bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Widespread allelic heterogeneity in complex traits

Farhad Hormozdiari[1], Anthony Zhu[1], Gleb Kichaev[2], Ayellet V. Segrè[3], Chelsea J.-T. Ju[1], Jong Wha Joo[1], Hyejung Won[4], Sriram Sankararaman[1,6], Bogdan Pasaniuc[5,6], Sagiv Shifman[7,*,**], Eleazar Eskin[1,6,*,**]

[1] Department of Computer Science, University of California, Los Angeles, CA

[2] Bioinformatics IDP, University of California, Los Angeles, CA

[3] Cancer Program, The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA

[4] Neurogenetics Program, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, CA

[5] Department of Pathology and Laboratory Medicine, University of California, Los Angeles, CA

[6] Department of Human Genetics, University of California, Los Angeles, CA

[7] Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem

* Correspondence: sagiv@vms.huji.ac.il, eeskin@cs.ucla.edu

** These authors contributed equally to this work

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Recent successes in genome-wide association studies (GWAS) make it possible to address important questions about the genetic architecture of complex traits, such as allele frequency and effect size. One lesser-known aspect of complex traits is the extent of allelic heterogeneity (AH) arising from multiple causal variants at a locus. We developed a computational method to infer the probability of AH and applied it to three GWAS and four expression quantitative trait loci (eQTL) datasets. We identified a total of 4152 loci with strong evidence of AH. The proportion of all loci with identified AH is 4-23% in eQTLs, 35% in GWAS of High-Density Lipoprotein (HDL), and 23% in schizophrenia. For eQTL, we observed a strong correlation between sample size and the proportion of loci with AH (R^2=0.85, P = 2.2e-16), indicating that statistical power prevents identification of AH in other loci. Understanding the extent of AH may guide the development of new methods for fine mapping and association mapping of complex traits.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Allelic heterogeneity (AH), the presence of multiple variants at the same locus that influence a particular disease or trait, is the rule for Mendelian conditions. For example, approximately 100 independent mutations are known to exist at the cystic fibrosis locus[1], and even more independent mutations are present at loci causing inherited haemoglobinopathies[2]. In contrast, the extent of AH at loci contributing to common, complex disease is almost unknown. Indeed, current GWAS studies assume the presence of a single causal variant at a locus and report the strongest signal. Most fine mapping methods that assume a single-variant model lack the power to detect the true causal variants[3].

We developed and applied a new method to quantify the number of independent causal variants at a locus that are responsible for the observed association signals in GWAS. Our method is incorporated into the CAusal Variants Identification in Associated Regions (CAVIAR) software[3]. The method is based on the principle of jointly analyzing association signals (i.e., summary level Z-score) and LD structure in order to estimate the number of causal variants (see Methods). Our method computes the probability of having multiple independent causal variants by summing the probability of all possible sets of SNPs for being causal. We compared results from our method to results produced using the standard conditional method (CM)[13], that tests for independent association of a variant after conditioning on its significantly associated neighbors.

To evaluate the performance of our method, we first simulated datasets with different number of causal variants (see Method section for about a detailed description of our simulated datasets). Our method detected AH with high accuracy (>80%) and outperformed the standard CM approach (see **Supplementary Figures** 1 and 2). Using

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

our method, the rate of false positive (FP) was very low using different input parameters (**Supplementary Figure** 3), even when the true causal variant was not included or tagged (**Supplementary Figure** 4). We provide a more detailed description of our simulation and evaluation in the Method section.

We used seven datasets to examine the extent of AH in complex traits. We first examined quantitative trait loci contributing to variation in transcript abundance (eQTL). In the Genotype Tissue Expression (GTEx) dataset[4], we estimated the number of causal variants for genes that are known to have a significant cis eQTLs (eGene). We found that 4%-23% of the eGenes show evidence for AH (with probability > 80%) (**Figure 1**, **Supplementary Table** 1). The proportion of eGenes with AH for each tissue has a linear relationship with the sample size (R^2=0.85, P = 2.2e-16), indicating that statistical power prevents the identification of AH at other loci. To check the reproducibility of these results, we compared the GTEx blood data with results from two other blood eQTL studies: GEUVADIS[5] and Wester et al. (2013)[6]. We tested the overlap between genes with AH for skin and adipose tissues based on the GTEx and MuTHER dataset[7]. We only considered eGenes that are common between the studies. In all comparisons, we observed a high reproducibility for the detection of AH in blood (Figure 1B, P=7.9e-97), skin (Figure 1C, P=4.9e-63), and adipose (Figure 1D, P=1.1e-69) tissues.

To measure the level of AH in a human quantitative trait, we applied our method to a GWAS of High-Density Lipoprotein (HDL)[8]. Out of 37 loci, 13 (35%) showed evidence for AH with probability > 80% (see **Supplementary Table** 2). We also studied the results of GWASs focused on two psychiatric diseases, major depression[9] and schizophrenia[10]. For depression, we found evidence for AH at one of two loci. For

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

schizophrenia (SCZ), we identified 25 loci out of 108 (23%) with high probability of AH (see **Supplementary Table** 3). One example of AH in SCZ is the locus on chromosome 18 that includes the *TCF4* gene (**Figure** 2a). The locus contains multiple associated SNPs that are distributed in different LD blocks (**Figure** 2b). According to our analysis, there are three or more causal variants in this locus with high probability (**Figure** 2c) (for similar results in other loci, see **Supplementary Figures** 6-41 for HDL and **Supplementary Figures** 42-169 for SCZ).

We have shown that allelic heterogeneity is widespread and more common than previously estimated in complex traits. Since our method is influenced by statistical power and by uncertainty induced by LD, the proportions of loci with AH detected in this study are just a lower bound on the true amount of AH. Thus, our study suggest that many, if not most, loci are affected by allelic heterogeneity. Our results highlight the importance of accounting for the presence of multiple causal variants when characterizing the mechanism of genetic association in complex traits.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

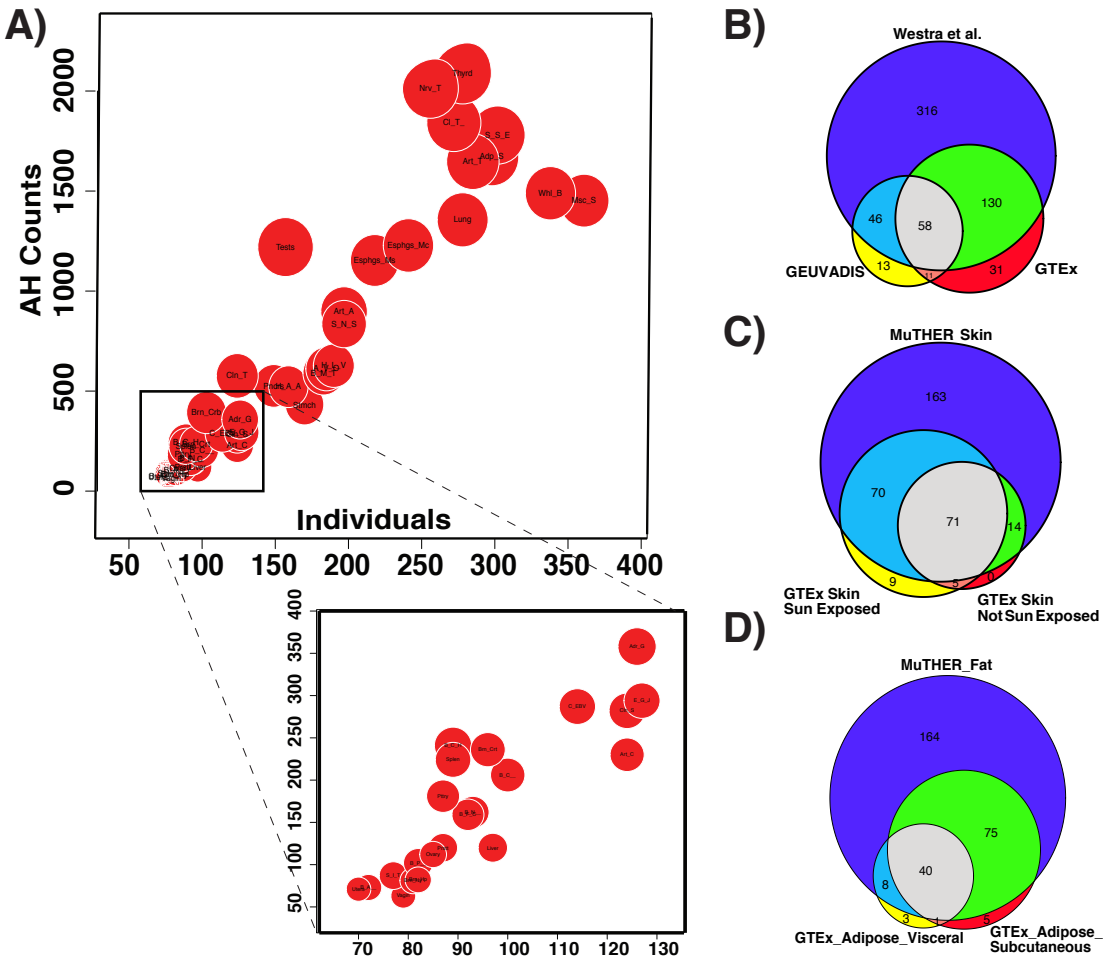


Figure1. Levels of allelic heterogeneity in eQTL studies. (A) Linear relationship between the amount of allelic heterogeneity and sample size. Each red circle indicates a different type of tissue from the GTEx dataset. The size of each red circle is proportional to the number of genes that harbor a significant eQTL (eGenes). (B-D) Significant overlap between allelic heterogeneity estimations for different eQTL datasets, shown for (B) blood (P=7.9e-97), (C) skin (P=4.9e-63), and (D) adipose (P=1.1e-69).

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

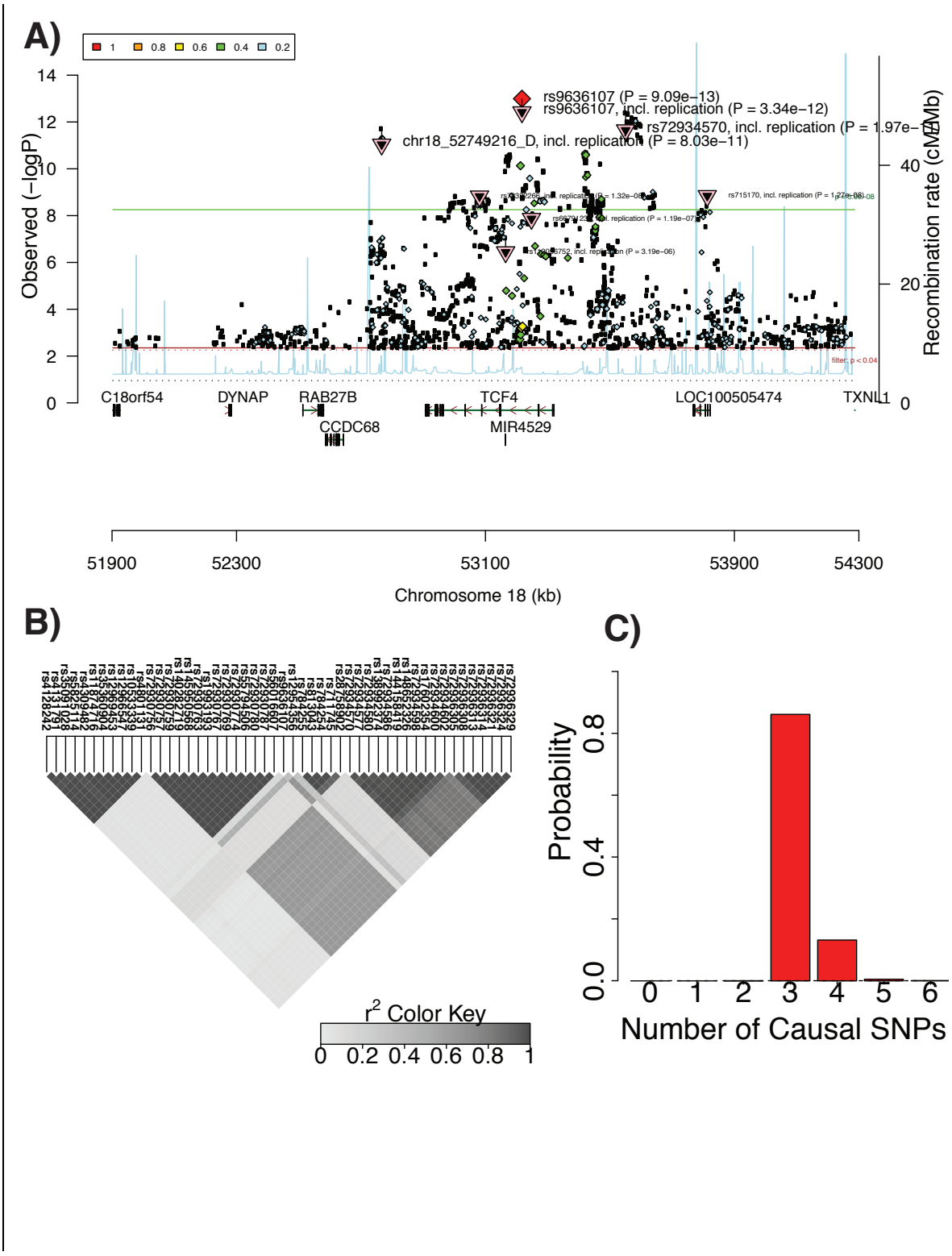


Figure2. Allelic heterogeneity in the *TCF4* locus associated with schizophrenia. (A) Manhattan plot obtained from Ricopili (http://data.broadinstitute.org/mpg/ricopili/)

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consists of all the variants in a 1Mbp window centered on the most significant SNP in the locus (rs9636107). This plot indicates multiple significant variants that are not in tight LD with the peak variant. (B) LD plot of the 50 most significant SNPs showing several distinct LD blocks. (C) Histogram of the estimated number of causal variants.

Online Methods

Overview of Methods.

The input to our method is the LD structure of the locus and the marginal statistics for each variant in the locus. The LD between each pair of variants is computed from genotyped data or is approximated from HapMap[11] or 1000G[12] data. We use the fact that the joint distribution of the marginal statistics follows a multivariate normal distribution (MVN) to compute the posterior probability of each set of variants being causal, as described below. Then, we compute the probability of having i independent causal variants in a locus by summing the probability of all possible sets of size i (sets that have i causal variants). We consider a locus to be AH when the probability of having more than one independent causal variant is more than 80%.

We would like to emphasize that using only summary statistics is not sufficient to detect AH in some cases. For example, it is impossible to detect the true number of causal variants using only the summary statistics and LD for a locus that contains several causal variants with perfect pair-wise LD. Therefore, our estimates are just a lower bound on the amount of AH for a given complex trait.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Generative Model

We utilize the fact that the vector of observed marginal statistics (e.g., the z-score) follows a multivariate normal (MVN) distribution[3] . Let $S = [s_1, s_2, s_3, \ldots s_m]^T$ indicate the observed marginal statistics for a locus with m variants. In addition, Σ is a (m × m) matrix that encodes the genotype pair-wise correlations. We denote the LD structure with Σ. The joint distribution of observed marginal statistics is as follows:

$$(S|\Lambda) \sim N(\Sigma\Lambda, \Sigma)$$

, where $\Lambda = [\lambda_1, \lambda_2, \ldots \lambda_m]^T$ is a vector of effect sizes and λ_i is the effect size of the ith variant. In this setting, λ_i is set to zero if the i-th variant is non-causal, and it is set to non-zero if the variant is causal. Let $C=[c_1, c_2, ..c_m]^T$ denote the vector of causal status. Causal status for one variant can have two possible values: zero and one. Causal status of zero indicates the variant is non-causal, and causal status of one indicates the variant is causal. Similar to the method of our previous study[3], we define the joint distribution of effect sizes following a MVN distribution:

$$(\Lambda|C) \sim N(0, \Sigma_c)$$

, where Σ_c is a (m × m) diagonal matrix. Diagonal elements of matrix Σ_c are set to zero for variants that are non-causal. Thus, using the conjugate prior, we have:

$$(S|\Lambda) \sim N(0, \Sigma\Sigma_c\Sigma + \Sigma)$$

We provide a more detailed description of the model in Supplementary Text.

Computing the Number of Independent Causal Variants

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We compute the probability of having i independent causal variants in a locus as the summation of all possible causal configurations where exactly i variants are causal. Let N_c indicate the number causal variants in a locus. We have:

$$P(N_c = i \mid S) = \frac{\sum_{|C|=i} P(S|C)P(C)}{\sum_{C \in \mathbf{C}} P(S|C)P(C)}$$

, where P(C) is the prior on the causal configuration C, $\boldsymbol{C}$ is the set of all possible causal configurations – including the configuration all the variants are not causal – and |C| indicates the number of causal variants in the causal configuration C. The numerator in the above equation considers all possible causal configurations that have i causal variants. The denominator is a normalization factor to ensure that the probability definition holds. We define the prior probability as $P(C) = \prod \gamma^{|C_i|} (1-\gamma)^{1-|C_i|}$. In our experiment, we set γ to 0.001 (see Supplementary Text).

Although computing the probability of each causal configuration requires $O(m^3)$ operations, we utilize the matrix structure of the marginal statistics to reduce this computation (see Supplementary Text). Our reduction is in a factor of m/k, where k is the number of causal variants for each causal configuration. Thus, in our method, we require $O(m^2k)$ operations to compute the likelihood of each causal configuration.

Conditional Method (CM)

A standard method to detect allelic heterogeneity (AH) is the conditional method (CM). In CM, we identify the SNP with most significant association statistics. Then, conditioning on that SNP, we re-compute the marginal statistics of all the remaining variants in the locus. We consider a locus to have AH when the re-computed marginal statistics for at least one of the variants is more

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significant than a predefined threshold. Similarly, we consider a locus to not have AH when the re-computed marginal statistics of all variants fall below the predefined threshold. The predefined threshold is referred to as the stopping threshold for CM. This standard method can be applied to either summary statistics or individual level data[13]. GCTA-COJO[13] performs conditional analysis while utilizing the summary statistics.

When applying CM to individual level data, we re-compute the marginal statistics by performing linear regression where we add the set of variants that are selected as covariates.

We utilize the LD between the variants, which we obtain from a reference dataset, when applying CM to summary statistics data. In this case, we re-compute the marginal statistics for the ith variant as follows:

$$Z_i^{new} = \frac{z_i - z_j r_{ij}}{\sqrt{1 - r_{ij}^2}}$$

when we have selected the jth variant as causal. Let z_i indicate the marginal statistics for the ith variant and r_{ij} the genotype correlations between the ith and jth variants.

Datasets

Genotype-Tissue Expression (GTEx): We obtained the summary statistics for GTEx[4] eQTL dataset (Release v6, dbGaP Accession phs000424.v6.p1) at http://www.gtexportal.org. We estimated the LD structure using the available genotypes in the GTEx dataset. We considered 44 tissues and applied our method to all eGenes, genes that have at least one significant eQTL, in order to detect loci that harbor allelic heterogeneity..

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genetic European Variation in Disease (GEUVADIS): We obtained the summary statistics of blood eQTL for 373 European individuals from the GEUVADIS[5] website (ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/). We approximated LD structure from the 1000G CEU population. We applied our method to the 2954 eGenes in GEUVADIS to detect AH loci.

Multiple Tissue Human Expression Resource (MuTHER): We obtained the summary statistics from MuTHER[6] website (http://www.muther.ac.uk/Data.html). We utilized the skin and fat (adipose) tissues. We then approximated LD from the 1000G CEU population. We obtained 1433 eGenes for skin and 2769 eGenes for adipose.

High-Density Lipoprotein Cholesterol (HDL-C): We used the High-Density Lipoprotein Cholesterol (HDL-C) trait[8]. We only considered the GWAS hits, which are reported in a previous study[8]. We applied ImpG-Summary[14] to impute the summary statistics with 1000G as the reference panel. We identified 37 loci that have at least one causal variant. Following common protocol in fine-mapping methods, we assumed at least one causal variant.. Then, we applied our method to each locus.

Psychiatric diseases: We analyzed the recent GWAS on major depression[9] and schizophrenia[10]. The major depression study has 2 and the schizophrenia study has 108 loci identified to contain at least one significant variant. We utilized the summary statistics provided by each study and approximated the LD using the 1000G CEU population.

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data simulation

We first simulated genotypes using HAPGEN2[15], where we utilized the 1000G CEU population as initial reference panels. Then, we simulated phenotypes using the Fisher's polygenic model, where the effects of causal variants are obtained from the normal distribution with a mean of zero. We let Y indicate the phenotypes and X indicate the normalized genotypes. In addition, β is the vector of effect sizes where $\boldsymbol{\beta}_i$ is the effect of the i-th variant. Thus, we have:

$$Y = X\,\beta + \mathrm{e}$$

, where e models the environment and measurement noise. Under the Fisher's polygenic model, the effect size of the causal variants is obtained from $N(0, \frac{\sigma_g^2}{N_c})$, where N_c is the number of causal variants and σ_g is the genetic variation. In addition, the effect size for variants that are non-causal is zero. We set the effect size in order to obtain the desired statistical power. We implanted one, two, or three causal variants in our simulated datasets.

We use false positive (FP) and true positive (TP) as metrics to compare different methods. FP indicates the fraction of loci that harbor one causal variant and are incorrectly detected as loci that harbor AH. TP indicates the fraction of loci that harbor AH and are correctly detected.

Code availability. Our method is incorporated into CAVIAR software that is available at http://genetics.cs.ucla.edu/caviar

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1 Estivill, X. *et al*. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Human mutation* **10**, 135–154 (1997).

2 Hardison, RC. *et al.* HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Human mutation* **19.3**, 225-233 (2002).

3 Hormozdiari, F. *et al*. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

4 GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**(6235), 648–660, doi: 10.1126/science.1262110 (2015).

5 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional. *Nature* 501(7468),506-11. doi: 10.1038/nature12531 (2013).

6 Westra, H. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**,1238–1243, doi:10.1038/ng.2756 (*2013*).

7 Nica, A. *et al.* MuTHER Consortium the architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLos Genetics 7(2)*, http://dx.doi.org/10.1371/journal.pgen.1002003 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8 Teslovich, T. M. *et al*. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

9 CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**(7562):588–591 (2015).

10 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

11 The International HapMap Consortium. The International HapMap Project. *Nature* **18**, 789–796 (2003).

12 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **467**, 1061–1073 (2010).

13 Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **18**:44(4):369–75 (2012).

14 Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30** pp. 2906–2914 (2014).

15 Su, Z., *et al.* Hapgen2: Simulation of multiple disease snps. *Bioinformatics* **27**, 2304–5 (2011).

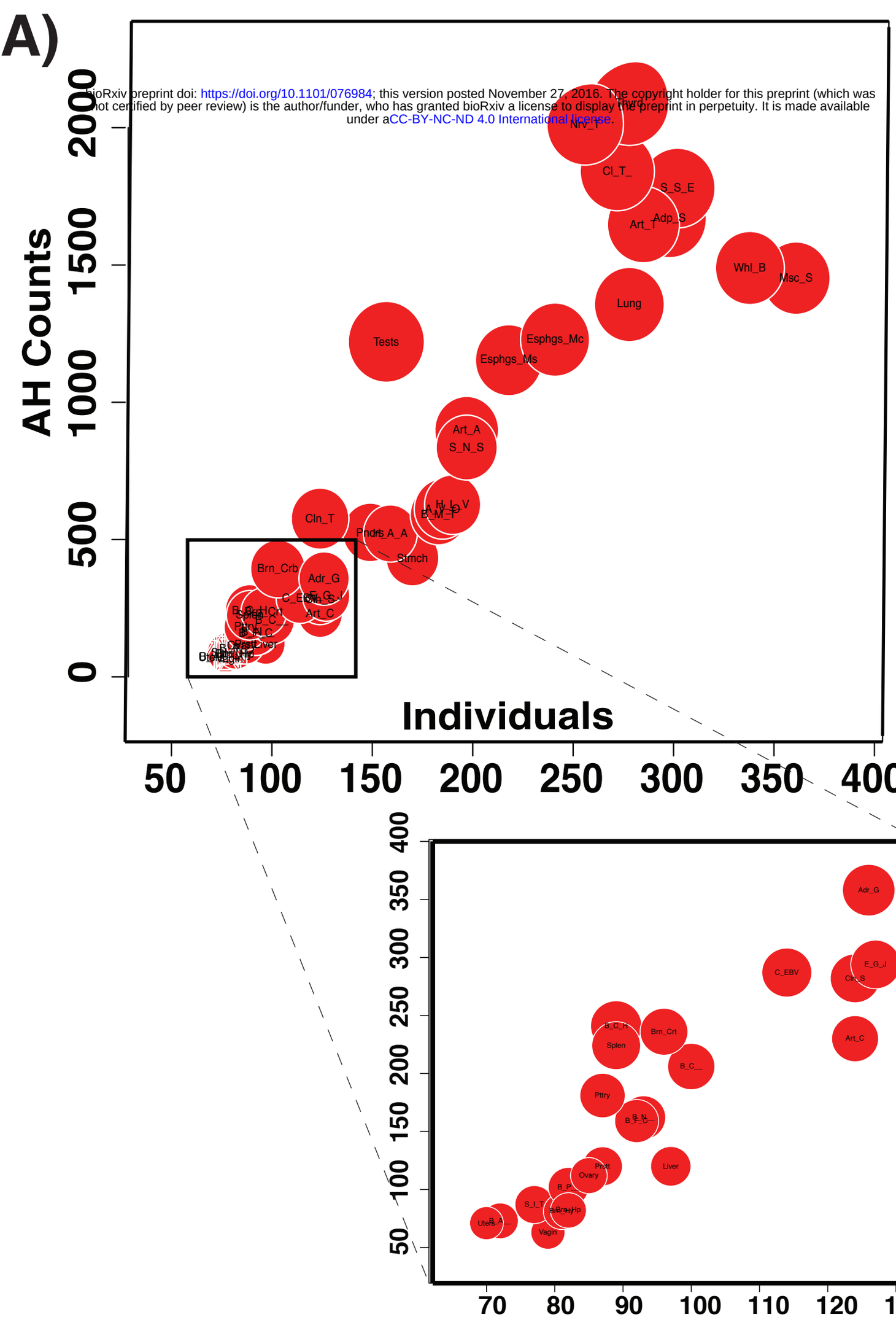
bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

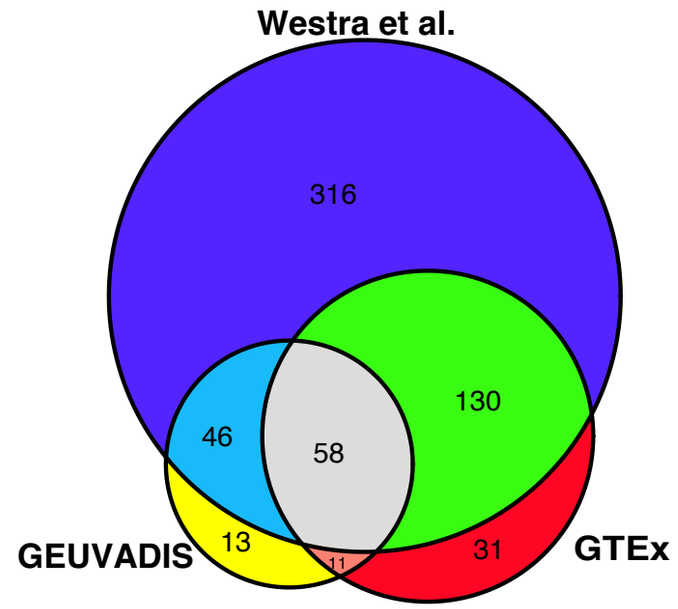
FH, JWJJ, and EE are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589 and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. EE is supported in part by the NIH BD2K award, U54EB020403. AVS is supported by a contract (HHSN268201000029C) to the Laboratory, Data Analysis, and Coordinating Center (LDACC) at The Broad Institute, Inc. SS was supported in part by NIH grant R00-GM 111744-03. GK is supported by the Biomedical Big Data Training Program (NIH-NCI T32CA201160). We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

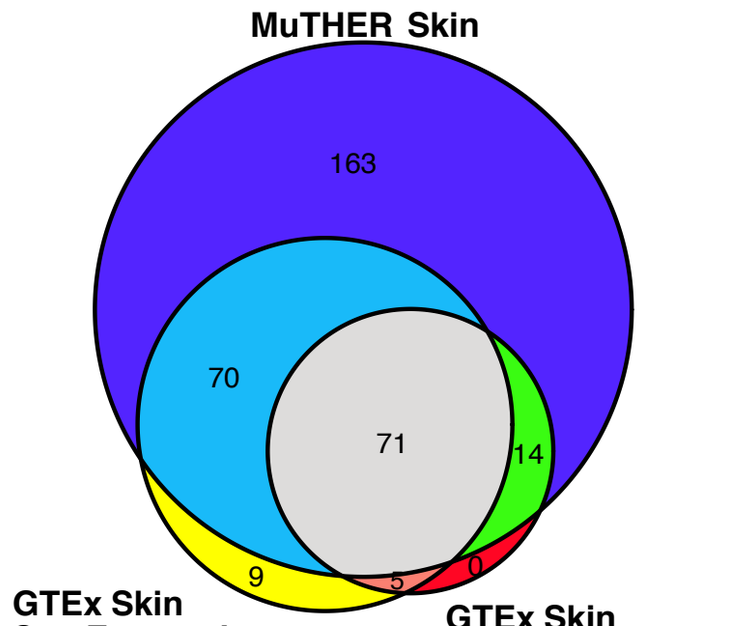
A)



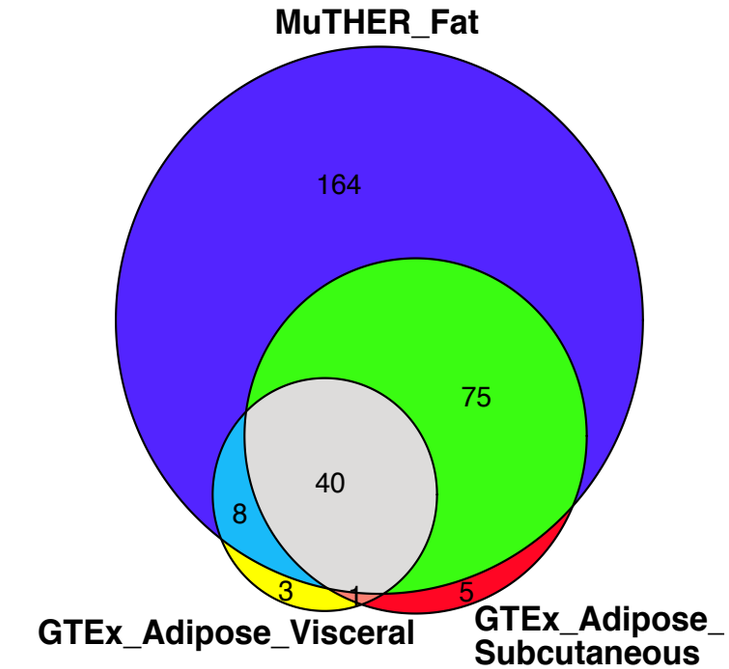
B)



C)



D)



bioRxiv preprint doi: https://doi.org/10.1101/076984; this version posted November 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.