [87]

X-Axis Zoom 4.46752e-06 - 4.98772



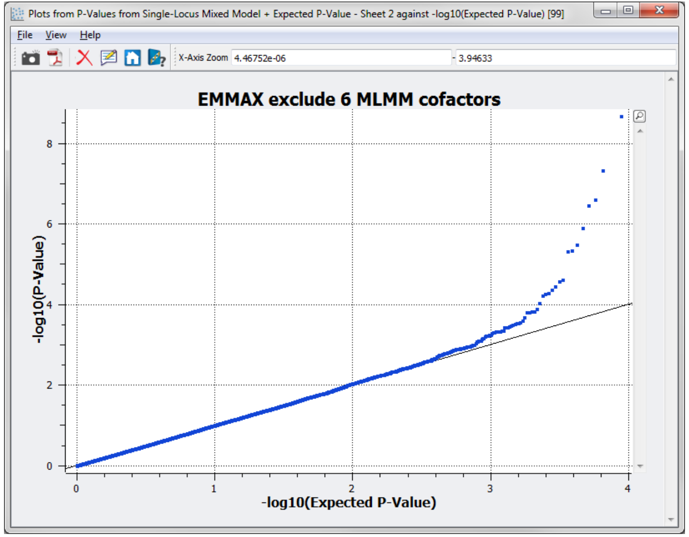
Plots from P-Values from Single-Locus Mixed Model + Expected P-Value against -log10(Expected P-Value) [80]

X-Axis Zoom 4.46752e-06 - 4.98772



Plots from P-Values from Multi-Locus Mixed Model - Removed Cofactors from Step 6 + Expected P-Value (Before Step 7) ag...

X-Axis Zoom 4.46807e-06 - 4.98767

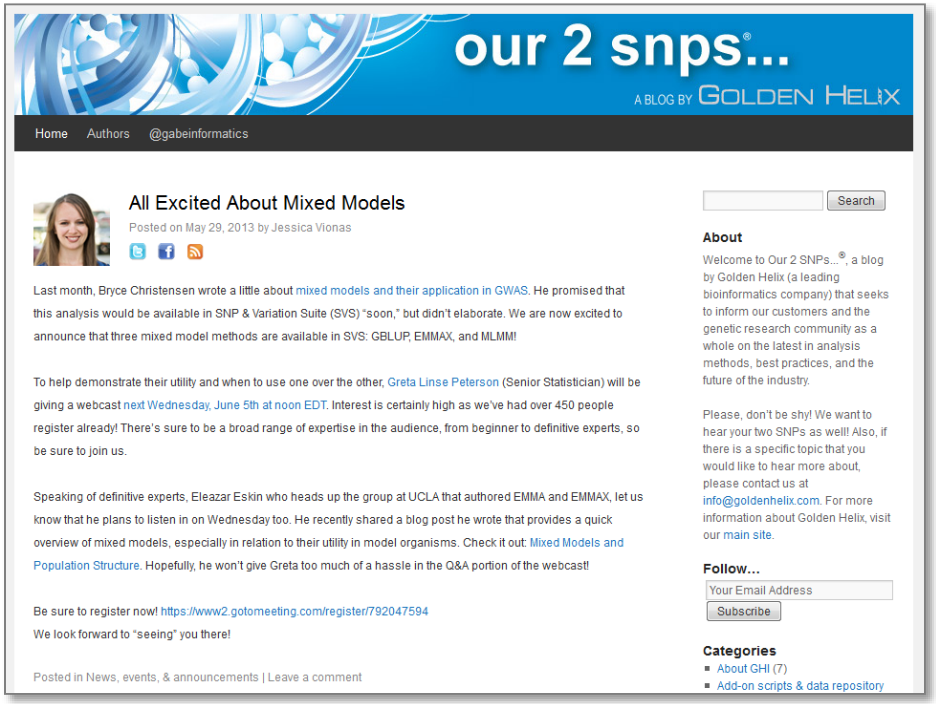


Plots from P-Values from Single-Locus Mixed Model + Expected P-Value - Sheet 2 against -log10(Expected P-Value) [99]

X-Axis Zoom 4.46752e-06 - 3.94633

Conclusion

- Mixed models can be a useful tool when applied appropriately.
- Use the method best suited for your data.
- Mixed models are not the "cure all" for bad data.
- Watch for a blog post to come out later this week on more mixed model methods!



our 2 snps...®

A BLOG BY GOLDEN HELIX

All Excited About Mixed Models

Posted on May 29, 2013 by Jessica Vionas

Last month, Bryce Christensen wrote a little about mixed models and their application in GWAS. He promised that this analysis would be available in SNP & Variation Suite (SVS) "soon," but didn't elaborate. We are now excited to announce that three mixed model methods are available in SVS: GBLUP, EMMAX, and MLMM!

To help demonstrate their utility and when to use one over the other, Greta Linse Peterson (Senior Statistician) will be giving a webcast next Wednesday, June 5th at noon EDT. Interest is certainly high as we've had over 450 people register already! There's sure to be a broad range of expertise in the audience, from beginner to definitive experts, so be sure to join us.

Speaking of definitive experts, Eleazar Eskin who heads up the group at UCLA that authored EMMA and EMMAX, let us know that he plans to listen in on Wednesday too. He recently shared a blog post he wrote that provides a quick overview of mixed models, especially in relation to their utility in model organisms. Check it out: Mixed Models and Population Structure. Hopefully, he won't give Greta too much of a hassle in the Q&A portion of the webcast!

Be sure to register now! https://www2.gotomeeting.com/register/792047594
We look forward to "seeing" you there!

Posted in News, events, & announcements | Leave a comment

About

Welcome to Our 2 SNPs...®, a blog by Golden Helix (a leading bioinformatics company) that seeks to inform our customers and the genetic research community as a whole on the latest in analysis methods, best practices, and the future of the industry.

Please, don't be shy! We want to hear your two SNPs as well! Also, if there is a specific topic that you would like to hear more about, please contact us at info@goldenhelix.com. For more information about Golden Helix, visit our main site.

Follow...

Categories

- About GHI (7)
- Add-on scripts & data repository

Acknowledgements

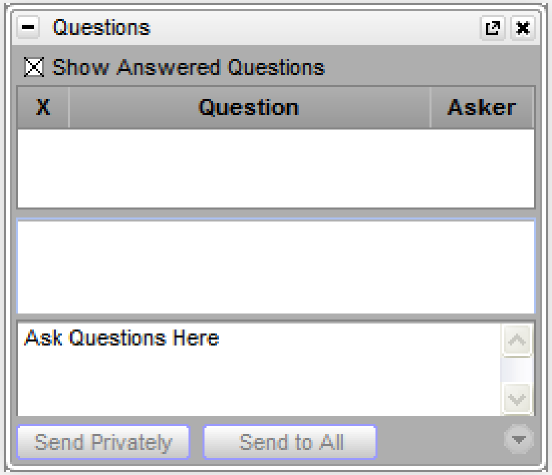
- Bjarni Vihjálmsson
- Christopher Seabury
- John McEwan of ISGC

References

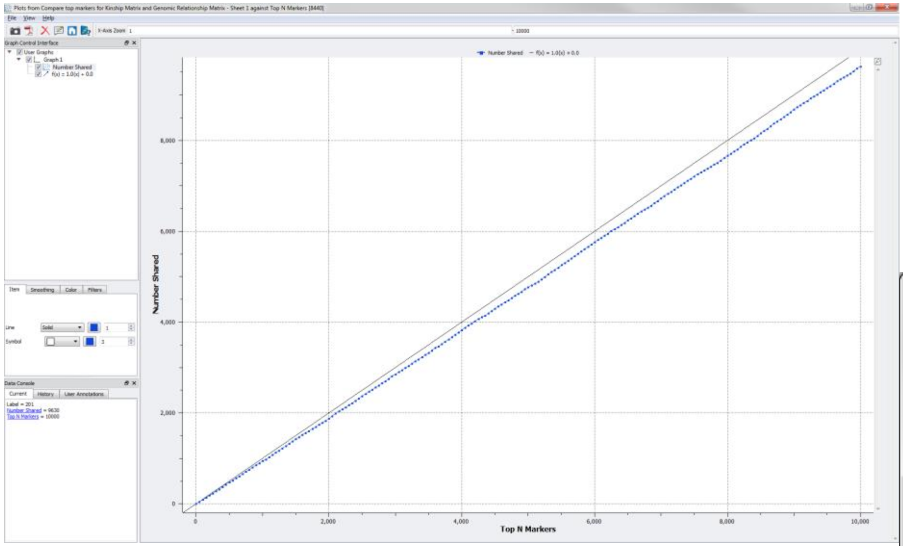
- Kang HM, et al (2008). ‘Efficient control of population structure in model organism association mapping’, Genetics, 178, 1709–1723.
- Kang HM, et al (2010). ‘Variance component model to account for sample structure in genome-wide association studies’, Nature Genetics 42, 348–354.
- Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190.
- Segura V, Vihjálmsson BJ, Platt A, Korte A, Seren Ü, et al. (2012) ‘An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations’, Nature Genetics, 44, 825–830.
- Taylor, J.F. (2013) ‘Implementation and accuracy of genomic selection’, Aquaculture, http://dx.doi.org/10.1016/j.aquaculture.2013.02.017
- VanRaden, P.M. (2008) ‘Efficient Methods to Compute Genomic Predictions’, J. Dairy Sci, 91, pp. 4414–4423.
- Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, Paiva S, Barendse W, Ciani E, Raadsma H, McEwan J, Dalrymple B; International Sheep Genomics Consortium Members. “Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection.” PLoS Biol. 2012 Feb;10(2):e1001258. doi: 10.1371/journal.pbio.1001258. Epub 2012 Feb 7

Do You Have Any Questions?

Use the Questions pane in your GoToWebinar window



IBS vs Genomic Relationship Matrix



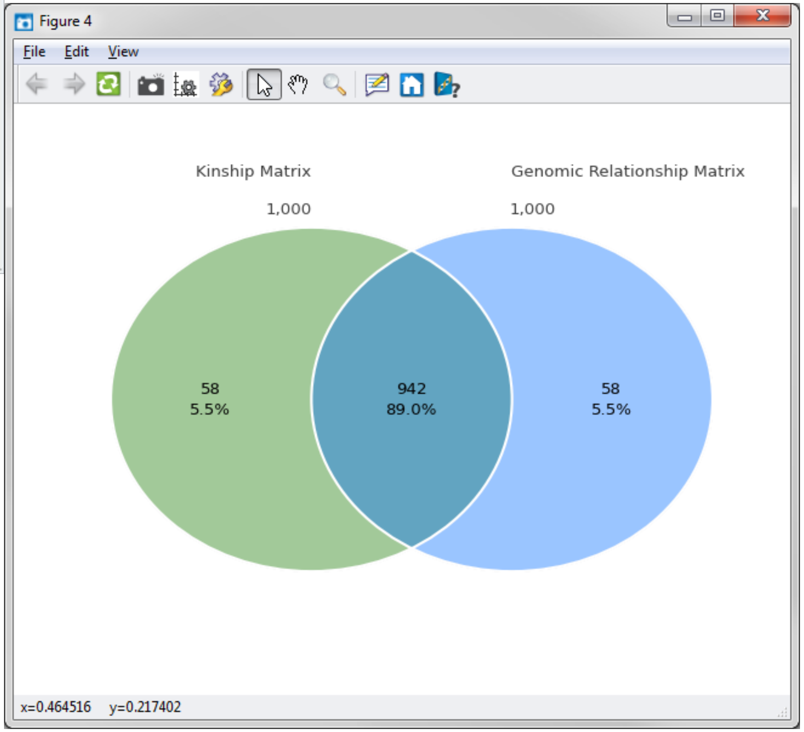
Plots from Compare top markers for Kinship Matrix and Genomic Relationship Matrix - Sheet 1 against Top N Markers [8440]

X-Axis Zoom 1 - 10000

Line Solid 1

Symbol 3

Label = 201
Number Shared = 9630
Top N Markers = 10000



x=0.464516 y=0.217402