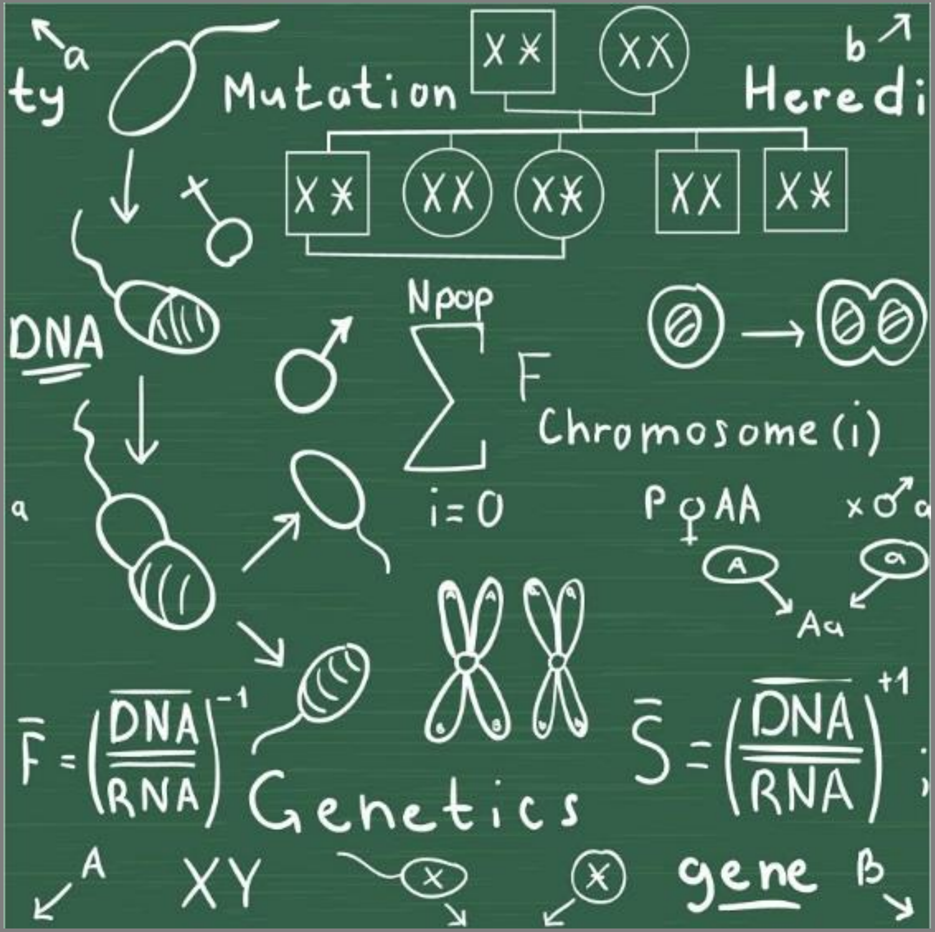


MM-KBAC: Using mixed models to adjust for population structure in a rare-variant burden test

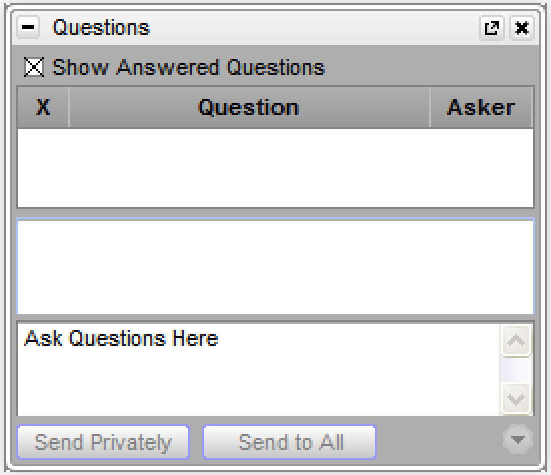
Tuesday, June 10, 2014

Greta Linse Peterson
Director of Product Management and Quality

GOLDEN HELIX
Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



Questions

☒ Show Answered Questions

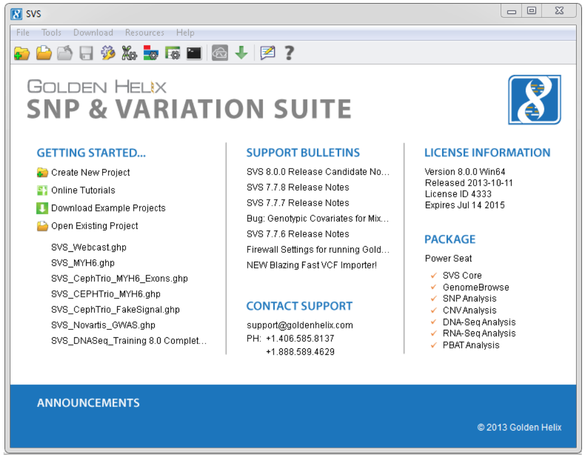
X	Question	Asker

Ask Questions Here

Golden Helix Offerings

Services

- Genomic Analytics
- Genotype Imputation
- Workflow Automation
- SVS Certification & Training



GOLDEN HELIX
SNP & VARIATION SUITE

GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CEPHTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Novartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

ANNOUNCEMENTS

© 2013 Golden Helix

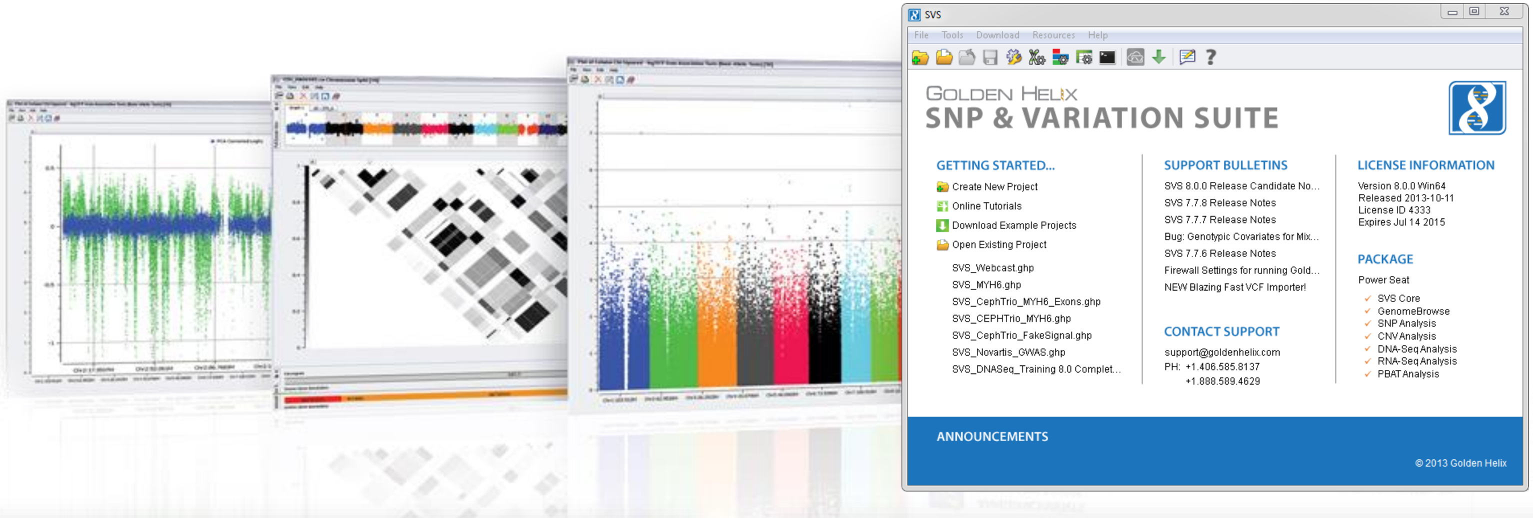
Software

- SNP & Variation Suite (SVS) for NGS, SNP, & CNV data
- GenomeBrowse
- New products in development

Support

- Support comes standard with software
- Customers rave about our support
- Extensive online materials including tutorials and more

SNP & Variation Suite (SVS)



GOLDEN HELIX
SNP & VARIATION SUITE

GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CEPHTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Novartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

ANNOUNCEMENTS

© 2013 Golden Helix

Core Features

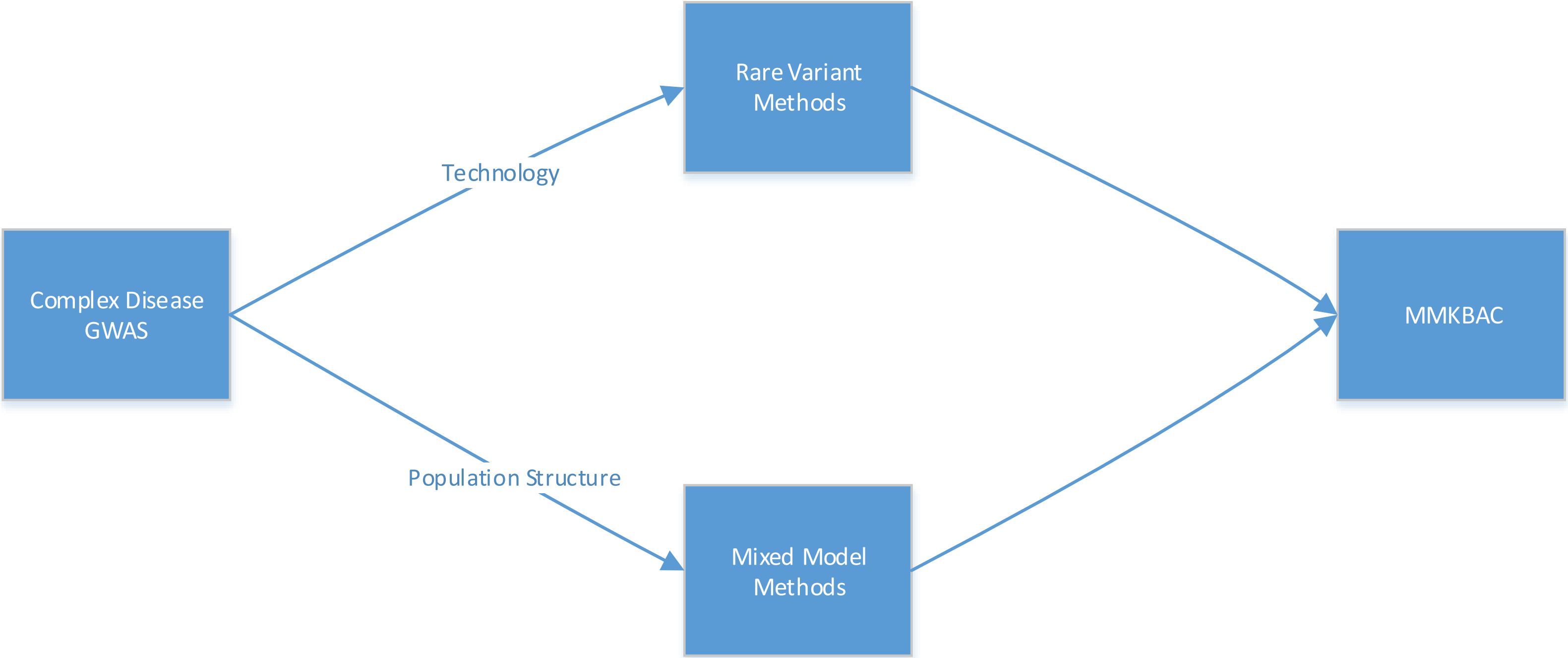
- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible
- Easy-to-use

Applications

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

GOLDEN HELIX
Accelerating the Quest for Significance™

Timeline


Complex Disease GWAS

Technology

Population Structure

Rare Variant Methods

Mixed Model Methods

MMKBAC

Study Design

- **Large cohort population based design (cases with matched controls or quantitative phenotypes and complex traits)**
  - Assumes: independent and well matched samples
  - Can interrogate complex traits
- **Small families (trios, quads, small extended pedigrees)**
  - Can only analyze a single family at a time, looking for de Novo, recessive or compound het variants unique to an affected sample in a single family
  - Looking for highly penetrant variants

What If????

- **What if we have:**
  - Known population structure
  - Cannot guarantee independence between samples
  - Controls were borrowed from a different study
  - Multiple families with affected offspring all exhibiting the same phenotype
  - Multiple large extended pedigrees of unknown structure

Just Add Random Effects!

- Why can’t we just add random effects to our regression models for our rare-variant burden testing algorithms?
- Existing mixed model algorithms assume a linear model
- Kernel-based adaptive clustering (KBAC) uses a logistic regression model
- Hmm what to do….?

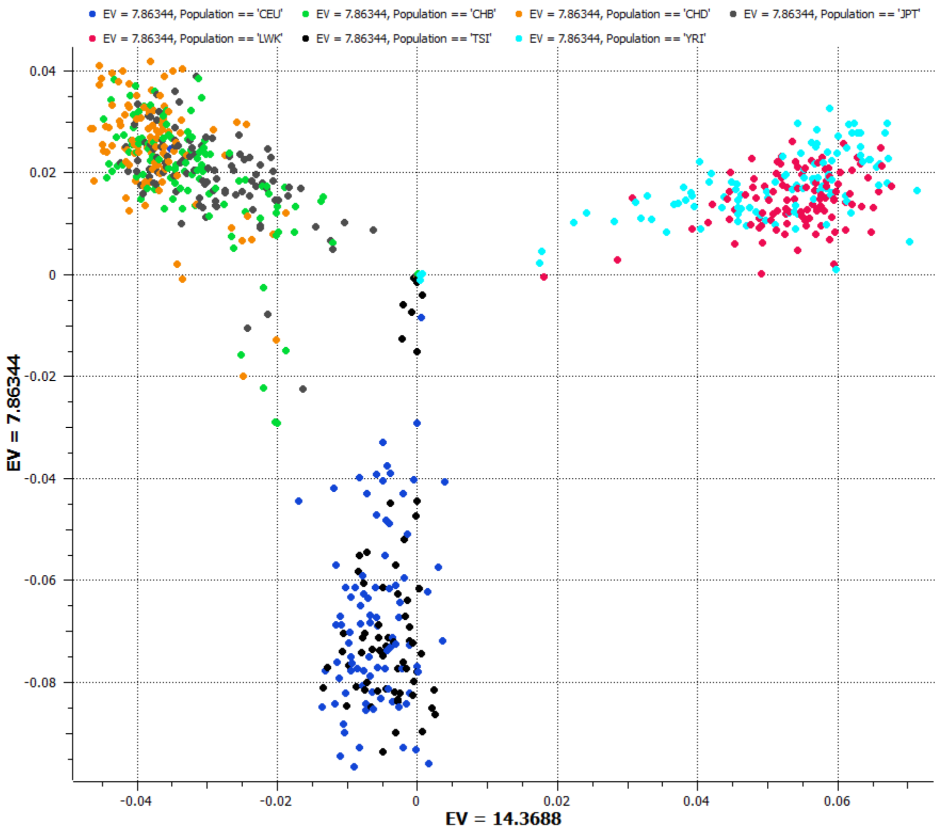
WARNING!

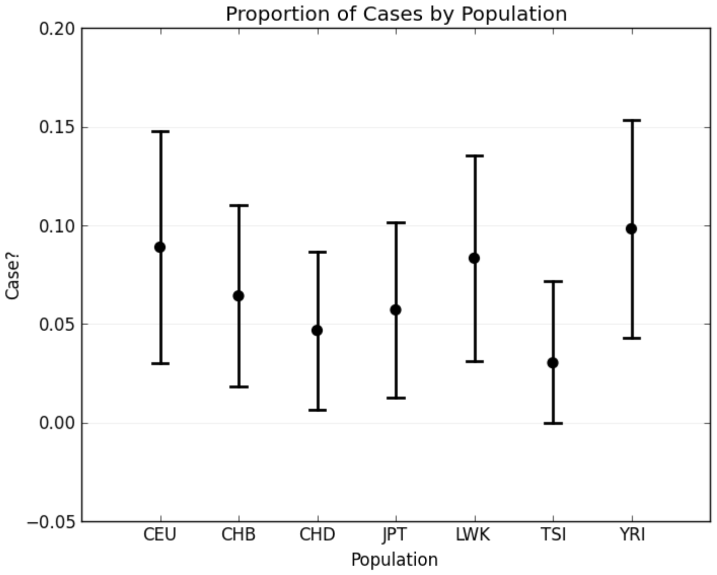
What is about to follow are formulas and statistics, specifically matrix algebra…

But don't worry we'll end the webcast with a presentation of some preliminary results! So hang in there!

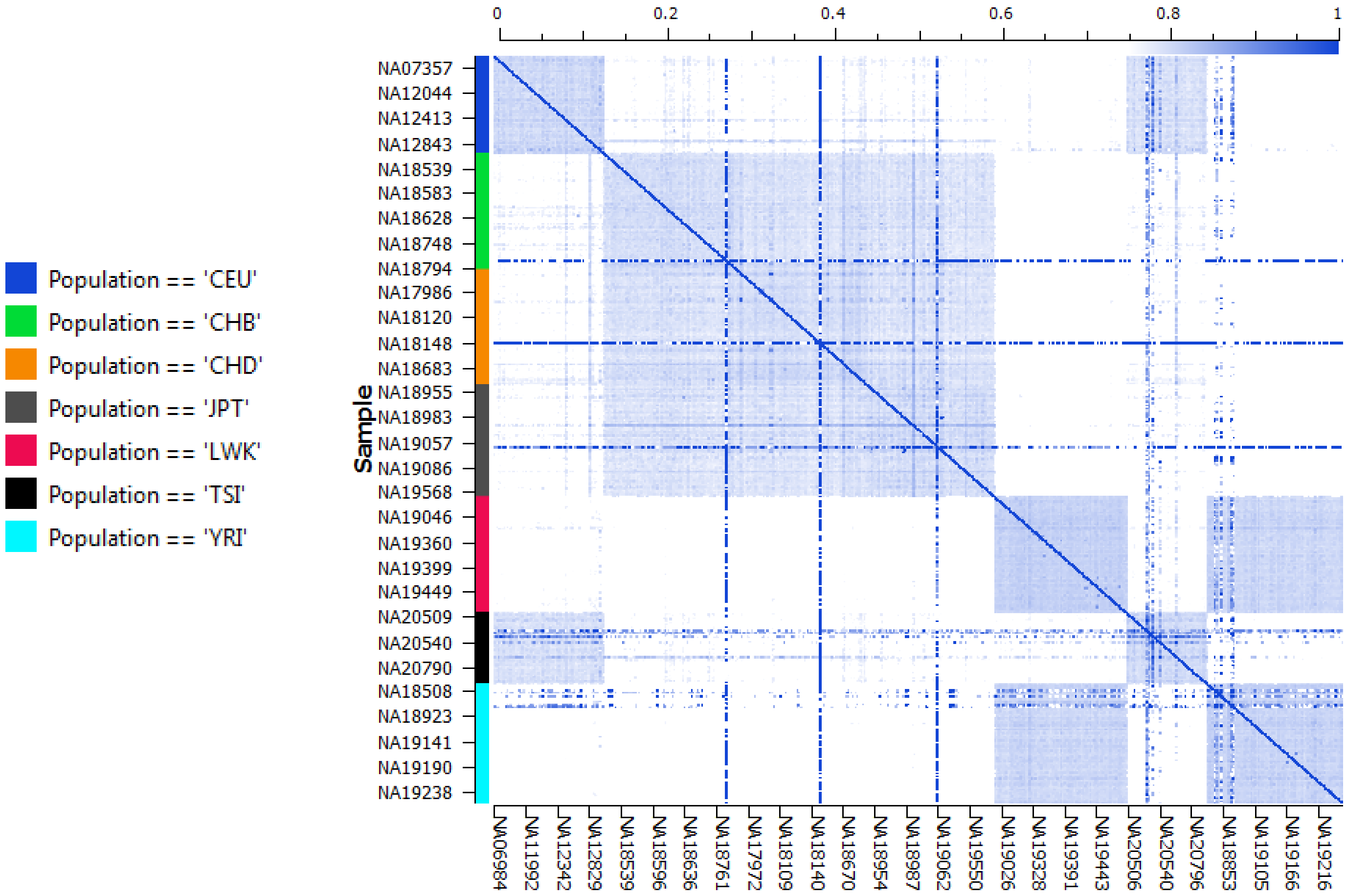
But first….

The dataset we have chosen for today is the 1000 Genomes Pilot 3 Exons dataset with a simulated phenotype.





Relatedness of samples



Why Mixed Models + KBAC?

- OK Mixed Models makes sense, but why KBAC?
- KBAC was chosen as our proof of concept rare-variant burden test for complex traits
- KBAC uses a score test which is trivial to calculate once you compute the reduced model
- Mixed models can be added to other burden and kernel tests using the same principles

What is KBAC?

- KBAC = Kernel-based Adaptive Clustering
- Catalogs and counts multi-marker genotypes based on variant data
- Assumes the data has been filtered to only rare variants
- Performs a special case/control test based on the counts of variants per region (aka gene)
- Test is weighted based on how often each genotype is expected to occur according to the null hypothesis
- Genotypes with higher sample risks are given higher weights
- One-sided test primarily, which means it detects higher sample risks

OPEN ACCESS Freely available online PLoS GENETICS

A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions

Dajiang J. Liu[1,2], Suzanne M. Leal[1,2]*

[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, [2] Department of Statistics, Rice University, Houston, Texas, United States of America

Abstract

There is solid evidence that rare variants contribute to complex disease etiology. Next-generation sequencing technologies make it possible to uncover rare variants within candidate genes, exomes, and genomes. Working in a novel framework, the kernel-based adaptive cluster (KBAC) was developed to perform powerful gene/locus based rare variant association testing. The KBAC combines variant classification and association testing in a coherent framework. Covariates can also be incorporated in the analysis to control for potential confounders including age, sex, and population substructure. To evaluate the power of KBAC: 1) variant data was simulated using rigorous population genetic models for both Europeans and Africans, with parameters estimated from sequence data, and 2) phenotypes were generated using models motivated by complex diseases including breast cancer and Hirschsprung's disease. It is demonstrated that the KBAC has superior power compared to other rare variant analysis methods, such as the combined multivariate and collapsing and weight sum statistic. In the presence of variant misclassification and gene interaction, association testing using KBAC is particularly advantageous. The KBAC method was also applied to test for associations, using sequence data from the Dallas Heart Study, between energy metabolism traits and rare variants in *ANGPTL* 3,4,5 and 6 genes. A number of novel associations were identified, including the associations of high density lipoprotein and very low density lipoprotein with *ANGPTL4*. The KBAC method is implemented in a user-friendly R package.

Citation: Liu DJ, Leal SM (2010) A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. PLoS Genet 6(10): e1001156. doi:10.1371/journal.pgen.1001156

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received June 13, 2010; **Accepted** September 10, 2010; **Published** October 14, 2010

Funding: This research is supported by National Institutes of Health grant R01-DC03594 and RC2-HL102926 (SML). DJL is partially supported by a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant No. 5 R90 DK071505-04). Computation for this research was supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467 and by a partnership between Rice University, Sun Microsystems, and Sigma Solutions. The sequence data from ANGPTL genes were provided by Drs. Jonathan Cohen and Helen Hobbs and supported by National Institute of Health grant RL1HL092550 (JC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sleal@bcm.edu

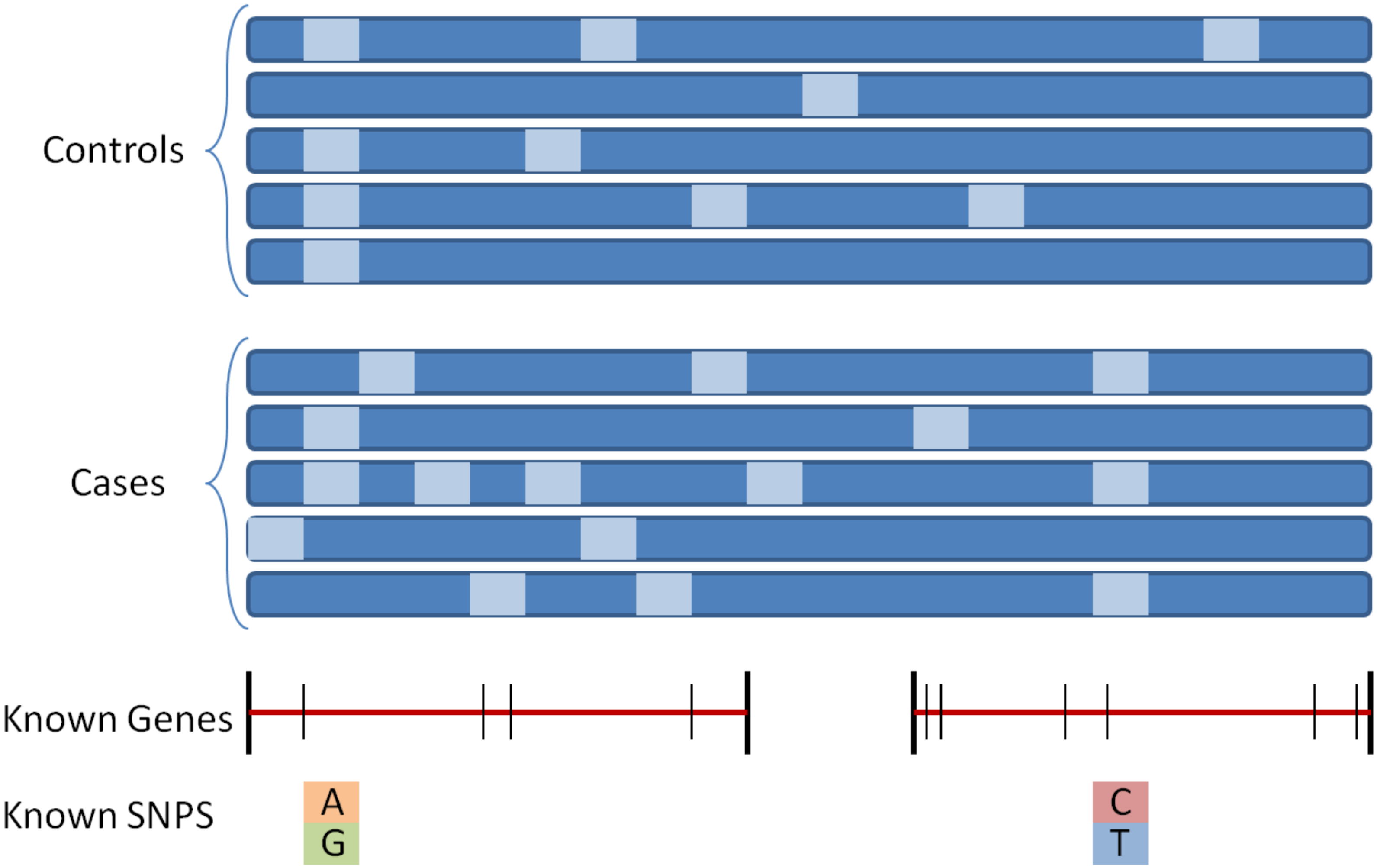
Introduction

Currently there is great interest in investigating the etiology of complex disease due to rare variants [1–6]. Until recently, indirect mapping of common variants has been the emphasis of complex trait association studies. It has been demonstrated that common variants tend to have modest phenotypic effects while rare variants are likely to have stronger phenotypic effects [7], although not strong enough to cause familial aggregation [8]. For mapping complex diseases due to common variants, instead of genotyping functional variants, tagSNPs are genotyped which act as a proxy for the underlying causal variants. For rare variant association studies, indirect mapping is not an optimal approach due to low correlations (r^2) between tagSNPs and rare variants. Instead, direct mapping should be used, where functional variants are analyzed. In order to implement direct mapping, variants must first be identified. Large scale sequencing efforts have begun including the 1000 Genome Project, which will provide a better understanding of the allelic architecture of the genome and a detailed catalog of human variants. Next-generation sequencing technologies e.g. Roche 454, ABI SOLiD, and Illumina HiSeq, have made it feasible to carry-out rare variant association studies of candidate regions, exomes and genomes.

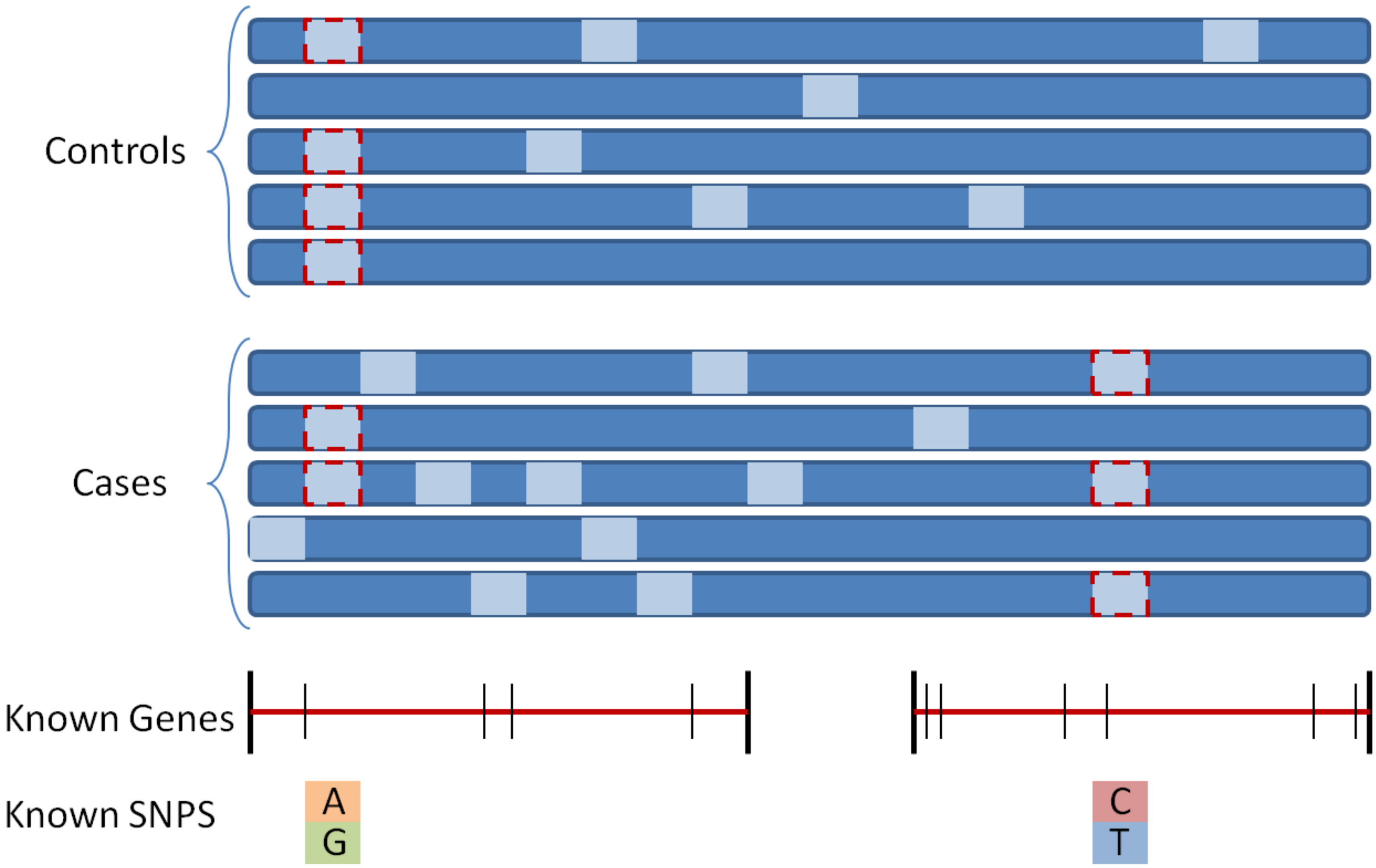
Gene interactions are believed to be involved in a broad spectrum of complex disease etiologies [9]. Although a number of methods have been developed to detect gene interactions between common variants [10–13], their detection has been limited [10]. There is evidence that rare variant interaction also plays a role in disease etiology. In direct association mapping of rare variants, one or more genetic loci are commonly jointly analyzed in order to aggregate information, for example genes with similar functions or residing in the same pathway [3,4]. Therefore it is necessary to account for potential interactions between rare variants in different loci [14] and interactions between common and rare variants [15,16].

Ideally, when carrying out direct mapping, only causal variants should be tested for associations. When DNA samples are sequenced, both causal and non-causal variants are uncovered. Bioinformatics tools [17,18] or filters [1] can be used to predict functionality of

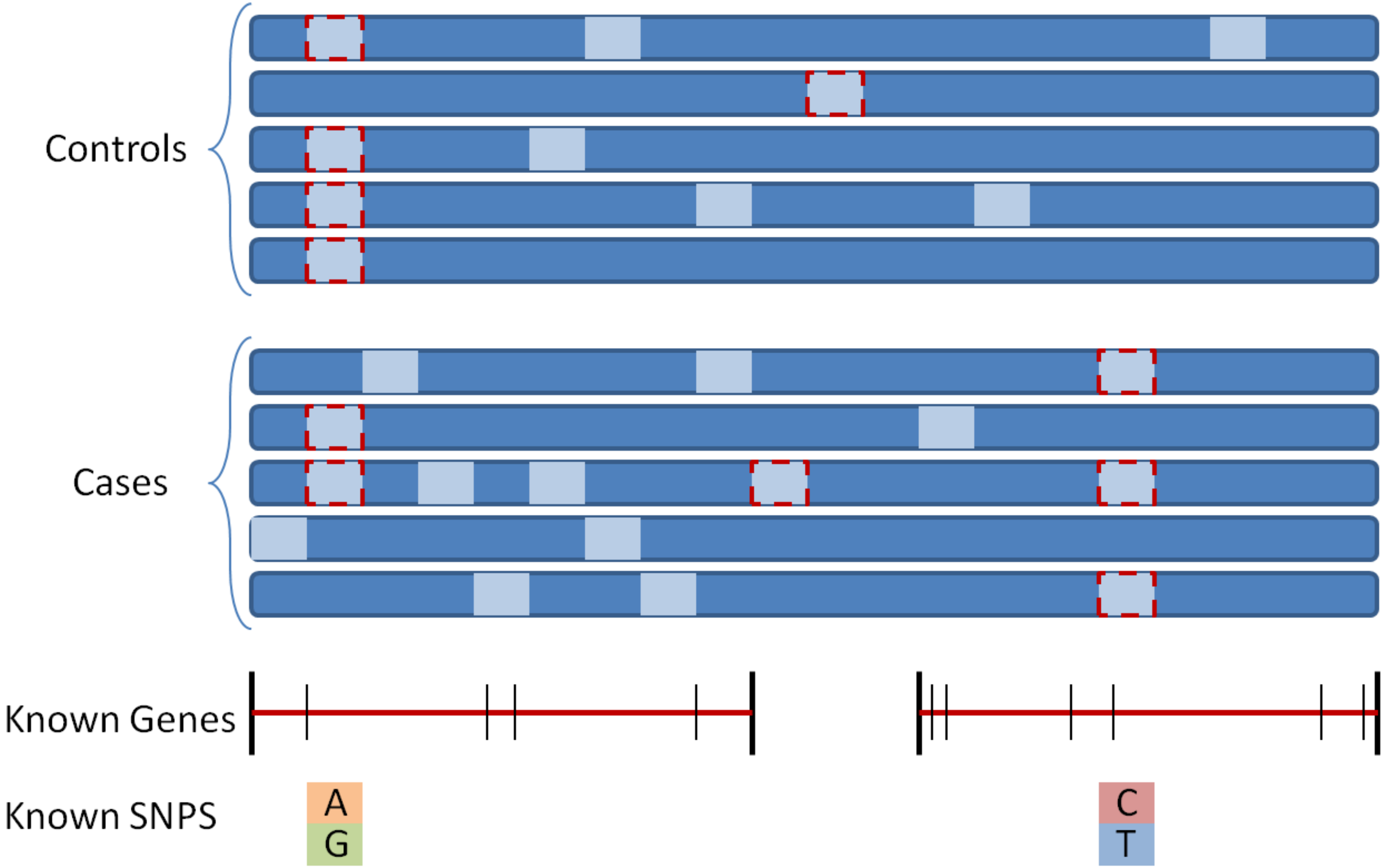
Pictorial Overview of Theory



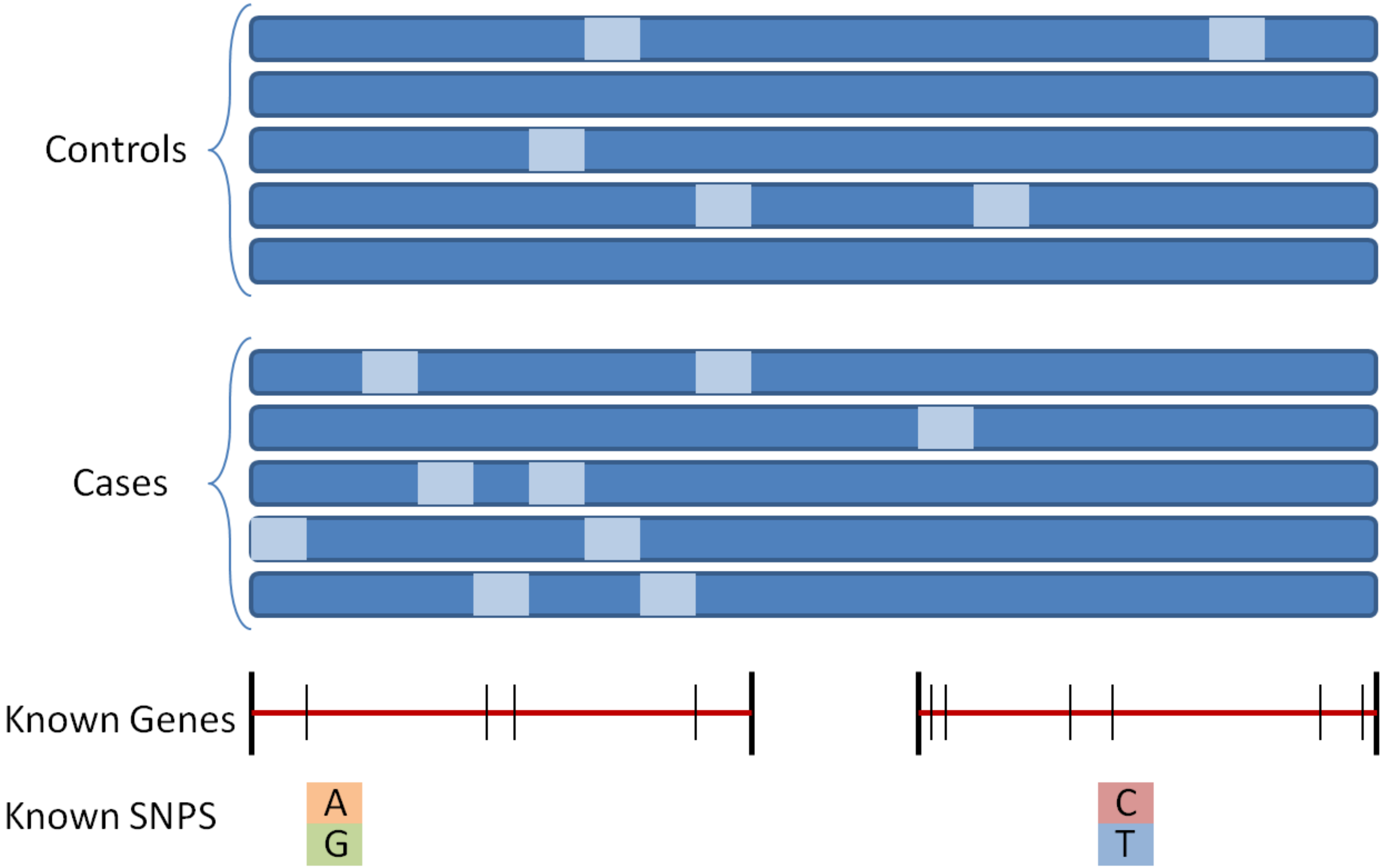
Filter Common/Known SNPs



Filter by Gene Membership



Rare Sequence Variants



KBAC Statistic

- $KBAC_1 = \sum_{i=1}^{k} \left(\frac{N_i^A}{N^A} - \frac{N_i^U}{N^U} \right) K_i^0(\hat{R}_i)$
- **Where the weight is defined as:**

$$w_i = K_i^0(\hat{R}_i) = \int_0^{\hat{R}_i} k_i^0(r)dr$$

The weight can be calculated as a:

- **Hyper-geometric kernel**
- **Marginal binomial kernel**
- **Asymptotic normal kernel**

Determining the KBAC p-value

- **Monte-Carlo Method is used as an approximation for finding the p-value**
- **The number of cases** n_i^A **for each genotype** G_i **approximates a binomial distribution** $n_i^A \sim Binom\left(n_i, \frac{n^A}{n}\right)$
- **The case status is permuted among all samples. The covariates and genotypes are held fixed.**

Logistic Mixed Model Equation

$$\log\left(\frac{P\left(Y_j = 1 \mid X_j, X_{fjl}, u_j\right)}{1 - P\left(Y_j = 1 \mid X_j, X_{fjl}, u_j\right)}\right) = \beta_0 + \beta_1 X_j + \sum_l \left(\beta_{fl} X_{fjl}\right) + u_j$$

Null hypothesis: $H_0 : \beta_1 = 0$

The score statistic to test the null of the independence of the model from X_j is:

$U = \sum_j X_j\left(Y_j - \mu_j\right)$**, where**

$\mu_j = h\left(\eta_j\right) = \frac{e^{\eta_j}}{1+e^{\eta_j}}$**, and**

$\eta_j = \beta_0 + \sum_l \beta_{fl} X_{fjl} + u_j$**, and**

u_j **is the random effect for the** j^(th) **sample.**

Logistic (Reduced) Mixed Model Equation

$$\log\left(\frac{P\left(Y_j = 1 \mid X_{fjl}, u_j\right)}{1 - P\left(Y_j = 1 \mid X_{fjl}, u_j\right)}\right) = \beta_0 + \sum_l \beta_{fl} X_{fjl} + u_j$$

Which can be rewritten as:

$$E[Y|u] = h\left(X_f\beta + u\right) = h(\eta) = \mu$$

And

$$Var[Y|u] = A = A^{1/2}A^{1/2}$$

Where A is the variance of the binomial distribution itself, where

$$\begin{cases} A_{jj} = \mu_j\left(1 - \mu_j\right)\ for\ j = 1..n \\ A_{ij} = 0\ for\ i \neq j \end{cases}$$

And the linear predictor for the model is $\eta = X\beta + u$

While $h(\blacksquare)$ is the inverse link function for the model

Solving the Logistic Mixed Model

Iterate between creating a linear pseudo-model and solving for the pseudo-model's coefficients

$$h(\eta) \doteq h(\tilde{\eta}) + \tilde{\Delta} X \left(\beta - \tilde{\beta}\right) + \tilde{\Delta}(u - \tilde{u})$$

Where

$$\tilde{\Delta} = \left(\frac{\partial h(\eta)}{\partial \eta}\right)_{\tilde{\beta},\tilde{u}}$$

Rearranging yields

$$\tilde{\Delta}^{-1}\left(\mu - h(\tilde{\eta})\right) + X\tilde{\beta} + \tilde{u} \doteq X\beta + u$$

The left side is the expected value, conditional on u, of

$$P \equiv \tilde{\Delta}^{-1}\left(Y - h(\tilde{\eta})\right) + X\tilde{\beta} + \tilde{u}$$

The variance of P given u is

$$Var[P|u] = \tilde{\Delta}^{-1}\tilde{A}^{1/2}\tilde{A}^{1/2}\tilde{\Delta}^{-1}$$

Where

$$\tilde{A}_{jj} = \tilde{\mu}_j\left(1 - \tilde{\mu}_j\right), \tilde{A}_{ij} = 0\ (for\ i \neq j)$$

Transform Pseudo-Model to use EMMA

Pseudo-model: $P = X\beta + u + \epsilon$ **and** $Var[\epsilon] = Var[P|u]$

> NOTE: As an alternative, rather than using the prediction of u from the pseudo-model, we can use the expected value of u, which is zero

Want to solve using EMMA (Kang 2008)

Find T **such that** $Var[T\epsilon] = I$

So that we can write

$$TP = TX\beta + Tu + T\epsilon$$

And use EMMA to solve the mixed model

$$TP = TX\beta + Tu + \epsilon^*$$

Where the variance of ϵ^***is proportional to** I

It can be shown that this is solved by letting

$$T = \tilde{A}^{-1/2}\tilde{\Delta}$$

Summary of the Algorithm

First pick starting values of $\tilde{\beta}$ and $\tilde{u}$, such as all zeros. Repeat the following steps until the changes in $\tilde{\beta}$ and $\tilde{u}$ are sufficiently small:

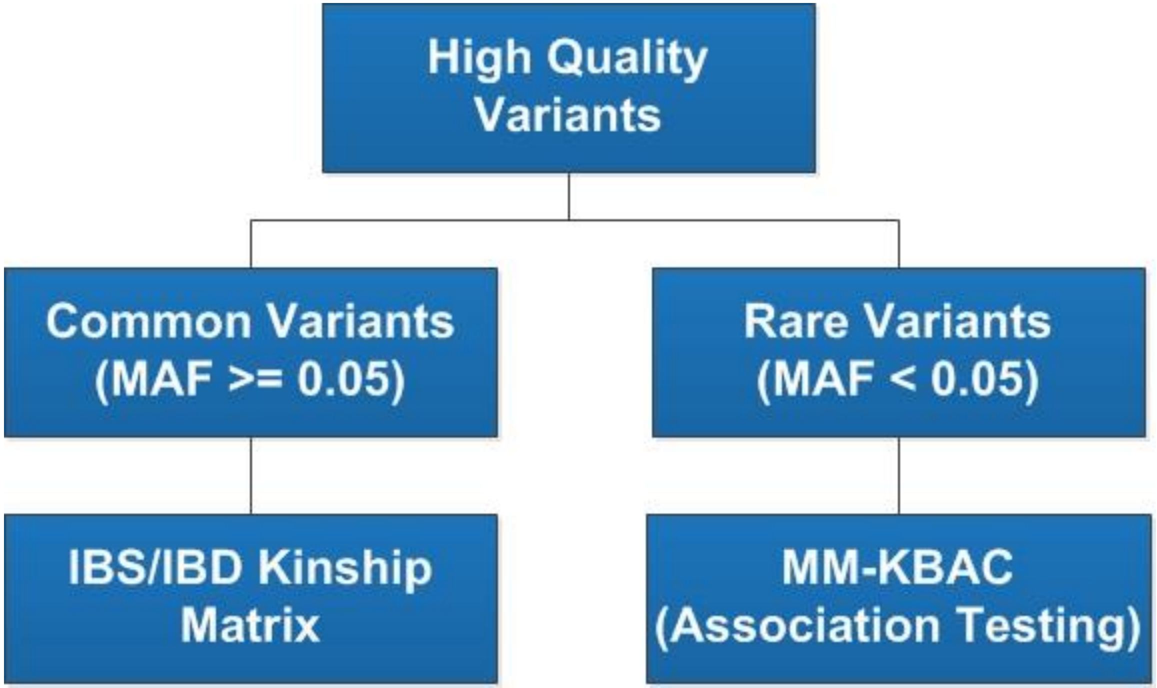
1. **Find $\tilde{\eta}$ and $\tilde{\mu}$ from the original linear predictor equation and the definition of $h(\blacksquare)$**
2. **Find the (diagonal) $\tilde{\Delta}$ matrix**
3. **Find the pseudo-model P**
4. **Find the (diagonal) matrix T**
5. **Solve the following for new values of $\tilde{\beta}$ and $\tilde{u}$ using EMMA:**

$$TP = TX\beta + Tu$$

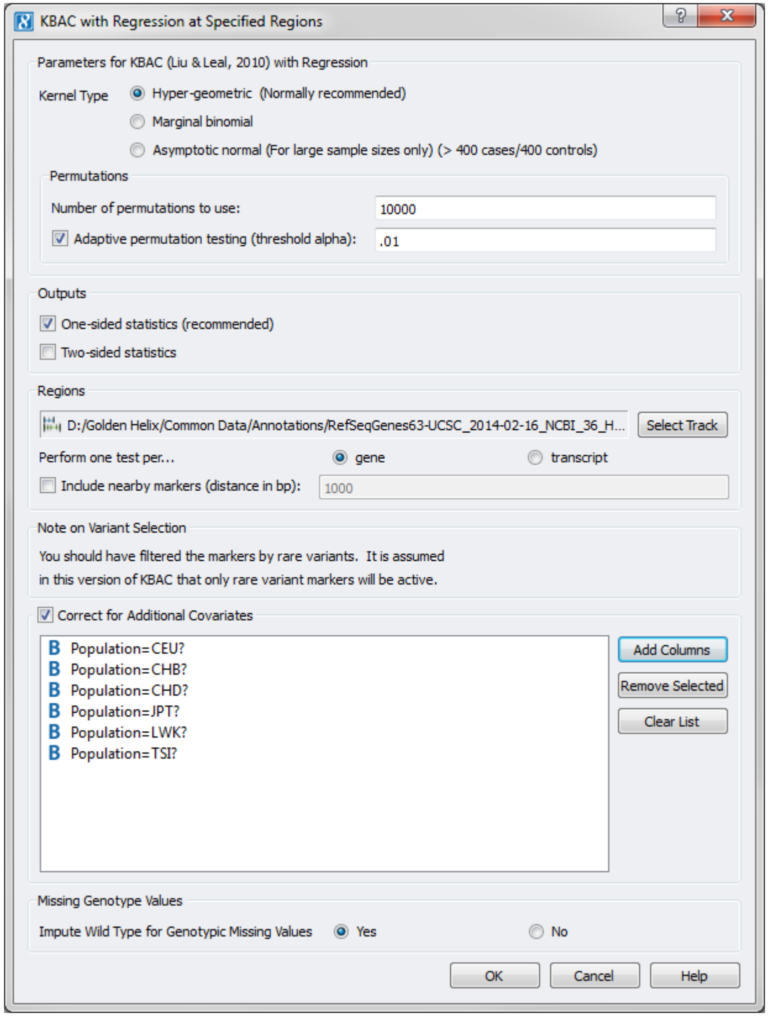
NOTE: The alternative method modifies Step 5 to use EMMA to determine the variance components and to find a new value for $\tilde{\beta}$, while leaving the value of $\tilde{u}$ at its expected value of zero.

After convergence, the alternative method predicts the values of u, and computes the final values of η and μ from this prediction

Computing the Kinship Matrix


High Quality Variants
Common Variants (MAF >= 0.05)
Rare Variants (MAF < 0.05)
IBS/IBD Kinship Matrix
MM-KBAC (Association Testing)

KBAC and MM-KBAC SVS Interface



KBAC with Regression at Specified Regions

Parameters for KBAC (Liu & Leal, 2010) with Regression

Kernel Type
- ● Hyper-geometric (Normally recommended)
- ○ Marginal binomial
- ○ Asymptotic normal (For large sample sizes only) (> 400 cases/400 controls)

Permutations

Number of permutations to use: 10000

☑ Adaptive permutation testing (threshold alpha): .01

Outputs

☑ One-sided statistics (recommended)

☐ Two-sided statistics

Regions

D:/Golden Helix/Common Data/Annotations/RefSeqGenes63-UCSC_2014-02-16_NCBI_36_H...

Perform one test per... ● gene ○ transcript

☐ Include nearby markers (distance in bp): 1000

Note on Variant Selection

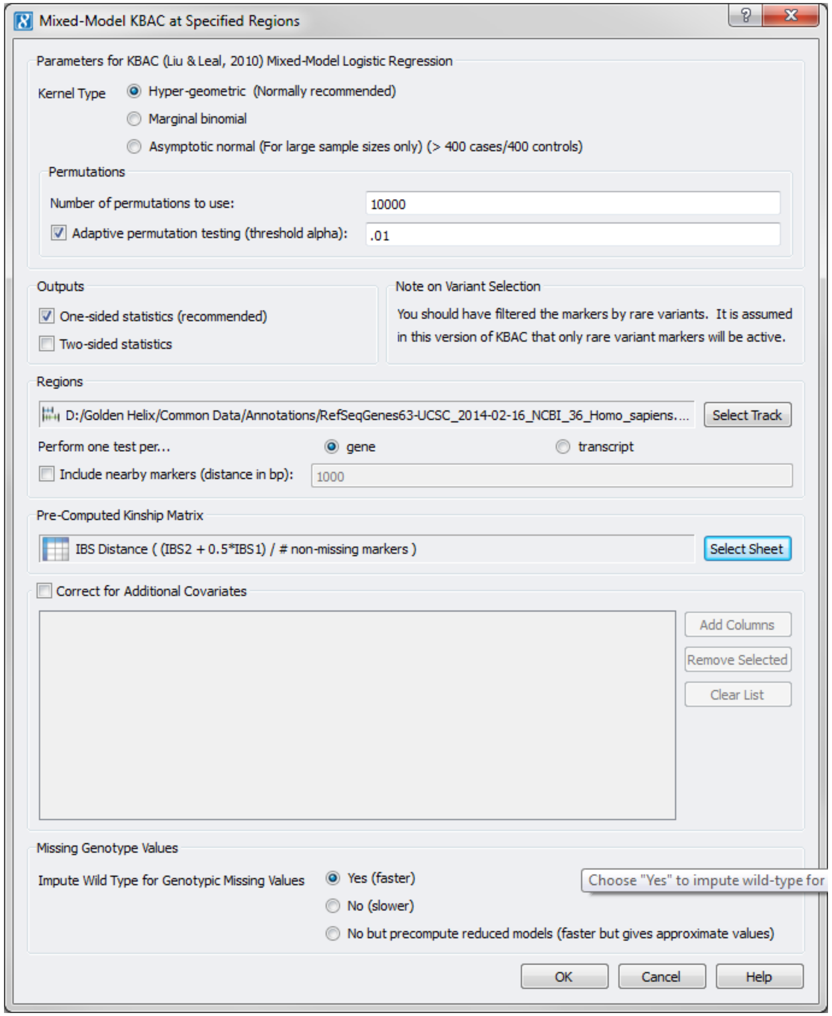
You should have filtered the markers by rare variants. It is assumed in this version of KBAC that only rare variant markers will be active.

☑ Correct for Additional Covariates

- B Population=CEU?
- B Population=CHB?
- B Population=CHD?
- B Population=JPT?
- B Population=LWK?
- B Population=TSI?

Missing Genotype Values

Impute Wild Type for Genotypic Missing Values ● Yes ○ No



Mixed-Model KBAC at Specified Regions

Parameters for KBAC (Liu & Leal, 2010) Mixed-Model Logistic Regression

Kernel Type
- ● Hyper-geometric (Normally recommended)
- ○ Marginal binomial
- ○ Asymptotic normal (For large sample sizes only) (> 400 cases/400 controls)

Permutations

Number of permutations to use: 10000

☑ Adaptive permutation testing (threshold alpha): .01

Outputs

☑ One-sided statistics (recommended)

☐ Two-sided statistics

Note on Variant Selection

You should have filtered the markers by rare variants. It is assumed in this version of KBAC that only rare variant markers will be active.

Regions

D:/Golden Helix/Common Data/Annotations/RefSeqGenes63-UCSC_2014-02-16_NCBI_36_Homo_sapiens....

Perform one test per... ● gene ○ transcript

☐ Include nearby markers (distance in bp): 1000

Pre-Computed Kinship Matrix

IBS Distance ((IBS2 + 0.5*IBS1) / # non-missing markers)

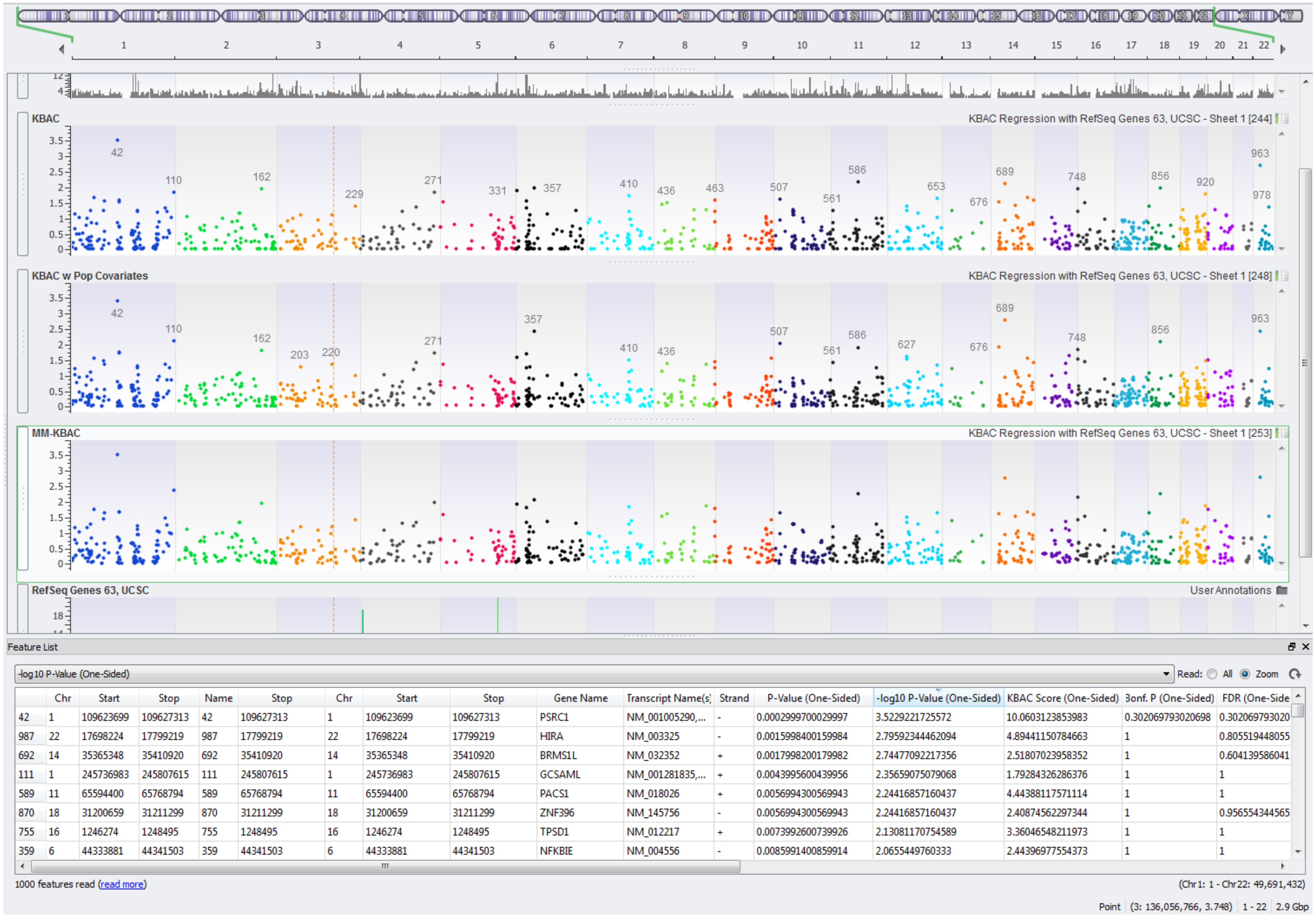
☐ Correct for Additional Covariates

Missing Genotype Values

Impute Wild Type for Genotypic Missing Values
- ● Yes (faster)
- ○ No (slower)
- ○ No but precompute reduced models (faster but gives approximate values)

Choose "Yes" to impute wild-type for

Applying MMKBAC to a real study



-log10 P-Value (One-Sided)

	Chr	Start	Stop	Name	Stop	Chr	Start	Stop	Gene Name	Transcript Name(s)	Strand	P-Value (One-Sided)	-log10 P-Value (One-Sided)	KBAC Score (One-Sided)	Bonf. P (One-Sided)	FDR (One-Side
42	1	109623699	109627313	42	109627313	1	109623699	109627313	PSRC1	NM_001005290,...	-	0.0002999700029997	3.5229221725572	10.0603123853983	0.302069793020698	0.302069793020
987	22	17698224	17799219	987	17799219	22	17698224	17799219	HIRA	NM_003325	-	0.0015998400159984	2.79592344462094	4.89441150784663	1	0.805519448055
692	14	35365348	35410920	692	35410920	14	35365348	35410920	BRMS1L	NM_032352	+	0.0017998200179982	2.74477092217356	2.51807023958352	1	0.604139586041
111	1	245736983	245807615	111	245807615	1	245736983	245807615	GCSAML	NM_001281835,...	+	0.0043995600439956	2.35659075079068	1.79284326286376	1	1
589	11	65594400	65768794	589	65768794	11	65594400	65768794	PACS1	NM_018026	+	0.0056994300569943	2.24416857160437	4.44388117571114	1	1
870	18	31200659	31211299	870	31211299	18	31200659	31211299	ZNF396	NM_145756	-	0.0056994300569943	2.24416857160437	2.40874562297344	1	0.956554344565
755	16	1246274	1248495	755	1248495	16	1246274	1248495	TPSD1	NM_012217	+	0.0073992600739926	2.13081170754589	3.36046548211973	1	1
359	6	44333881	44341503	359	44341503	6	44333881	44341503	NFKBIE	NM_004556	-	0.0085991400859914	2.0655449760333	2.44396977554373	1	1

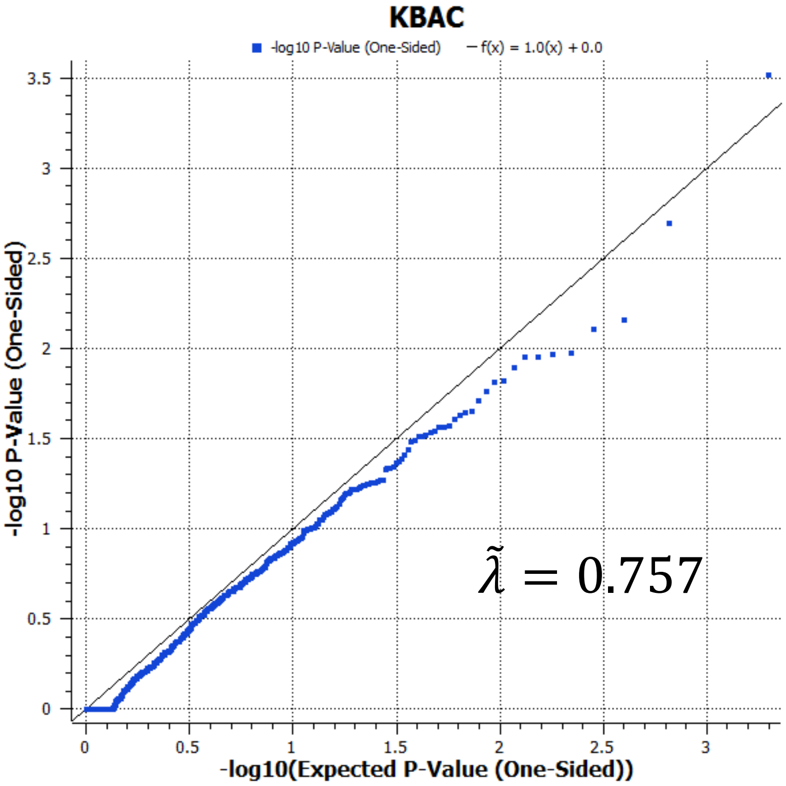
1000 features read (read more)

(Chr1: 1 - Chr22: 49,691,432)

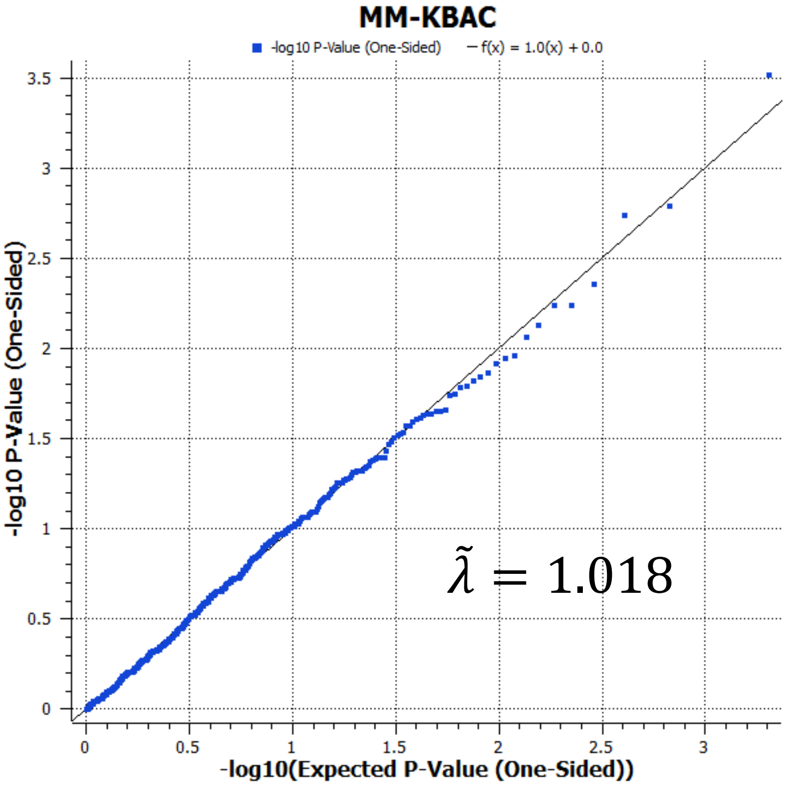
Point | (3: 136,056,766, 3.748) | 1 - 22 | 2.9 Gbp

GOLDEN HELIX
Accelerating the Quest for Significance™

KBAC vs MM-KBAC QQ Plots



$\tilde{\lambda} = 0.757$

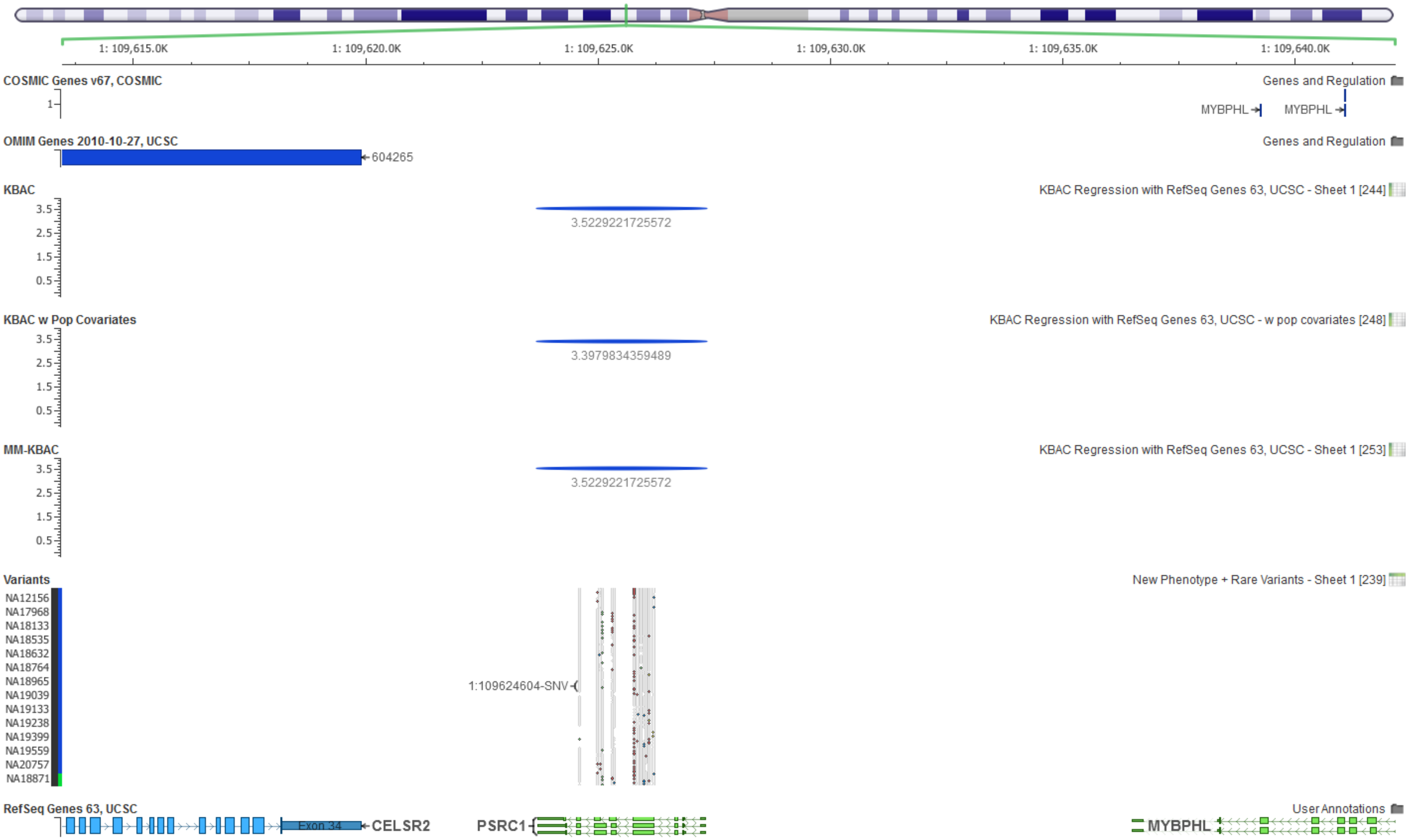


$\tilde{\lambda} = 1.018$

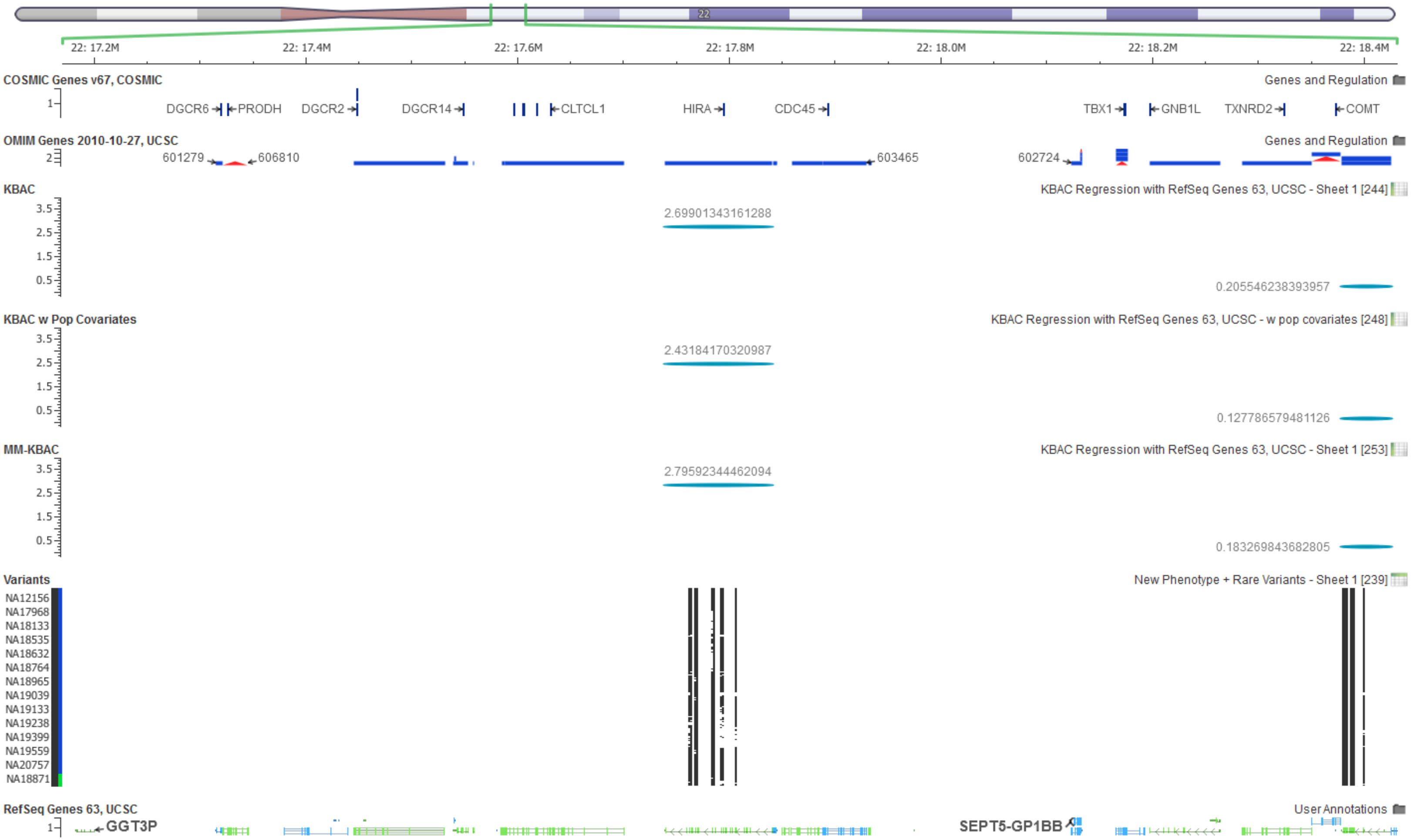
KBAC w Pop. Covariates: $\tilde{\lambda} = 0.902$

Golden Helix
Accelerating the Quest for Significance™

Signal at PSRC1



Signal at HIRA



Conclusion

- This will method will be added into SVS in the near future…

In the meantime…

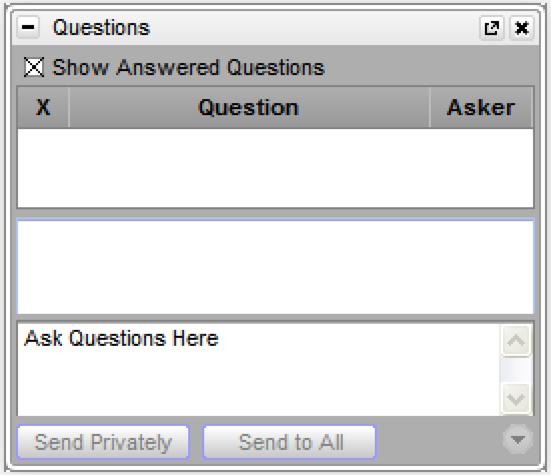
- Like to try it out on your dataset – ask us to be part of our early-access program!
- We have submitted an abstract to ASHG, hope to see you there!

Announcements

- **Webcast recording and slides will be up on our website tomorrow.**
- **T-shirt Design Contest! Details at www.goldenhelix.com/events/t-shirtcontest.html**
- **Next scheduled webcast is July 22nd, but Heather Huson of Cornell University.**

Questions?

Use the Questions pane in your GoToWebinar window



Questions

☒ Show Answered Questions

X	Question	Asker

Ask Questions Here