Fine-mapping gene-based associations via knockoff analysis of biobank-scale data with applications to UK Biobank

Shiyang Ma[1], Chen Wang[1], Atlas Khan[2], Linxi Liu[3], James Dalgleish[1], Krzysztof Kiryluk[2], Zihuai He[4,5], Iuliana Ionita-Laza[1,#]

[1] Department of Biostatistics, Columbia University, New York, NY
[2] Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY
[3] Department of Statistics, University of Pittsburgh, Pittsburgh, PA
[4] Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA
[5] Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA
[#]Correspondance: ii2135@columbia.edu

Abstract

We propose BIGKnock (BIobank-scale Gene-based association test via Knockoffs), a gene-based testing approach that leverages long-range chromatin interaction data, is applicable to biobank-scale data, and performs conditional testing genome-wide via knockoffs. Thereby BIGKnock reduces the confounding effect of linkage disequilibrium relative to existing gene-based tests and can prioritize causal genes over proxy associations at a locus. We applied BIGKnock to the UK Biobank data with 405,296 British subjects for multiple binary and quantitative traits, and show that relative to conventional gene-based tests BIGKnock produces smaller sets of significant genes that contain the causal gene(s) with high probability (> 90%). We further illustrate its ability to pinpoint potentially causal genes at ∼ 80% of the associated loci (4,829 loci across 24 diseases and traits), including genes with well established causal links in the literature such as *ASGR1* and *ANGPTL4* and cholesterol, and *ALDH2* and coronary artery disease, as well as plausible novel links such as *NGFR* and asthma, *AGPAT1* and type 2 diabetes, *DBH* and blood pressure, *ZHX3* and calcium, *PPARγ* and LDL cholesterol. Relative to several methods for causal gene prioritization such as closest gene, cS2G and L2G, we show that BIGKnock produces more interpretable results and improves precision on two sets of gold standard causal genes. Finally, we show that the prioritized genes have several interesting properties relative to non-significant genes that are consistent with them being putative causal genes.

Introduction

Gene-based tests that incorporate regulatory variation from proximal and distal regulatory elements are appealing given that most genetic variants associated with complex traits reside in non-coding regions. Unlike single variant testing which requires follow-up investigations to

identify the causal gene(s), gene-based testing that incorporates putative regulatory elements provides a unified test at the gene level. Transcriptome-wide association tests (TWAS) are typical examples of gene-based tests that leverage expression quantitative trait loci (eQTL) data from reference datasets such as GTEx[1]. However, a main challenge is the high false positive rate for such tests caused by confounding due to linkage disequilibrium (LD) and co-regulation. Although fine-mapping approaches have been proposed for TWAS[2], these approaches are limited to eQTLs being present in the reference datasets, and the majority of genetic associations cannot be clearly assigned to existing eQTLs[3,4].

In our previous work[5] we have proposed a new gene-based test that incorporates genetic variation in proximal and distal regulatory elements (not restricted to eQTLs) and which performs genome-wide conditional tests via knockoffs in order to reduce the confounding effect of LD. In this paper we propose a scalable implementation that improves the computational efficiency while maintaining the statistical performance (i.e., FDR control and power) of the knockoff framework, making it applicable to biobank sized datasets, and demonstrate its ability to prioritize likely causal genes for several binary and continuous traits in the UK biobank data. Relative to recent causal gene prioritization methods such as combined SNP-to-gene (cS2G)[6] and Locus-to-gene (L2G)[7] which are based on supervised machine learning methods to integrate various functional features predictive of the causal gene(s) at a locus, and which are therefore dependent on good quality training data and high quality fine-mapping results, our gene-based test avoids these limitations, produces more interpretable results (in terms of q-values and FDR control) and naturally restricts false positives due to LD confounding.

Biobanks with comprehensive genetic and phenotypic data from electronic medical records provide a powerful resource for genomic studies. For example, the UK biobank is comprised of genotype and phenotype data on about 500,000 individuals and millions of genetic variants[11]. The size of such data poses challenges in terms of computational time and memory requirements for conventional linear mixed models, and recent methods have been proposed to make such models scalable to biobank sized datasets[12,13]. Furthermore, the proposed gene-based test is based on knockoff inference, a statistical framework for variable selection in high-dimensional settings[14]. The idea behind knockoff-based inference is to generate synthetic, noisy copies (knockoffs) of the original genetic variants that resemble the true variants in terms of preserving correlations but are conditionally independent of the phenotype given the true genetic variants. The knockoffs serve as negative controls and help select significant genetic risk loci while controlling the false discovery rate (FDR). Constructing multiple knockoff genotype features is time consuming and recently efficient methods for knockoff generation have been proposed[15].

In this paper we leverage these methodological improvements, and propose a computationally and memory efficient gene-based test via knockoffs for biobank sized data, BIGKnock. Its ability to prioritize causal genes, e.g. those for which regulatory changes mediate genome-wide association signals, through genome-wide conditional testing is a unique feature for our test and we demonstrate its ability to prioritize putative causal genes at $\sim$ 80% loci for four binary and twenty quantitative traits in the UK biobank. We illustrate with several examples of well known causal genes along with new plausible ones that BIGKnock is able to identify. We also show that the prioritized genes have interesting properties relative to non-significant genes that are consistent with them being putative causal genes.

Results

Overview of BIGKnock

We provide here a brief overview of the proposed gene-based test, BIGKnock, and the conventional test GeneScan3D which we compare against. GeneScan3D incorporates classical Burden and SKAT tests to test for association between genetic variation in the gene body (i.e., the interval between the transcription start site and the end of 3' UTR) and in proximal/distal regulatory elements, and a trait of interest. GeneScan3D scans the gene body region and the associated regulatory elements with varied window sizes and combines results using the Cauchy combination method[16] to produce one single p-value per gene. BIGKnock extends GeneScan3D by implementing the knockoff framework. BIGKnock computes for each gene a knockoff statistic W that measures the importance of each gene (similar to a p-value), and then uses the knockoff filter to detect genes with sufficiently large W, i.e. those genes significant at a specified FDR target level[14]. We also compute a q-value for each gene. A q-value is similar to a p-value, except that it measures significance in terms of FDR rather than FWER, and already incorporates correction for multiple testing. The details on these specific tests can be found in the Methods section.

Applications to UK Biobank: Binary Traits

We applied BIGKnock to four binary traits in the UK Biobank, including Hypertension (Phecode 401), Coronary Artery Disease (Phecode 411), Asthma (Phecode 495) and Type 2 diabetes (Phecode 250.2) (See Table S1 for sample size information). Note that we have previously[5] compared the performance of the original knockoff-based test, and GeneScan3D with other commonly-used tests including STAAR-O[8] and MAGMA/H-MAGMA[9,10], and have shown improved power and FDR control relative to these existing methods. Since BIGKnock is a scalable implementation of the previous test, we only compare with GeneScan3D (see Methods and[5]), the conventional 3D gene-based test without knockoff-based inference, to illustrate the advantages of the knockoff-based testing approach. We use a Bonferroni adjusted threshold of 2.5×10^{-6} for GeneScan3D and an FDR threshold of 0.01 or 0.05 (depending on the size of the study) for BIGKnock. For four binary traits we consider here, we identify 1,209 gene-trait associations for GeneScan3D and 801 associations for BIGKnock (Supplemental Tables 1-4). Among the 1,209 significant associations under GeneScan3D, only 688 (57%) are significant under BIGKnock, despite the more liberal FDR threshold used by BIGKnock owing to LD adjustement.

We use the significant GWAS SNPs ($p < 5 \times 10^{-8}$) to define 1Mb loci centered at the most significant SNP. For each gene-based test (GeneScan3D and BIGKnock), we count the number of loci that contain at least one significant gene for each test respectively. In terms of the number of significant loci, GeneScan3D and BIGKnock show similar results, with most of the significant loci shared between GeneScan3D and BIGKnock (Table S2). However, one of our main interests in employing the knockoff framework is to filter out false positive genes that appear in the conventional GeneScan3D test. We therefore consider shared loci that contain at least one significant gene for both GeneScan3D and BIGKnock, and compare the number of significant genes identified by the two methods at such loci. The knockoff test discovers a smaller number of significant genes than GeneScan3D despite the more liberal FDR threshold (Figure 1). We provide further evidence below that BIGKnock, by conditioning on nearby variants, can prioritize genes more likely to be causal.

BIGKnock can prioritize putative causal genes at significant loci. We demonstrate that significant genes detected by BIGKnock tend to be enriched among genes nearest to the lead GWAS SNP at significant loci, the class of genes most likely to be the causal genes[17,18]. We first perform the enrichment analysis (Methods) based on 136 BIGKnock significant loci for multiple binary traits. Knockoff significant genes are 4.6-fold (range 2.3-8.8 for four binary traits) more likely to be the nearest gene relative to the rest of the genes at a locus (Figure 2a). Similar results hold when we restrict the analyses to 127 loci shared between BIGKnock and GeneScan3D (5.1-fold with range 2.3-10.7, Figure S1a).

Next, we focus on several loci where the knockoff-based test can prioritize only a few genes at a locus relative to the conventional GeneScan3D test (Table 1), and there is compelling literature support for a mechanistic role of the selected genes in the pathogenesis of the corresponding traits.

***ALDH2* (Aldehyde Dehydrogenase 2) and coronary artery disease.** We illustrate first in detail the association between *ALDH2* and coronary artery disease. Although GeneScan3D identifies 11 significant genes at this locus, BIGKnock identifies only two of them as significant including *ALDH2* and *BRAP* (Figure 3(a)). The additional associations detected by the conventional GeneScan3D test are likely due to LD between variants in those genes and putative causal variants in the *BRAP-ALDH2* neighborhood. *ALDH2* is expressed across many tissues in GTEx but is most abundant in the liver and adipose tissues (Figure 3(c)). The role of *ALDH2* in cardiovascular disease is well-documented in the literature[19]. The *ALDH2* Glu504lys polymorphism is widely considered as a risk factor for the development of coronary artery disease, especially in Asian populations[20–22]. Furthermore, mitochondrial ALDH2 has emerged as a key enzyme for removal of ethanol-derived acetaldehyde, and has been shown to play a role in inflammation regulation and macrophages accumulation[23]. Epidemiological studies in humans carrying an inactivating mutation in *ALDH2*, combined with genetic and pharmacological studies in animal models, have implicated *ALDH2* in the development and prognosis of coronary heart disease, hypertension, type 2 diabetes, and stroke, and suggest *ALDH2* as an important target for generating new treatments for heart diseases[24].

Additional loci with strong literature support. *NGFR* (nerve growth factor receptor) and asthma (Figure 4a): Nerve growth factor has been implicated in both the immune and neuronal components of allergic asthma pathogenesis. Furthermore, the nerve growth factor (*NGF*) targeting treatment may be an important therapy for antigen-induced airway hyper responsiveness via attenuation of airway innervation and inflammation in asthma[25].

AGPAT1 (1-acylglycerol-3-phosphate O-acyltransferase 1) and type 2 diabetes (Figure 4b): *AGPAT1* is a metabolism (lipid biosynthesis) gene and plays important functions in the physiology of multiple organ systems. In particular, Agpat1-deficient mouse developed widespread disturbances of metabolism including low body weight and low plasma glucose levels[26]. Furthermore, Agpat1 mouse knockout has low circulating glucose and increased urine glucose and urine microalbumin (International Mouse Phenotyping Consortium).

MARCHF5 (membrane-associated RING-CH-type finger 5) type 2 diabetes (Figure 4c): *MARCHF5* is a PPARγ target gene that influences mitochondrial and cellular metabolism in adipocytes[27]. These functions likely alter the utilization of lipid, which subsequently impacts glucose metabolism.

Effector BIGKnock Genes. We further restrict the list of BIGKnock significant genes by identifying those that coincide with the closest gene (among all genes) to the top significant GWAS SNP at a locus. Among 136 significant loci across four binary traits, we identify 91 (67%) such loci (Supplemental Table 5). We call these genes effector BIGKnock genes. For loci that do not have effector BIGKnock genes, 22 loci have only one BIGKnock significant gene. Therefore, we prioritize one potentially causal gene at 83% of the loci.

Mouse phenotype enrichment analyses. Using ToppFun[28] we have tested whether the effector BIGKnock genes are enriched in sets of genes associated with mouse phenotypes. The mouse phenotype data are extracted from the Mammalian Phenotype Ontology, and consists of mouse genes that cause phenotypes in genetically engineered or mutagenesis experiments. Effector BIGKnock genes are enriched in gene sets corresponding to relevant mouse phenotypes (Figure S2). For example, among the most significantly enriched phenotypes were abnormal circulating insulin levels, and increased circulating glucose levels for Type 2 diabetes, abnormal systemic arterial blood pressure for hypertension, abnormal CD4-positive, alpha-beta T cell physiology and abnormal T-helper 2 physiology for asthma, and increased susceptibility to atherosclerosis and abnormal hepatobiliary system physiology for coronary artery disease.

Applications to UK Biobank: Quantitative Traits

We have also applied BIGKnock to 20 quantitative traits in the UK Biobank, including estimated glomerular filtration rate (eGFR), Body Mass Index (BMI), Diastolic Blood Pressure Automated Reading (BP-Diastolic), Systolic Blood Pressure Automated Reading (BP-Systolic), Cystatin C, Platelet count, Mean platelet volume (MPV), Apolipoprotein A, HDL cholesterol, Cholesterol, Glycated haemoglobin (HbA1c), Mean reticulocyte volume (MRV), Mean sphered cell volume (MSCV), Red blood cell (erythrocyte) distribution width (RDW), Neutrophil count, Reticulocyte count, Calcium, insulin-like growth hormone factor-1 (IGF-1), LDL direct (LDL cholesterol) and Direct bilirubin (samples sizes for individual traits are in Table S1). For quantitative traits we use more stringent FDR thresholds (0.001 or 0.005) due to large sample sizes and consequently large number of significant findings. For these 20 quantitative traits, we identify 57,043 gene-trait associations for GeneScan3D and 37,391 associations for BIGKnock (Supplemental Tables 6-25). Among 57,043 associations significant under GeneScan3D, only 36,086 (63%) are significant under BIGKnock, due to LD adjustement.

We report the number of significant loci/genes per trait in Table S3. As with the binary traits, for most of the significant shared loci, BIGKnock can reduce the number of significant associations despite the more liberal (FDR) thresholds being used (Figures S3-S6).

BIGKnock can prioritize putative causal genes at significant loci. As with binary traits, we demonstrate that significant genes detected by BIGKnock tend to be enriched among genes nearest to the lead GWAS SNP at significant loci. We first perform the enrichment analysis (Methods) on 6,195 BIGKnock significant loci for multiple quantitative traits. In particular, knockoff significant genes are 2-fold (range 1.6-2.9 for individual traits) more likely to be the nearest gene relative to the rest of the genes at a locus (Figure 2b). When we restrict the analyses to 6,001 loci shared between BIGKnock and GeneScan3D, similar enrichment can be observed (Figure S1b).

Next, we focus on several loci where the knockoff-based test can prioritize few genes at a

locus relative to GeneScan3D (Table 1), and there is compelling literature support for a mechanistic role of the selected genes in the pathogenesis of the corresponding traits.

***ASGR1* (asialoglycoprotein receptor 1) and cholesterol.** We illustrate first in detail the example of *ASGR1* and cholesterol. At the 1 Mb locus containing *ASGR1*, BIGKnock prioritizes two genes including *ASGR1* among 43 genes significant using the conventional GeneScan3D test (Figure 3(b)). Most of the GeneScan3D associations are due to gene-enhancer links for two enhancers (Figure S7). Specifically, 18 associations are due to variants in an enhancer GH17F007167 just upstream of gene *ASGR1*, and when accounting for LD with nearby variants, BIGKnock no longer detects them as significant. Furthermore, additional associations that are removed by BIGKnock are 12 genes linked to ABC enhancer chr17:7,144,929-7,146,587 (hg19) downstream of gene *ASGR1*, and 6 genes linked to 4 other enhancers (Supplemental Table 26). Therefore, at this locus, BIGKnock is able to prioritize two genes by adjusting for linkage disequilibrium in the region. *ASGR1* is also highly expressed in liver (Figure 3(d)). The role of *ASGR1* in the control of non-HDL cholesterol levels and in regulation of the endogenous levels of at least some asialoglycoproteins has been established[29]. Specifically, Nioi et al.[29] have identified rare loss-of-function variants in *ASGR1* that are associated with lowering of non-HDL cholesterol levels and a reduced risk of coronary artery disease. Recent mechanistic studies also support a role of *ASGR1* in cholesterol. For example, ASGR1-deficient pigs show lower levels of non-HDL cholesterol and less atherosclerotic lesions than that of controls, therefore targeting ASGR1 might be an effective strategy to reduce hypercholesterolemia and atherosclerosis[33].

Additional loci with strong literature support. *SLC39A8* (solute carrier family 39 member 8) and diastolic blood pressure (Figure 4d): Slc39a8 deletion in mice results in increased nitric oxide (NO) production, decreased blood pressure, and protection against high-salt-induced hypertension, while homozygosity of the *SLC39A8* loss-of-function variant in humans is associated with increased NO, providing a plausible explanation for the association of *SLC39A8* with blood pressure[34,35].

DBH (dopamine beta-hydroxylase) and diastolic blood pressure (Figure 4e): Dbh(-/-) mice had a low heart rate, were severely hypotensive, and displayed an attenuated circadian blood pressure rhythm[36].

ANGPTL4 (angiopoietin-like protein 4) and cholesterol (Figure 4f): *ANGPTL4* was uncovered as a novel modulator of plasma lipoprotein metabolism. In 24-hour fasted mice, Angptl4 overexpression increased plasma triglycerides (TG) by 24-fold, which was attributable to elevated VLDL-, IDL/LDL- and HDL-TG content[37].

RAB11A (ras-related protein Rab-11A) and neutrophil counts (Figure 4g): In mice challenged with endotoxin, intratracheal instillation of Rab11a-depleted macrophages reduced neutrophil count in bronchoalveolar lavage fluid, increased the number of macrophages containing apoptotic neutrophils, and prevented inflammatory lung injury[38].

ZHX3 (zinc fingers and homeoboxes 3) and calcium (Figure 4h): Zhx3-KO mice have increased bone mineral density (International Mouse Phenotyping Consortium), and ZHX3 may be useful as an early osteogenic differentiation marker[39].

PPARγ (peroxisome proliferator- activated receptor gamma) and LDL cholesterol (Figure 4i): PPARγ regulates fatty acid storage and glucose metabolism. The genes activated by PPARγ stimulate lipid uptake and adipogenesis by fat cells. PPARγ plays a regulatory role in the first steps of the reverse-cholesterol-transport pathway through the activation of ABCA1-mediated cholesterol efflux in human macrophages.[40]

POLDIP2 (polymerase delta-interacting protein 2) and LDL cholesterol (Figure 4j): Poldip2 was shown to increase Nox4 enzymatic activity by 3-fold and to positively regulates basal reactive oxygen species production in vascular smooth muscle cells[41]. The authors suggest that Poldip2 may be a novel therapeutic target for vascular pathologies with a significant vascular smooth muscle cell migratory component, such as restenosis and atherosclerosis.

BIGKnock can prioritize putative effector genes in Backman et al.[30]. We use data on putative effector genes identified in a recent study by Backman et al.[30] using rare-variant exome-wide association studies in 454,787 participants in the UK Biobank study. Specifically, Backman et al. first identify common variants independently associated with each trait (i.e., GWAS sentinel variant), which are then included as additional covariates for Burden association tests with rare variants focusing on pLOF (including stop-gain, frameshift, stop-loss, start-loss and essential splice variants) and deleterious missense variants with a minor allele frequency (MAF) of up to 1%. Overall, 168 significant genes adjusting for GWAS signals (with Burden p-values $\leq 2.18 \times 10^{-11}$) and that are nearest to the GWAS sentinel variant are defined as the likely effector genes[30]. Here we consider the 120 effector gene-trait associations corresponding to the quantitative traits considered in our analyses (Supplemental Table 27). We identify 116 effector gene associations that are significant under GeneScan3D with 106 (91%) also significant under BIGKnock. Note that this is a significantly higher retention rate for effector gene associations than the expected rate (63%, see above) based on all genes significant under GeneScan3D (two-sided p= 6.4×10^{-10}), and supports the claim that BIGKnock retains the truly causal genes but removes many of the false associations due to LD. Several examples include *ASGR1* and *SH2B3* and cholesterol, and *APOB* and Apolipoprotein A. *ANGPTL4* was also prioritized by BIGKnock for cholesterol, and identified as effector gene for HDL cholesterol (Table S4 and Figure S8).

In addition, Backman et al.[30] identified 564 genes associated with traits using rare variant association tests focusing as above on pLOF and deleterious missense variants with a MAF of up to 1%. Among 134 genes that correspond to the quantitative traits considered in our analyses (Supplemental Table 28), we identify 111 GeneScan3D significant genes with 99 (89%) being significant under BIGKnock. Again, this is a significantly higher proportion than expected based on all GeneScan3D associations (two-sided p=2.6×10^{-8}). Several example include *DBH* associated with BP-Diastolic, gene *SLC5A3* associated with Cystatin C and gene *POLE* associated with MRV.

Another recent study using whole-exome sequencing data on 200,337 UK Biobank participants and focused on cardiometabolic traits has also performed exome-wide rare variant analyses with rare (pLOF and deleterious missense) variants[42]. Restricting to the traits included in our analyses (BMI, HDL, LDL and IGF-1) and the 19 gene-trait associations with q-value< 0.05 in[42], we find that 17 of them are significant in GeneScan3D, of which 16 (94%) are significant in BIGKnock (two-sided p= 1.7×10^{-2}).

Effector BIGKnock Genes. We further restrict the list of BIGKnock significant genes by identifying those that coincide with the closest gene (among all genes) to the top significant GWAS SNP at a locus. Among 6,195 significant loci across 20 quantitative traits, we identify 3,839 (62%) such loci (Supplemental Table 29). Effector BIGKnock genes have significantly higher pLI scores relative to GeneScan3D significant genes as well as genes that are never selected by BIGKnock across a variety of binary and quantitative traits considered here (Figure S9). Furthermore, for significant loci that do not contain effector BIGKnock genes, an additional 877 (14%) loci have only one gene significant under BIGKnock. Therefore, using the BIGKnock significant genes we can prioritize one potentially causal gene for 76% of the loci.

Mouse phenotype enrichment analyses. Using ToppFun[28] we have tested whether the effector BIGKnock genes are enriched in sets of genes associated with mouse phenotypes (Figures S10-S11). Effector BIGKnock genes are enriched in gene sets corresponding to relevant mouse phenotypes. For example, among the most significantly enriched phenotypes were abnormal systemic arterial blood pressure for BP-diastolic, abnormal erythroid progenitor cell morphology for RDW, abnormal calcium level for Calcium, abnormal circulating LDL cholesterol level for LDL cholesterol, decreased circulating HDL cholesterol level for HDL cholesterol, abnormal circulating hormone level and abnormal postnatal growth for IGF-1.

Comparisons with other locus-to-gene linking methods on gold-standard gene sets. We have compared the accuracy of effector BIGKnock genes to other methods to prioritize putative causal genes at GWAS loci, including the closest gene footprint to the top GWAS SNP as well as more recent methods such as combined SNP-to-gene (cS2G)[6] and Locus-to-gene (L2G)[7], using two gold-standard gene sets from the literature.

Specifically, we first consider 36 expert-curated genes with high confidence[7], as well as 120 effector genes identified using rare pLOF variants in[30]. For our analyses we focus on 138 gene-trait associations overlapping loci that are significant using the BIGKnock test. As control genes we consider the remaining genes at those loci for a total of 2,013 genes. For cS2G, we focus on a subset of 84 gold-standard genes and 1,303 control genes for 10 traits analyzed both here and in[6]. We compare methods in terms of precision and recall, where precision for a method is computed as the fraction of positive genes among the genes prioritized by that method, and recall is computed as the fraction of positive genes prioritized by that method (Figure 5a). BIGKnock effector genes have the highest precision among all methods considered, i.e. 0.67; the recall is also high (0.77). By comparison, cS2G achieves a higher recall (0.9 for cS2G score > 0.5) but at a greatly reduced precision (0.3). Closest gene footprint has similar recall (0.83) as effector genes, but reduced precision (0.59). L2G has lower recall (0.7) and precision (0.56) relative to effector genes. Furthermore, combining BIGKnock with other scores (such as cS2G and L2G) generally leads to improved precision over the individual cS2G and L2G scores (Supplemental Table 30).

We consider a second set of stringently defined positive genes, including Mendelian disease genes and drug targets as described in Forgetta et al.[31]. Specifically, we consider 199 genes that corresponded to traits Type 2 diabetes, BP-Systolic, BP-Diastolic, LDL-Cholesterol, Calcium, Direct bilirubin and RDW considered in our analyses. We focus on 62 genes residing at BIGKnock significant loci. As control genes we consider all genes at these 1Mb loci for a total of 973 genes. For cS2G, we focus on a subset of 55 gold-standard genes and 831 control genes for 5 traits analyzed both here and in[6] (Calcium and Direct bilirubin do not have cS2G gene scores). BIGKnock effector genes have the highest precision among all individual methods

considered, i.e. 0.41; the recall is also relatively high (Figure 5b, 0.42). By comparison, cS2G achieves a higher recall (0.53) but at a greatly reduced precision (0.2). Closest gene footprint has higher recall (0.5), but slightly lower precision (0.39). L2G has slightly lower recall (0.4) and precision (0.39). Furthermore, combining BIGKnock with other scores (such as cS2G and L2G) generally leads to improved precision over the individual methods (Supplemental Table 31).

Finally, we have compared BIGKnock with L2G and cS2G for several known causal genes, including *ASGR1*-Cholesterol, *ANGPTL4*-Cholesterol and *ALDH2*-CAD, as previously discussed (Figure S12). For *ASGR1*, all three methods identify *ASGR1* with high scores; however, cS2G identifies four such genes at the locus. For *ANGPTL4*, only BIGKnock and cS2G identify it among high scoring genes. However cS2G identifies three other genes with similar high score at this locus. For *ALDH2*, only BIGKnock and L2G identify it among the highest scoring gene; however L2G identifies five such genes at this locus. Results for all putative causal genes and loci highlighted before are similar (Figures S13-S14).

Characteristics of prioritized genes. We have focused here on prioritizing genes at $\sim 80\%$ loci that have either effector genes, i.e. the gene closest to the most significant GWAS SNP is significant using the BIGKnock test, or loci where BIGKnock prioritizes only one gene. We show that these genes have certain interesting properties: (1) they have significantly longer CDS (Coding DNA Sequence), and (2) higher LOF mutation rates than genes that are not significant using the BIGKnock test (Figure S15). Note that these properties are true more generally for BIGKnock significant genes relative to genes that are not significant using the BIGKnock test at effector loci, but not at non-effector loci. However focusing on those non-effector loci where BIGKnock prioritizes only one gene we find that the prioritized genes have significantly longer CDS and higher LOF mutation rates than non-significant genes. Our results are consistent with previous studies that showed that highly conserved genes (including putative disease causing genes) have, rather counterintuitively, higher mutation rates[32]. Specifically, Michaelson et al. showed that hypermutability is correlated with highly conserved sequence using whole genome sequencing data. Although the exact mechanisms underlying this relationship are not known, one possible explanation is that these genes, on account of their essential nature, are highly transcribed and consequently more susceptible to transcription-mediated mutagenic events.

Discussion

A main limitation of gene-based tests when incorporating putative regulatory variants, such as eQTLs or variants residing in regulatory elements such as promoters and enhancers, is the potentially high false positive rate due to LD confounding and co-regulation. We propose here a scalable gene-based test that reduces the LD confounding effect and can prioritize putative causal genes at GWAS significant loci. The proposed test goes beyond state-of-the-art gene-based tests by allowing integration of a wider class of regulatory variants than eQTLs, and by performing conditional analysis, thereby adjusting for LD, at genome-wide level.

We show that BIGKnock reduces the number of significant associations at a locus relative to conventional tests despite a more liberal FDR adjustment, and retains with high probability ($> 90\%$) the likely causal genes as shown using the effector and rare variant association results in[30]. Furthermore, between $62\% - 67\%$ of loci with BIGKnock significant genes have the closest gene to the top GWAS SNP at the locus being significant under BIGKnock (Supplemental

Tables 5 and 29). In addition to such effector BIGKnock genes, BIGKnock also prioritizes genes that are not necessarily nearest to the top GWAS SNP. Overall, approximately 80% of loci have one single gene prioritized based on significant genes detected by BIGKnock.

BIGKnock is complementary to other locus-to-gene strategies in the literature that are based on supervised machine learning models and fine-mapping results. BIGKnock prioritizes causal genes via a formal gene-based test that limits confounding due to LD relative to existing tests in the literature. Therefore BIGKnock is less functionally informed relative to existing locus-to-gene strategies, and therefore less affected by potential biases in existing training datasets. Combining significant genes in BIGKnock with other functionally informed causal gene prioritization methods is a promising avenue for increasing performance. We show that relative to other causal gene prioritization approaches, the proposed method has improved precision while achieving high recall, which is important in this setting due to costly follow-up functional studies. Note that for $\sim$ 20% of the loci we are not providing a single prioritized putative causal gene; such loci include those with multiple potentially causal genes, for example loci with co-regulation where a causal enhancer may regulate multiple genes, which further complicates the prioritization task.

Although it is a challenging task to prove that the prioritized genes from any method are indeed causal, we show multiple lines of evidence from mouse phenotype data, curated gold-standard gene lists, mutation rate data and supporting literature that BIGKnock is helpful in identifying putative causal genes including several examples with known causal links in the literature such as *ASGR1* and *ANGPTL4* and cholesterol, and *ALDH2* and coronary artery disease. These prioritized genes can serve as good candidates for further functional studies.

We have implemented BIGKnock in a computationally efficient R package that can be applied generally to the analysis of biobank scale data.

Methods

Overview of GeneScan3D and knockoff-based extensions

We first describe the details of a gene-based test (GeneScan3D) that incorporates noncoding variants using long-range chromatin data[5]. Assume there are n samples with p variants in a gene plus buffer region as well as the corresponding regulatory elements. For i-th individual, we denote Y_i as the phenotype, $\boldsymbol{G}_i$ as the $p \times 1$ genotype vector and $\boldsymbol{X}_i$ as the $d \times 1$ covariate vector including an intercept. We are interested in testing for association between the phenotype and the p variants, while adjusting for covariates. For unrelated individuals, we consider the generalized linear model (GLM):

$$g(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{\beta},$$

where μ_i is the conditional mean of phenotype Y_i conditional on covariates, $\boldsymbol{\alpha}$ is a $d \times 1$ vector of regression coefficients for d covariates (including an intercept) and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients for p variants.

We scan the gene plus buffer region ($\pm$ 5 Kb) using L 1D windows with sizes 1-5-10 Kb, then construct 3D windows by adding one enhancer to each 1D window. For each gene, we focus on GeneHancer and ABC enhancers[46,47] that are outside the gene plus buffer region, containing at least 2 variants and with length between 0.5 Kb to 10 Kb. In the ABC model[47],

we only incorporate predicted ABC enhancers with ABC scores ≥ 0.02 for 5 human cell types and tissues, i.e., K562, GM12878, NCCIT, LNCAP, hepatocytes.

Assuming R enhancers for a gene, then we construct $L \times R$ 3D windows. For each 3D window, we conduct (i) Burden and SKAT tests for common variants (MAF $\geq$ 0.01) with weights one; (ii) Burden and SKAT tests for rare variants (MAF $<$ 0.01 and MAC $\geq$ 10) with weights Beta(MAF$_j$; 1, 25); (iii) Burden test for aggregation of ultra-rare variant (MAC$<$ 10) and (iv) Single variant score tests for common and rare variants with MAC $\geq$ 10. The Cauchy combination method[16] is applied to combine p-values from the above tests within each 3D window. Finally, we compute the GeneScan3D p-value by combining $L \times R$ 3D window's p-values using Cauchy combination method.

Knockoffs-based extension. By incorporating distal regulatory elements, gene-based tests can leverage noncoding genetic variation to improve power of gene-based tests. However, due to linkage disequilibrium (LD) and/or co-regulation of multiple genes by the same regulatory element, many of the significant genes may be false positives. Hence, the multiple knockoff framework is implemented to attenuate the confounding effect of LD and prioritize putative causal genes with controlled false discovery rate (FDR). Note that co-regulation is still a problem and cannot be addressed by the proposed approach.

To generate multiple knockoff genotypes, we consider the general sequential conditional independent tuples approach[14,48,49]. Specifically, we sequentially sample $\tilde{G}_j^1, \ldots, \tilde{G}_j^M$ independently from $\mathcal{L}(G_j | G_{-j}, \tilde{G}^1_{1\ldots j-1}, \ldots, \tilde{G}^M_{1\ldots j-1})$, where M is the number of knockoffs. Note that we can leverage the approximate block structure for LD in the genome to only include variants in a neighborhood of the current variant j. The knockoff genotypes are exchangeable with the original genotypes G, and lead to guaranteed FDR control. With the assumption that genotypes can be approximately modeled by a multivariate normal distribution, we consider a computational efficient auto-regressive model to estimate:

$$\hat{G}_j = \hat{\alpha} + \sum_{k \neq j} \hat{\beta}_k G_k + \sum_{m=1}^{M} \sum_{k \leq j-1} \hat{\gamma}_k^m \tilde{G}_k^m. \tag{1}$$

By calculating the residual $\hat{\epsilon}_j = G_j - \hat{G}_j$ and its M permutation, the knockoff features $\tilde{G}_j^{\,m} = \hat{G}_j + \hat{\epsilon}_j^{*m}$ are obtained.

After generating multiple knockoffs, we conduct the proposed gene-based test on the original genotype and knockoff genotypes for each gene. The feature statistic for each gene G is then defined as

$$W_G = (T_G - \text{median } T^m_{\tilde{G}}) I_{T_G \geq \max_{1 \leq m \leq M} T^m_{\tilde{G}}},$$

where $T_G = -\log_{10}(p_G)$ and $T^m_{\tilde{G}} = -\log_{10}(p^m_{\tilde{G}})$ are the importance score for gene G in original genotype and knockoff cohort, and I is an indicator function. We compute the threshold τ for FDR control at a certain level q:

$$\tau = \min \left\{ t > 0 : \frac{\frac{1}{M} + \frac{1}{M} \#\{G : \kappa_G \geq 1, \tau_G \geq t\}}{\#\{G : \kappa_G = 0, \tau_G \geq t\}} \leq q \right\},$$

where $\kappa_G = \text{argmax}_{0 \leq m \leq M} T^m_{\tilde{G}}$ (note that $T^0_{\tilde{G}} = T_G$) and $\tau_G = T_G - \text{median } T^m_{\tilde{G}}$. Finally, we select as significant those genes with $W_G \geq \tau$ (Ma et al.[5]).

q-value. We additionally compute the corresponding q-value for a gene, q_G. The q-value already incorporates correction for multiple testing, and is defined as the minimum FDR that can be attained when all tests showing evidence against the null hypothesis at least as strong as the current one are declared as significant. In particular, we define the q-value for a gene G with feature statistic $W_G > 0$ as

$$q_G = \min_{t \le W_G} \frac{\frac{1}{M} + \frac{1}{M}\#\{G : \kappa_G \ge 1, \tau_G \ge t\}}{\#\{G : \kappa_G = 0, \tau_G \ge t\}},$$

where $\frac{\frac{1}{M} + \frac{1}{M}\#\{G:\kappa_G \ge 1, \tau_G \ge t\}}{\#\{G:\kappa_G = 0, \tau_G \ge t\}}$ is an estimate of the proportion of false discoveries for multiple knockoffs if we were to select all genes with $\kappa_G = 0, \tau_G \ge t$ (with $t > 0$). For genes with feature statistic $W_G = 0$ (i.e., $\kappa_G \ge 1$), we set $q_G = 1$ and never select those genes.

Shrinkage leveraging algorithm for knockoffs generation

The computational cost of knockoff generation is substantial for biobank-scale data with hundreds of thousands of samples and millions of genetic variants. To reduce the computational time, we optimize the knockoff generation using the shrinkage leveraging (SL) algorithm[15].

To filter out highly-correlated variants in the knockoff generation, we apply hierarchical clustering. We compute correlations for all pairs of variants in regions containing the gene plus buffer region (±100 Kb neighborhood) and enhancers (±50 Kb neighborhood). Variants with correlation ≥0.75 are clustered together and one representative variant is selected for each cluster. Specifically, if a cluster contains variants inside the gene plus buffer/enhancer region, we randomly select one representative variant inside the gene plus buffer/enhancer region. Otherwise, we randomly select one variant as representative.

We draw $r = 10n^{1/3} \log n$ subsamples from n samples with importance sampling probabilities:

$$\pi_i = 0.5\pi_i^{\text{Lev}} + 0.5\pi_i^{\text{Unif}}, i = 1, \ldots, n,$$

where $\pi_i^{\text{Unif}} = 1/n$ follows uniform distribution and $\pi_i^{\text{Lev}} = \sum_{j=1}^{p} U_{ij}^2 / \sum_{i=1}^{n}\sum_{j=1}^{p} U_{ij}^2$, U is the orthogonal singular vectors of $(\mathbf{1}, G, \tilde{G})$. We then form a weighted linear regression model (1) with weights $w_i = 1/(r\sqrt{\pi_i})$ for r subsamples and compute the least square estimates:

$$(\hat{\alpha}^{\text{SL}}, \hat{\boldsymbol{\beta}}^{\text{SL}}, \hat{\boldsymbol{\gamma}}^{\text{SL}}) = [\text{cov}(\mathbf{1}, G^{\text{SL}}, \tilde{G}^{\text{SL}})]^{-1}(\mathbf{1}, G^{\text{SL}}, \tilde{G}^{\text{SL}})^T G_j^{\text{SL}},$$

where $(\mathbf{1}, G^{\text{SL}}, \tilde{G}^{\text{SL}})$ are the weighted genotypes and knockoffs for r shrinkage leveraging subsampling from $(\mathbf{1}, G, \tilde{G})$ and G_j^{SL} corresponds to r sampling rows of G_j. The inverse term $[\text{cov}(\mathbf{1}, G^{\text{SL}}, \tilde{G}^{\text{SL}})]^{-1}$ can be approximated by spectral decomposition[15]. Finally, we generate the knockoff features for n samples using the least square estimates $(\hat{\alpha}^{\text{SL}}, \hat{\boldsymbol{\beta}}^{\text{SL}}, \hat{\boldsymbol{\gamma}}^{\text{SL}})$. In summary, we select a subset of "informative" samples to estimate intermediate parameters used for knockoff generation and thus improve the computational efficiency while maintaining the statistical performance (i.e., FDR control and power) of the knockoff framework[15].

To efficiently store the knockoff genotypes, we use the Genomic Data Structure compressed files based on *gdsfmt* R package[15,51].

Generalized linear mixed effects model for related samples

For very large sample sizes as in biobanks, we account for sample relatedness using the generalized linear mixed-effects model (GLMM). Specifically, we assume:

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\beta} + b_i,$$

where the random effect $\boldsymbol{b} = (b_1, \ldots b_n)^T \sim \text{MVN}(\boldsymbol{0}, \tau\boldsymbol{\psi})$ and $\boldsymbol{\psi}$ is the $n \times n$ genetic relationship matrix (GRM).

Following SAIGE-Gene[13], we consider three steps for the UK Biobank data. In step 1 we construct the sparse GRM $\boldsymbol{\psi}_S$ with cutoff 0.125 for $n = 405,296$ British samples using 106,256 pruned markers. In step 2 we fit the null GLMMs for binary and quantitative traits. Both steps are using the existing software implementation in SAIGE/SAIGE-Gene[12,13]. In step 3 we perform the gene-based test for each gene using the fitted values $\hat{\mu}$ and estimated variance ratio $\hat{r}$ obtained in step 2. Note that due to the light sample relatedness of UK Biobank data, one can use the sparse GRM to fit null GLMM and estimate variance ratio, which is much more computationally efficient than using the dense GRM[13].

To fit the GLMM under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ in a computationally efficient way, SAIGE uses the preconditioned conjugate gradient method[52] that allows calculating the log quasi-likelihood and average information without take the inverse of $n \times n$ matrix. Specifically, SAIGE maximizes the log quasi-likelihood using the average information restricted maximum likelihood algorithm (AI-REML)[53] to iteratively estimate $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}, \hat{\phi}, \hat{\tau})$ (note that the dispersion parameter $\hat{\phi} = 1$ for binary traits). Denote $\hat{\Sigma} = \hat{W}^{-1} + \hat{\tau}\boldsymbol{\psi}$, where $\hat{W} = \hat{\phi}^{-1}I$ for quantitative traits and $\hat{W} = \text{diag}(\hat{\mu}_1(1-\hat{\mu}_1), \ldots \hat{\mu}_n(1-\hat{\mu}_n))$ for binary traits. Denote the covariate-adjusted genotype matrix as $\tilde{G} = G - X(X^TWX)^{-1}X^TWG$ and the projection matrix $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}$.

After fitting the null GLMM, we obtain the variance ratio $\hat{r} = \tilde{\boldsymbol{g}}^T\hat{P}\tilde{\boldsymbol{g}}/\tilde{\boldsymbol{g}}^T\hat{P}_S\tilde{\boldsymbol{g}}$ where $\tilde{\boldsymbol{g}}$ is the covariate-adjusted single variant genotype vector, $\hat{P}_S = \hat{\Sigma}_S^{-1} - \hat{\Sigma}_S^{-1}X(X^T\hat{\Sigma}_S^{-1}X)^{-1}X^T\hat{\Sigma}_S^{-1}$ and $\hat{\Sigma}_S = \hat{W}^{-1} + \hat{\tau}\boldsymbol{\psi}_S$. The variance ratio, which is estimated using a set of 30 randomly selected variants and shown to be approximately constant for all variants[12], is used to calibrate the score test statistics and variance-covariance matrix of gene-based tests for GLMM.

For the single variant score test in GeneScan3D, $S_j = \sum_{i=1}^n \tilde{G}_{ij}(Y_i - \hat{\mu}_i)/\hat{\phi}$. We consider the variance-adjusted test statistic:

$$T_j^{\text{adj}} = \frac{S_j}{\sqrt{\tilde{\boldsymbol{g}}_j^T\hat{P}\tilde{\boldsymbol{g}}_j}},$$

where $\tilde{\boldsymbol{g}}_j$ is the covariate-adjusted genotype vector of j-th variant. The approximation of $\text{var}(S_j) = \tilde{\boldsymbol{g}}_j^T\hat{P}\tilde{\boldsymbol{g}}_j = \hat{r}\tilde{\boldsymbol{g}}_j^T\hat{P}_S\tilde{\boldsymbol{g}}_j \approx \hat{r}\tilde{\boldsymbol{g}}_j^T\hat{\Sigma}_S^{-1}\tilde{\boldsymbol{g}}_j$ and the score test p-value can be computed based on $S_j^2/\text{var}(S_j) \sim \chi^2_{\text{df}=1}$.

The Burden and SKAT test statistics in GeneScan3D can be written as:

$$Q_{\text{Burden}} = \left(\sum_{j=1}^p w_jS_j\right)^2, \quad Q_{\text{SKAT}} = \sum_{j=1}^p w_j^2S_j^2,$$

where w_j is the weight of each variant. The joint null distribution of $\boldsymbol{S} = (S_1, \ldots S_p)$ follows a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\tilde{G}^T\hat{P}\tilde{G} = G^T\hat{\Sigma}^{-1}G - (G^T\hat{\Sigma}^{-1}X)(X^T\hat{\Sigma}^{-1}X)^{-1}(X^T\hat{\Sigma}^{-1}G) = G^T\hat{P}G$. We adjust the covariance matrix for GLMM as $K =$

$\hat{r}G^T\hat{P}_S G$. Since both $\hat{\Sigma}$ and G are sparse matrices, K can be calculated by using the sparse LU decomposition (solve function in R) for each 3D window. Then the Burden p-value is obtained from a scaled chi-square distribution $\tilde{\lambda}_B\chi^2_1$, where $\tilde{\lambda}_B = (w_1, \ldots, w_p)K(w_1, \ldots, w_p)^T$. The SKAT p-value is obtained from a mixture of chi-square distribution $\sum_{j=1}^{p}\tilde{\lambda}_{S_j}\chi^2_1$ using Davies method[54], where $\tilde{\lambda}_{S_j}$ are the eigenvalues of $\text{diag}(w_1, \ldots, w_p)K\text{diag}(w_1, \ldots, w_p)$.

Saddlepoint approximation for gene-based test

One challenge for binary traits in biobanks is the possibility of highly unbalanced case:control ratios. In such cases we implement the saddlepoint approximation (SPA) to recalibrate the score test statistics for gene-based testing[55,56]. Specifically, under case-control imbalance, the distribution of score statistics $\boldsymbol{S} = (S_1, \ldots S_p)$ is skewed, in which case one needs to adjust the covariance matrix K using SPA. As in[13,56], we first compute the p-values of single-variant score test by SPA $\tilde{p}_j$, then the SPA-adjusted variance $\tilde{v}_j = S_j^2/Q(1-\tilde{p}_j)$, where Q is the quantile function of χ^2_1. The adjusted covariance matrix $\tilde{K} = \sqrt{\tilde{V}}K\sqrt{\tilde{V}}$, where $\tilde{V} = \text{diag}(\tilde{v}_1/\hat{v}_1, \ldots, \tilde{v}_p/\hat{v}_p)$ and $\hat{v}_j = K[j, j]$ is the estimated variance of S_j. The adjusted covariance matrix $\tilde{K}$ is used to compute the SPA gene-based p-values of SKAT and Burden.

UK Biobank data analyses

The UK Biobank data contains data on 488,377 individuals. All individuals underwent genome-wide genotyping with UK Biobank Axiom array from Affymetrix and UK BiLEVE Axiom arrays (~ 825,000 markers). Genotype imputation was carried out using a 1000 Genomes reference panel with IMPUTE4 software[11]. We apply several quality-control filters, keeping only variants with MAF≥ 0.01 imputed with high confidence ($R^2 \geq 0.8$). This resulted in 9,233,477 imputed variants that were available for the analyses. We restrict our analyses to 405,296 participants (218,068 females and 187,228 males) with British ancestry. We adjust for covariates including sex, age, age^2, age × sex and 5 principal components. For principal component analysis, we used a set of common genotypes (MAF> 0.01) pruned using the following command in PLINK –indep-pairwise 500 50 0.05 with 35,226 pruned variants using FlashPCA[58]. A total of 17,753 genes with gene length < 500 kb and with at least 2 variants in the gene plus buffer region were tested. The details on the traits analyzed are given in Table S1.

We use 106,256 pruned genotyped markers to construct the sparse GRM with relatedness coefficient cutoff ≥ 0.125, then fit null GLMMs for several binary and quantitative traits using SAIGE[12,13]. The 106,256 markers were pruned from the UK Biobank genotype data using PLINK with pairwise LD threshold $r^2 \leq 0.05$, MAF≥ 0.01, 95% genotyping rate, window size of 500 bp and step size 50 bp. Based on the sparse GRM, there are 21,397 related pairs among the 405,296 participants, including 8 duplicate twins (kinship coefficient >0.354), 8,275 1st-degree relatives (kinship coefficient between 0.177 to 0.354) and 13,114 2nd-degree relatives (kinship coefficient ≤ 0.177)[57].

Enrichment of BIGKnock associations among genes closest to lead GWAS SNPs

We consider the significant loci for different UK Biobank binary and quantitative traits. We use the SAIGE summary statistics from the existing UK Biobank studies for binary traits

(https://pheweb.org/UKB-SAIGE/) and the GWAS summary statistics for UK Biobank quantitative traits were obtained from the Neale Lab (https://www.nealelab.is/ukbiobank). For each significant locus, all genes within the locus are ranked according to the distance to the lead GWAS variant. The enrichment is then defined as the ratio of the proportion of BIGKnock significant genes that are ranked k-th and the proportion of the remaining genes at the locus that are ranked k-th, where $k = 1, \ldots, 10$.

Locus-to-gene scores

L2G. We selected GWAS analyses from the OpenTarget Genetics Portal[7] to match the 24 traits tested by BIGKnock. For the four binary traits we use summary statistics from SAIGE[12]. For ten quantitative traits (Apolipoprotein A, Calcium, Cholesterol, Cystatin C, Direct bilirubin, eGFR, Glycated hemoglobin HbA1c, HDL cholesterol, IGF-1, and LDL direct) we use summary statistics from[59]. The remaining ten quantitative traits are part of the Neale lab UKB GWAS round 2 results. OpenTarget used the "locus-to-gene" (L2G) model to prioritize likely causal genes at each GWAS locus detected by these studies. An L2G score is derived from gene distance, molecular QTL colocalization, chromatin interaction, and pathogenicity to quantify the causal probability of a gene. We downloaded the L2G scores and selected the gene with the highest L2G score for each GWAS locus for the 24 traits.

cS2G. The combined SNP-to-gene strategy (cS2G)[6] includes seven SNP-to-gene (S2G) linking strategies such as Exon, Promoter, two fine-mapped cis-eQTL strategies, EpiMap enhancer-gene linking, Activity-By-Contact, and Cicero. A cS2G score is computed for a SNP and a gene as a linear combination of linking scores from different S2G strategies, and the optimal weights are estimated to maximize the recall under a constraint of precision ≥ 0.75 with non-trait-specific training critical gene set. cS2G was applied to fine-mapping results of 49 UK Biobank diseases and traits; a cS2G score > 0.5 was used to identify high-confidence SNP-gene-disease triplets. In our analyses, we considered the cS2G predicted target genes of fine-mapping results for ten UKBB traits: CAD, Cholesterol, HDL cholesterol, LDL cholesterol, HbA1c, MPV, Platelet count, RDW, BMI and BP-Diastolic, all with cS2G scores > 0.5.

Gold-standard genes. For 4 binary traits and 20 quantitative traits considered in our analyses, we identified 36 expert-curated gold-standard genes with high confidence for CAD, Cholesterol, HDL cholesterol and LDL cholesterol[7]. 120 effector genes are identified in[30] for 18 quantitative traits (Cholesterol, HDL cholesterol, LDL cholesterol, HbA1c, MPV, Platelet count, RDW, BMI, BP-Diastolic, Apolipoprotein, Calcium, Direct bilirubin, MRV, MSCV, Reticulocyte count, IGF 1, Neutrophil count and Cystatin C). Among 36 gold-standard genes, there are 33 GeneScan3D significant genes, and among them, 32 are BIGKnock significant (with retention rate 97%). Among 120 Backman effector genes, there are 116 GeneScan3D significant genes, and among them, 106 are BIGKnock significant (with retention rate 91.4%).

Positive genes in Forgetta et al.[31]. The positive genes for 12 traits considered in Forgetta et al.[31] were selected based on Mendelian disease genes or positive control drug targets. There are in total 494 positive genes across 12 diseases and traits, with 381 known to cause Mendelian forms of the disease and 113 drug targets. We focus on 199 gene-trait associations for 7 traits considered in our paper (Type 2 diabetes, BP-Systolic, BP-Diastolic, LDL-Cholesterol, Calcium, Direct bilirubin and Red blood cell distribution width).

Genome build

All genomic coordinates are given in GRCh37/hg19.

Data Availability

The manuscript used UK Biobank data available at https://biobank.ndph.ox.ac.uk/showcase/ (under UKBB project ID number 41849), summary statistics from SAIGE: ftp://share.sph.umich.edu/UKBB_SAIGE_HRC/, Neale lab UKB GWAS round 2 results: http://www.nealelab.is/uk-biobank/, gold-standard genes: https://github.com/opentargets/genetics-gold-standards.

Code Availability

We have implemented BIGKnock in a computationally efficient R package that can be applied generally to the analysis of other large biobank datasets. The package can be accessed at: https://github.com/Iuliana-Ionita-Laza/BIGKnock. Results for additional UK Biobank traits will be made available at the same location.

References

1. Kim-Hellmuth, S. *et al*. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).

2. Mancuso, N., Freund, M.K., Johnson, R. *et al*. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).

3. Chun, S., Casparino, A., Patsopoulos, N. *et al*. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

4. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends. Genet.* **37**, 109–124 (2021).

5. Ma, S. *et al*. Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proc. Natl. Acad. Sci. USA* **118**, e2105191118 (2021).

6. Gazal S, Weissbrod O, Hormozdiari F, *et al*. Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. *medRxiv* 2021.08.02.21261488 (2021) doi: 10.1101/2021.08.02.21261488.

7. Mountjoy E, Schmidt EM, Carmona M, *et al*. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).

8. Li, X. *et al*. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).

9. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

10. Sey, N.Y.A., Hu, B. & Mah, W. *et al*. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020).

11. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, *et al*. The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **562**, 203–209 (2018).

12. Zhou, W. *et al*. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

13. Zhou, W. *et al*. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).

14. Candès, E, Fan, Y, Janson, L, Lv, J. Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection. *J. R. Stat. Soc. B.* **80**, 551-577 (2018).

15. He, Z., Guen, Y. L. *et al*. Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics *Am. J. Hum. Genet.* **108**, 1–18 (2021).

16. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E. & Lin, X. ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410-421 (2019).

17. Nasser, J. *et al*. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

18. Morris, J. A. *et al*. Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. Preprint at bioRxiv https://doi.org/10.1101/2021.04.07.438882v1 (2021).

19. Chen CH, Ferreira JCB, Mochly-Rosen D. ALDH2 and Cardiovascular Disease. *Adv Exp Med Biol.* **1193**: 53-67 (2019)

20. F. Xu, Y.G. Chen, L. Xue, R.J. Li, H. Zhang, Y. Bian, C. Zhang, R.J. Lv, J.B. Feng, Y. Zhang Role of aldehyde dehydrogenase 2 Glu504lys polymorphism in acute coronary syndrome *J. Cell. Mol. Med.* **15**, 1955-1962 (2011).

21. S. Takagi, N. Iwai, R. Yamauchi, S. Kojima, S. Yasuno, T. Baba, M. Terashima, Y. Tsutsumi, S. Suzuki, I. Morii, S. Hanai, K. Ono, S. Baba, H. Tomoike, A. Kawamura, S. Miyazaki, H. Nonogi, Y. Goto Aldehyde dehydrogenase 2 gene is a risk factor for myocardial infarction in Japanese men *Hypertens. Res.*, **25**, 677-681 (2002)

22. S.A. Jo, E.K. Kim, M.H. Park, C. Han, H.Y. Park, Y. Jang, B.J. Song, I. Jo A Glu487Lys polymorphism in the gene for mitochondrial aldehyde dehydrogenase 2 is associated with myocardial infarction in elderly Korean men *Clin. Chim. Acta*, **382**, 43-47 (2007)

23. Guo R, Xu X, Babcock SA, Zhang Y, Ren J. Aldehyde dedydrogenase-2 plays a beneficial role in ameliorating chronic alcohol-induced hepatic steatosis and inflammation through regulation of autophagy. *J Hepatol.* **62**:647–56 (2015).

24. Chen, C. H. *et al*. Mitochondrial aldehyde dehydrogenase and cardiac diseases. *Cardiovascular Research* **88**, 51–57 (2010).

25. Chen, Y. L. *et al*. Small Interfering RNA Targeting Nerve Growth Factor Alleviates Allergic Airway Hyperresponsiveness. *Mol. Ther. Nucleic. Acids.* **3**, E158 (2014).

26. Agarwal, A. K., Tunison, K., Dalal, J. S. *et al*. Metabolic, Reproductive, and Neurologic Abnormalities in Agpat1-Null Mice. *Endocrinology* **158**, 3954–3973 (2017).

27. Bond, S. T. *et al*. The E3 ligase MARCH5 is a PPARγ target gene that regulates mitochondria and metabolism in adipocytes. *Am. J. Physiol. Endocrinol. Metab.* **316**, E293–E304 (2019).

28. Chen, J., Bardes, E.E., Aronow, B.J. and Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305-W311 (2009).

29. Nioi, P. *et al*. Variant ASGR1 associated with a reduced risk of coronary artery disease. *New England Journal of Medicine* **374**, 2131–2141 (2016).

30. Backman, J. D. *et al*. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

31. Forgetta, V., Jiang, L., Vulpescu, N. A. *et al*. An effector index to predict target genes at GWAS loci. *Hum Genet* (2022).

32. Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151: 1431–1442 (2012).

33. Xie B, Shi X, Li Y, Xia B, Zhou J, Du M, Xing X, Bai L, Liu E, Alvarez F, Jin L, Deng S, Mitchell GA, Pan D, Li M, Wu J. Deficiency of ASGR1 in pigs recapitulates reduced risk factor for cardiovascular disease in humans. *PLoS Genet.* **17**:e1009891 (2021).

34. Lin, W. Slc39a8/zip8 Influences Complex Traits By Regulating Metal Ion Metabolism. Publicly Accessible Penn Dissertations. 2426 (2017).

35. Nebert, D. W. and Liu, Z. SLC39A8 gene encoding a metal ion transporter: discovery and bench to bedside. *Hum. Genomics.* **13**, 51 (2019).

36. Swoap, S. J., Weinshenker, D., Palmiter, R. D. and Garber, G. Dbh(-/-) mice are hypotensive, have altered circadian rhythms, and have abnormal responses to dieting and stress. *Am J Physiol Regul Integr Comp Physiol.* **286**, R108-113 (2004).

37. Lichtenstein, L. *et al*. Angptl4 upregulates cholesterol synthesis in liver via inhibition of LPL- and HL-dependent hepatic cholesterol uptake. *Arterioscler. Thromb. Vasc. Biol.* **27**, 2420–2427 (2007).

38. Jiang, C., Liu, Z., Hu, R. *et al*. Inactivation of Rab11a GTPase in Macrophages Facilitates Phagocytosis of Apoptotic Neutrophils. *J. Immunol.* **198**, 1660–1672 (2017).

39. Suehiro, F. *et al*. Impact of zinc fingers and homeoboxes 3 on the regulation of mesenchymal stem cell osteogenic differentiation. *Stem Cells Dev.* **20**, 1539–1547 (2011).

40. Chinetti, G. *et al*. PPAR-alpha and PPAR-gamma activators induce cholesterol removal from human macrophage foam cells through stimulation of the ABCA1 pathway. *Nat. Med.* **7**, 53–58 (2001).

41. Lyle, A. *et al*. Poldip2, a novel regulator of Nox4 and cytoskeletal integrity in vascular smooth muscle cells. *Circulation research* **105**, 249–259 (2009).

42. Jurgens, S. J., Choi, S. H., Morrill, V. N. *et al*. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat. Genet.* (2022).

43. Chen, J., Xu, H., Aronow, B. J. and Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC bioinformatics* **8**, 392 (2007)

44. Piero, J., Bravo, A., Queralt-Rosinach, N., Gutirrez-Sacristn, A., Deu-Pons, J., Centeno, E., Garca-Garca, J., Sanz, F. and Furlong, L.I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* **45**: D833-D839 (2017).

45. He, Z. Liu, L., Belloy, M. E. *et al*. Summary statistics knockoff inference empowers identification of putative causal variants in genome-wide association studies. *bioRxiv* 2021.12.06.471440 (2021) doi:10.1101/2021.12.06.471440.

46. Fishilevich, S. *et al*. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* Volume 2017, bax028 (2017).

47. Fulco, C.P., Nasser, J., Jones, T.R. *et al*. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

48. He, Z. *et al*. Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nat. Commun.* **12**, 3152 (2021).

49. Gimenez, J.R. & Zou, J. Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. *arXiv preprint arXiv:1810.11378* (2018).

50. Ma, P., Mahoney, M. W. & Yu, B. A Statistical Perspective on Algorithmic Leveraging. *Journal of Machine Learning Research* **16**, 861–911 (2015).

51. Zheng X. *et al*. SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).

52. Kaasschieter, E. F. Preconditioned conjugate gradients for solving singular systems. *J. Comput. Appl. Math.* **24**, 265–275 (1988).

53. Chen, H. *et al*. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).

54. Davies, R.B. The distribution of a linear combination of chi-square random variables. *Appl. Stat.* **29**, 323–333 (1980).

55. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).

56. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L. G. & Lee, S. UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *Am. J. Hum. Genet.* **106**, 3–12 (2020).

57. Zhou, W. *et al*. Set-based rare variant association tests for biobank scale sequencing data sets. *medRxiv* 2021.07.12.21260400 (2021) doi:10.1101/2021.07.12.21260400.

58. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, e93766 (2014).

59. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D. *et al*. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).

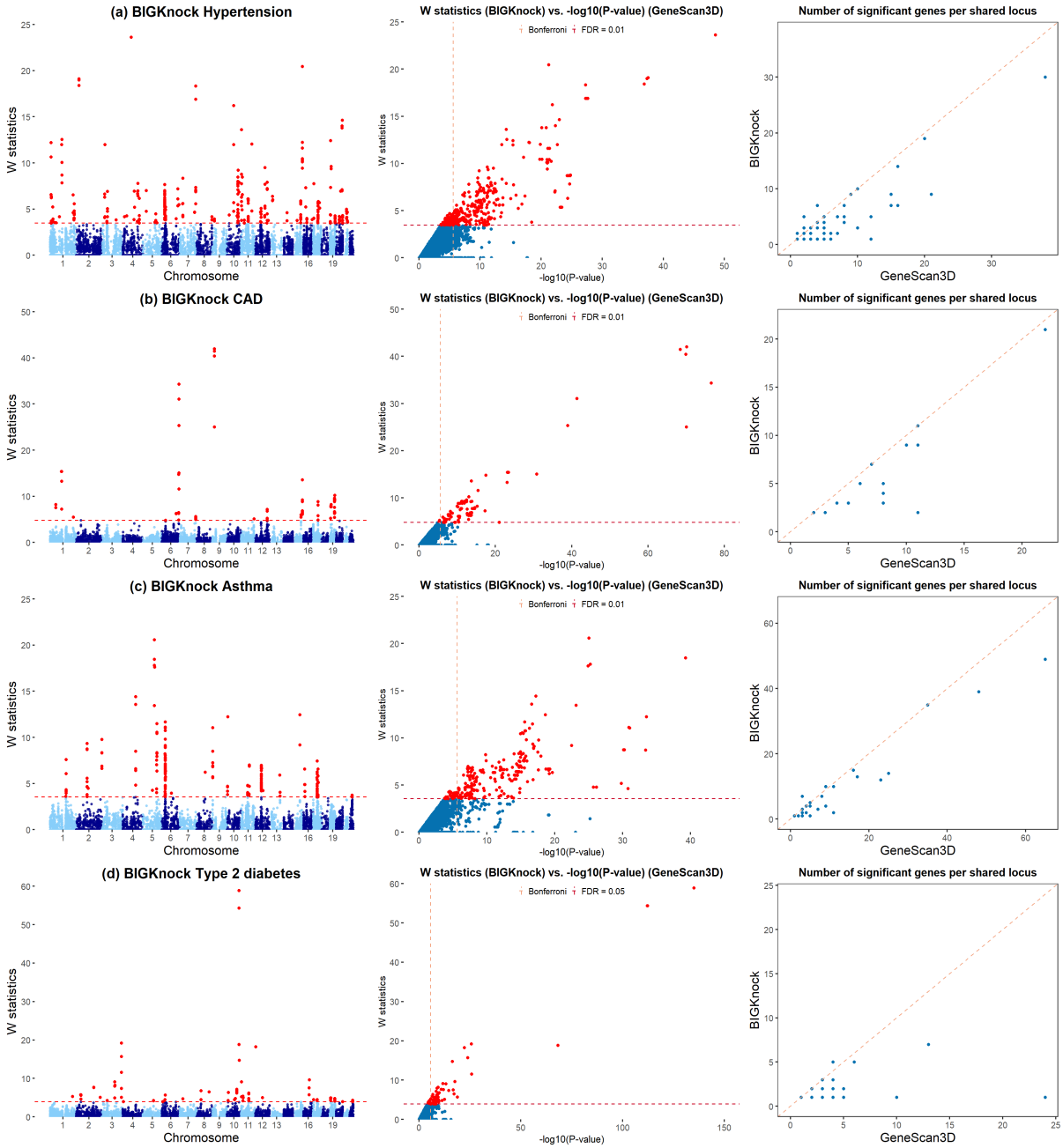

Figure 1: **Applications to UK Biobank binary traits**. a-d, Manhattan plots for BIGKnock, Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D), and Scatter plot of the number of significant genes per locus between conventional GeneScan3D and BIGKnock are shown for (a) Hypertension, (b) Coronary artery disease , (c) Asthma, and (d) Type 2 diabetes. The dashed lines in the left and middle panels show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

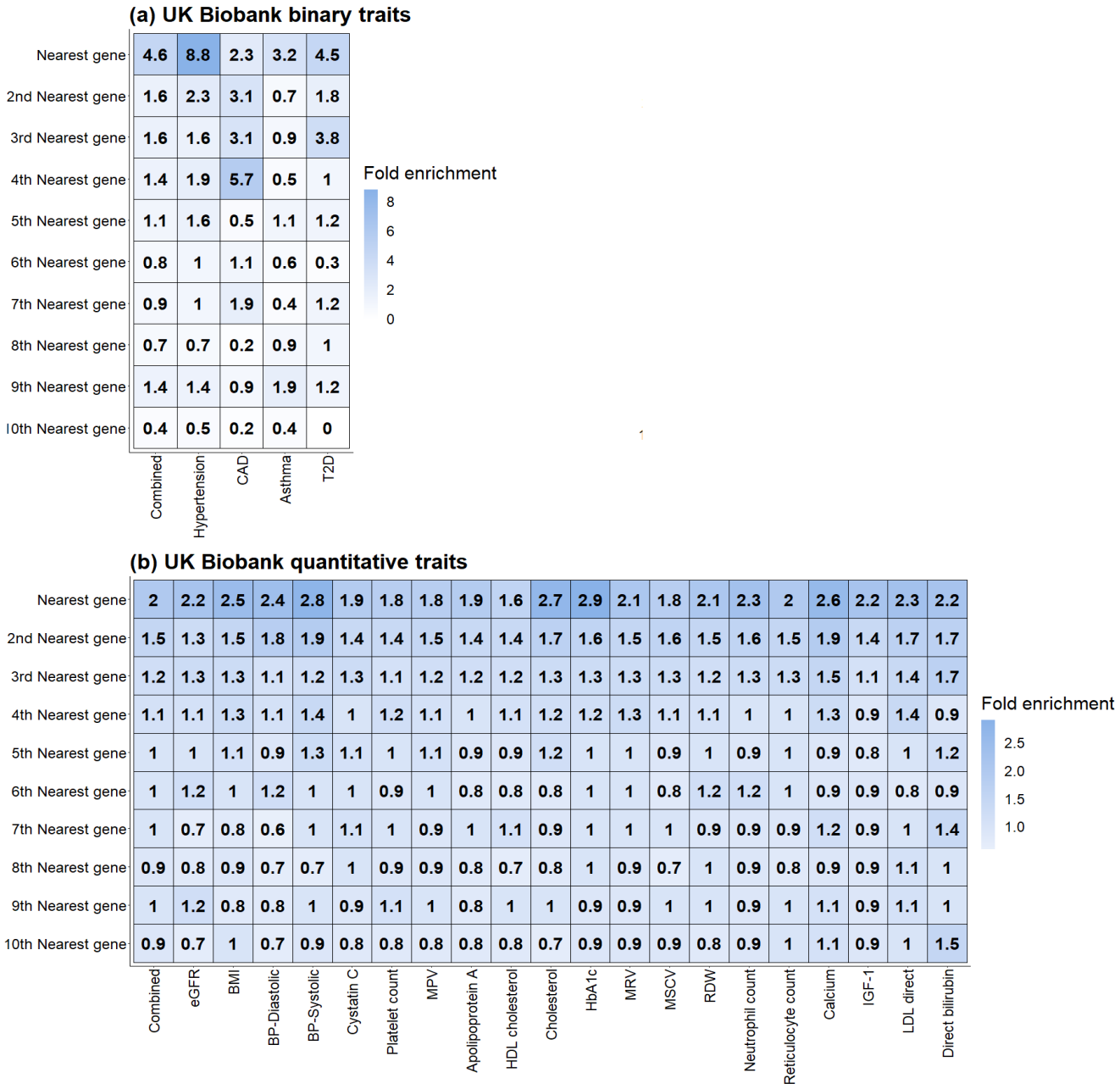

Figure 2: **Enrichment of BIGKnock significant genes among genes closest to the lead GWAS variant at BIGKnock significant loci**. Enrichment of BIGKnock significant genes for (a) the five combined binary traits and each binary trait separately: Hypertension, Coronary artery disease (CAD), Asthma and Type 2 diabetes (T2D); and (b) the twenty combined quantitative traits and each quantitative trait separately: eGFR, BMI, BP-Diastolic, BP-Systolic, Cystatin C, Platelet count, MPV, Apolipoprotein A, HDL cholesterol, Cholesterol, HbA1c, MRV, MSCV, RDW, Neutrophil count, Reticulocyte count, Calcium, IGF-1, LDL direct and Direct bilirubin.

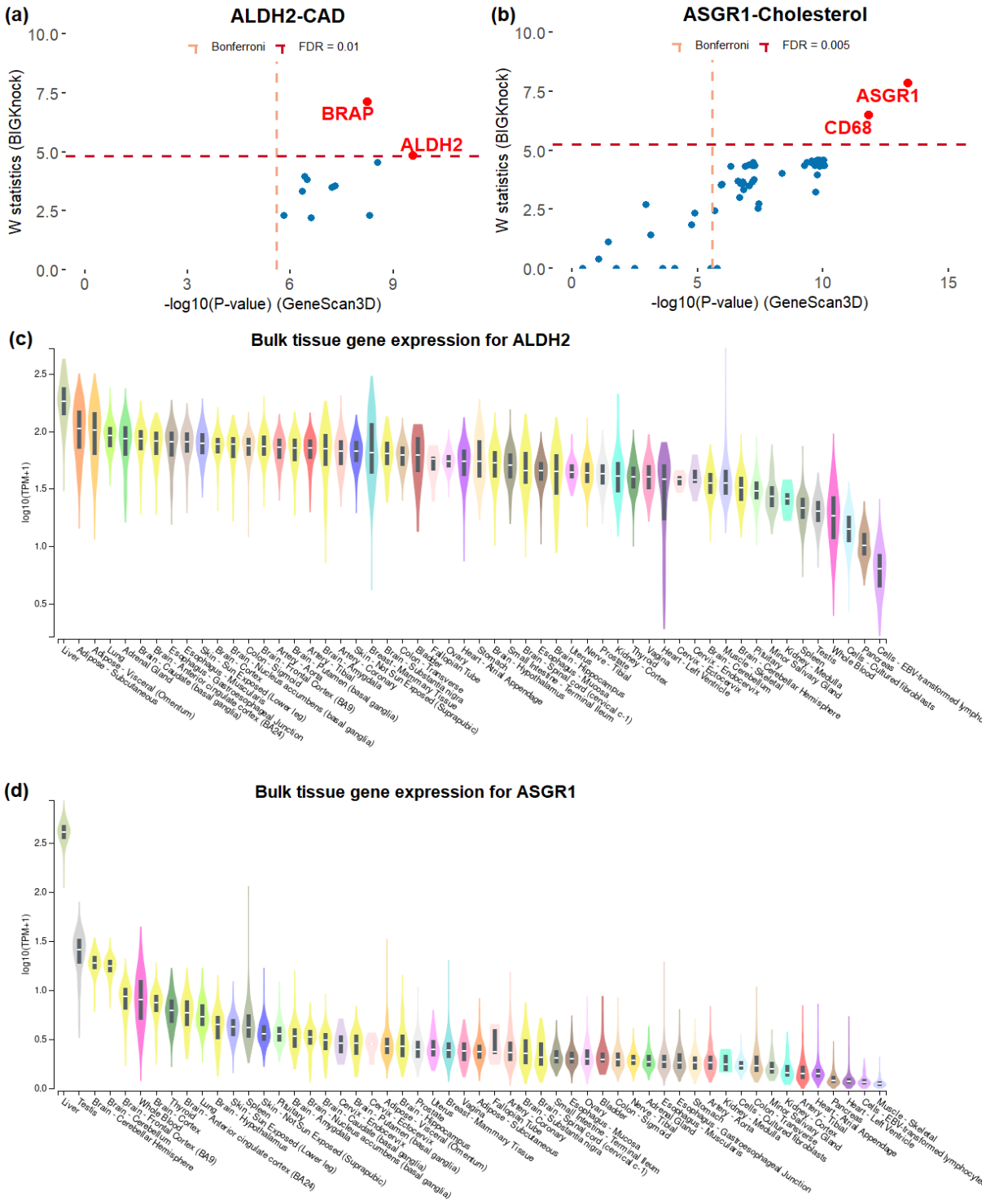

Figure 3: **ALDH2-CAD and ASGR1-Cholesterol loci.** (a) Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D) for the ALDH2-CAD locus, (b) Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D) for the ASGR1-Cholesterol locus, (c) GTEx gene expression across tissues for *ALDH2*, and (d) GTEx gene expression across tissues for *ASGR1*.

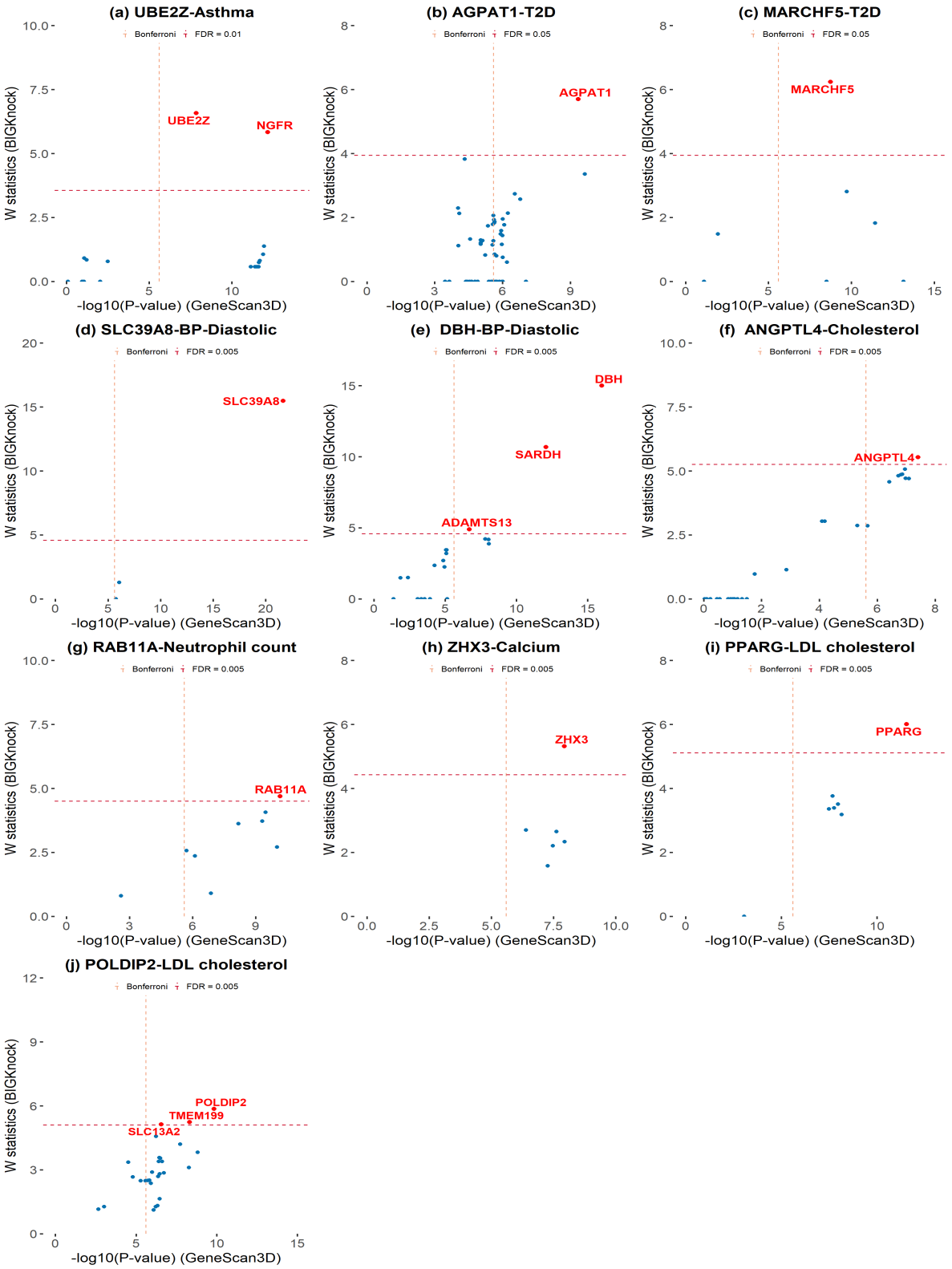

Figure 4: **Putative causal genes at selected loci for UK Biobank binary traits and quantitative traits**. Scatter plots of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D) for selected loci of (a) Asthma, (b-c) Type 2 diabetes (T2D), (d-e) BP-Diastolic, (f) Cholesterol, (g) Neutrophil count, (h) Calcium and (i-j) LDL cholesterol. Loci are named according to the most significant gene in BIGKnock. The dashed lines show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

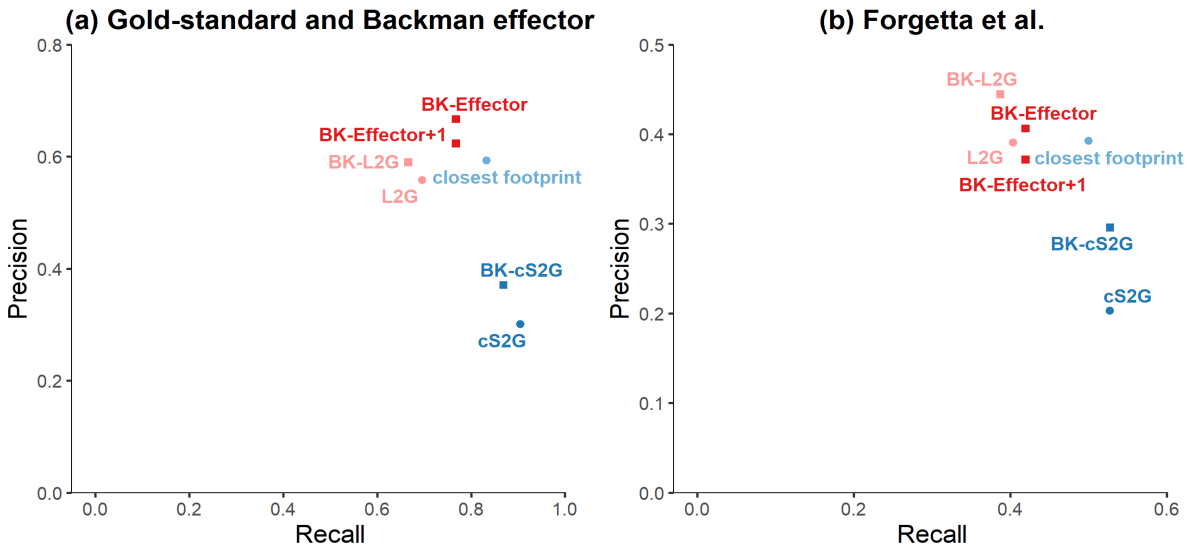

Figure 5: **Comparisons of different locus-to-gene prioritization methods.** Precision vs. Recall is shown for several representative methods including closest footprint, cS2G, L2G, BIGKnock Effector genes (BK-Effector), BIGKnock Effector genes and genes at BIGKnock significant loci with only one significant gene (BK-Effector+1), as well as combination of BIGKnock and cS2G (BK-cS2G) and L2G (BK-L2G). (a) Gold standard and Backman effector dataset including 138 positive genes at BIGKnock significant loci. The negative genes include 2,013 genes located within the 1Mb loci containing the 138 positive genes; (b) Forgetta et al. gene set including 62 positive genes at BIGKnock significant loci. The negative genes include 973 genes located at 1Mb loci containing the 62 positive genes.

Table 1: Selected loci for binary and quantitative traits. The number of significant genes per locus for GeneScan3D, BIGKnock, and BIGKnock significant genes are shown. The putative causal gene is shown in boldface font.

BIGKnock locus-trait	position (hg19)	# GeneScan3D	# BIGKnock	BIGKnock genes
BRAP-CAD	12: 111,986,818-112,986,818	11	2	***ALDH2***,*BRAP*
UBE2Z-Asthma	17: 46,948,346-47,948,346	11	2	***NGFR***,*UBE2Z*
AGPAT1-T2D	6: 31,618,085-32,618,085	24	1	***AGPAT1***
MARCHF5-T2D	10: 93,966,910-94,966,910	5	1	***MARCHF5***
ASGR1-Cholesterol	17: 6,569,412-7,569,412	43	2	***ASGR1***,*CD68*
SLC39A8-BP-Diastolic	4: 102,688,709-103,688,709	3	1	***SLC39A8***
DBH-BP-Diastolic	9: 136,001,756-137,001,756	6	3	*ADAMTS13*,***DBH***,*SARDH*
ANGPTL4-Cholesterol	19: 7,951,937-8,951,937	9	1	***ANGPTL4***
RAB11A-Neutrophil count	15: 65,544,465-66,544,465	8	1	***RAB11A***
ZHX3-Calcium	20: 39,455,078-40,455,078	6	1	***ZHX3***
PPARG-LDL cholesterol	3: 11,739,931-12,739,931	6	1	***PPARG***
POLDIP2-LDL cholesterol	17: 26,194,861-27,194,861	22	3	***POLDIP2***,*SLC13A2*,*TMEM199*

Supplemental Material

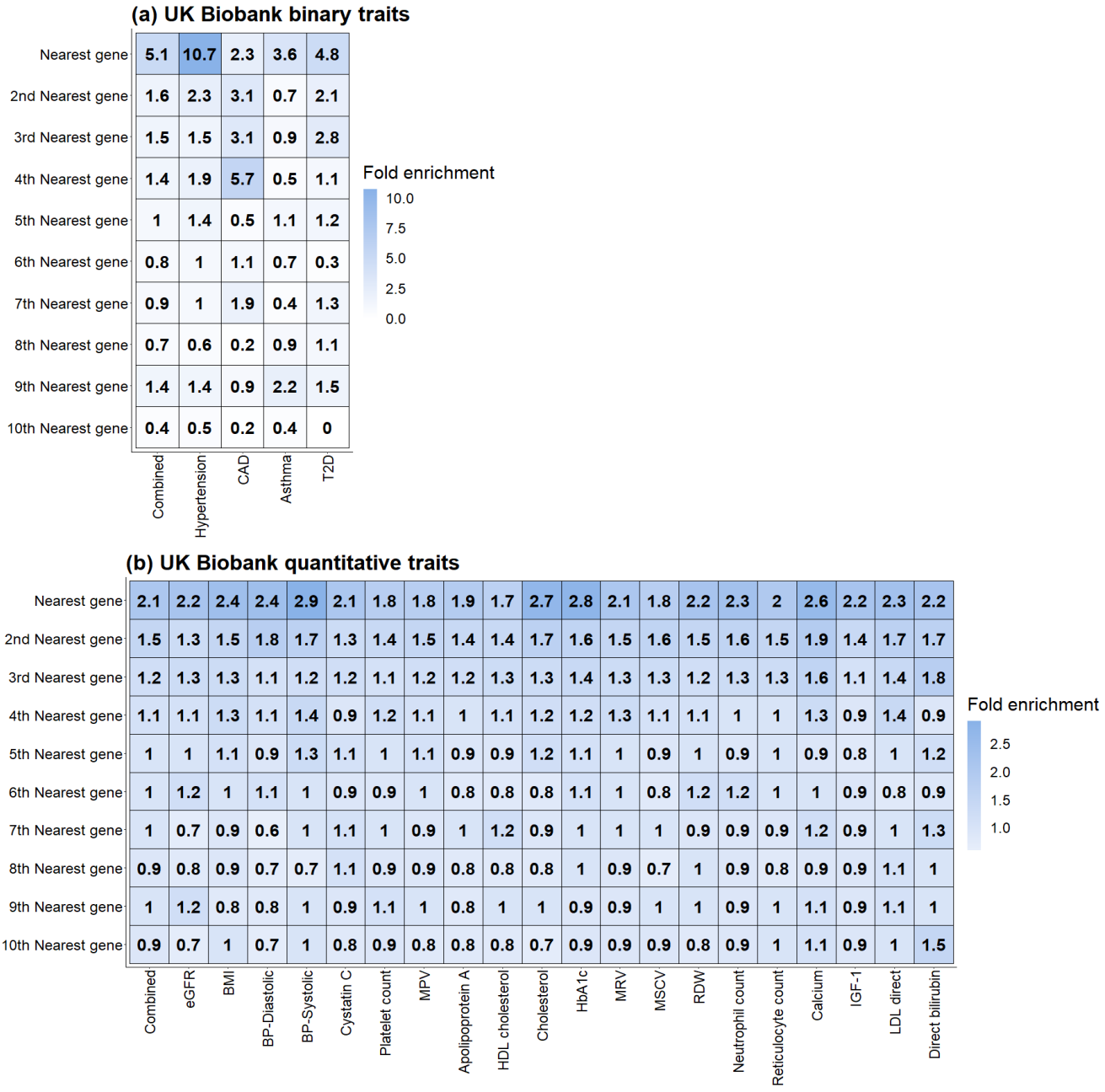

Figure S1: **Enrichment of BIGKnock significant genes among genes closest to the lead GWAS variant at shared loci between BIGKnock and GeneScan3D**. Enrichment of BIGKnock significant genes for (a) the combined five UK Biobank binary traits and each binary trait separately: Hypertension, Coronary artery disease (CAD), Asthma, Cataract and Type 2 diabetes (T2D); and (b) the combined twenty UK Biobank quantitative traits and each quantitative trait separately: eGFR, BMI, BP-Diastolic, BP-Systolic, Cystatin C, Platelet count, MPV, Apolipoprotein A, HDL cholesterol, Cholesterol, HbA1c, MRV, MSCV, RDW, Neutrophil count, Reticulocyte count, Calcium, IGF-1, LDL direct and Direct bilirubin.

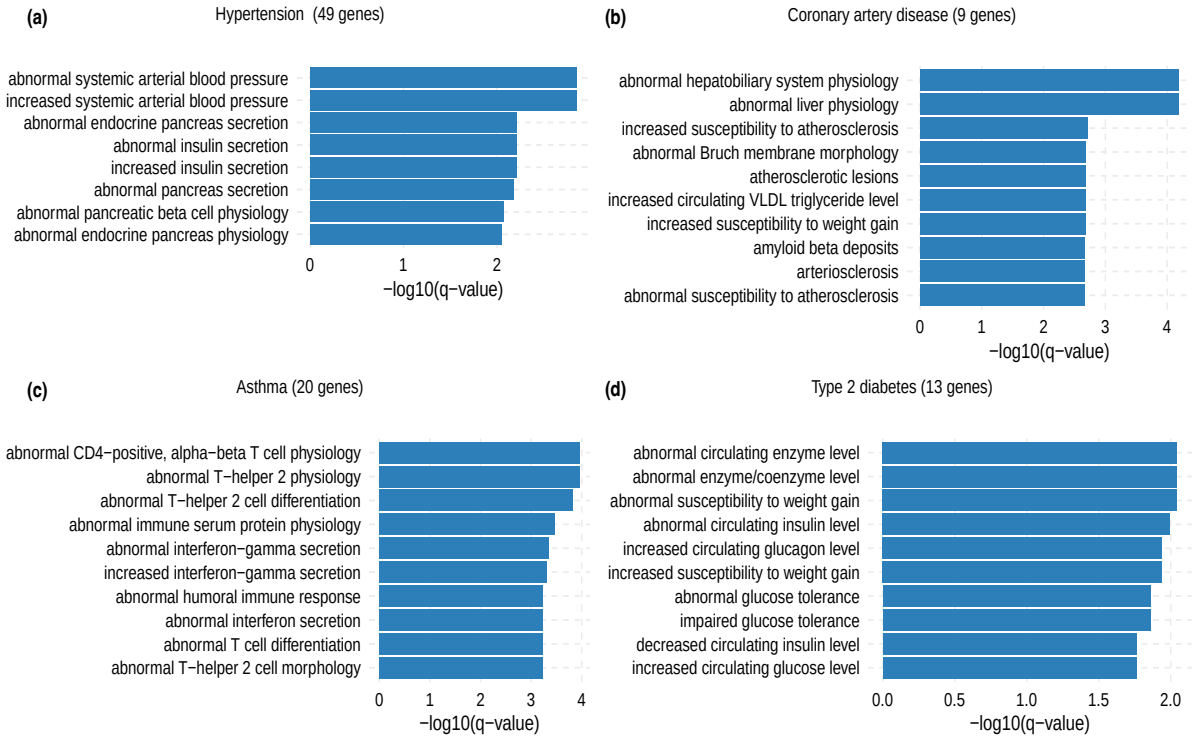

Figure S2: **Mouse phenotype enrichment analyses for four binary traits in ToppFun.** The top 10 mouse phenotypes in terms of q-value are shown for each trait. The number of effector BIGKnock genes used in these analyses is indicated for each trait.

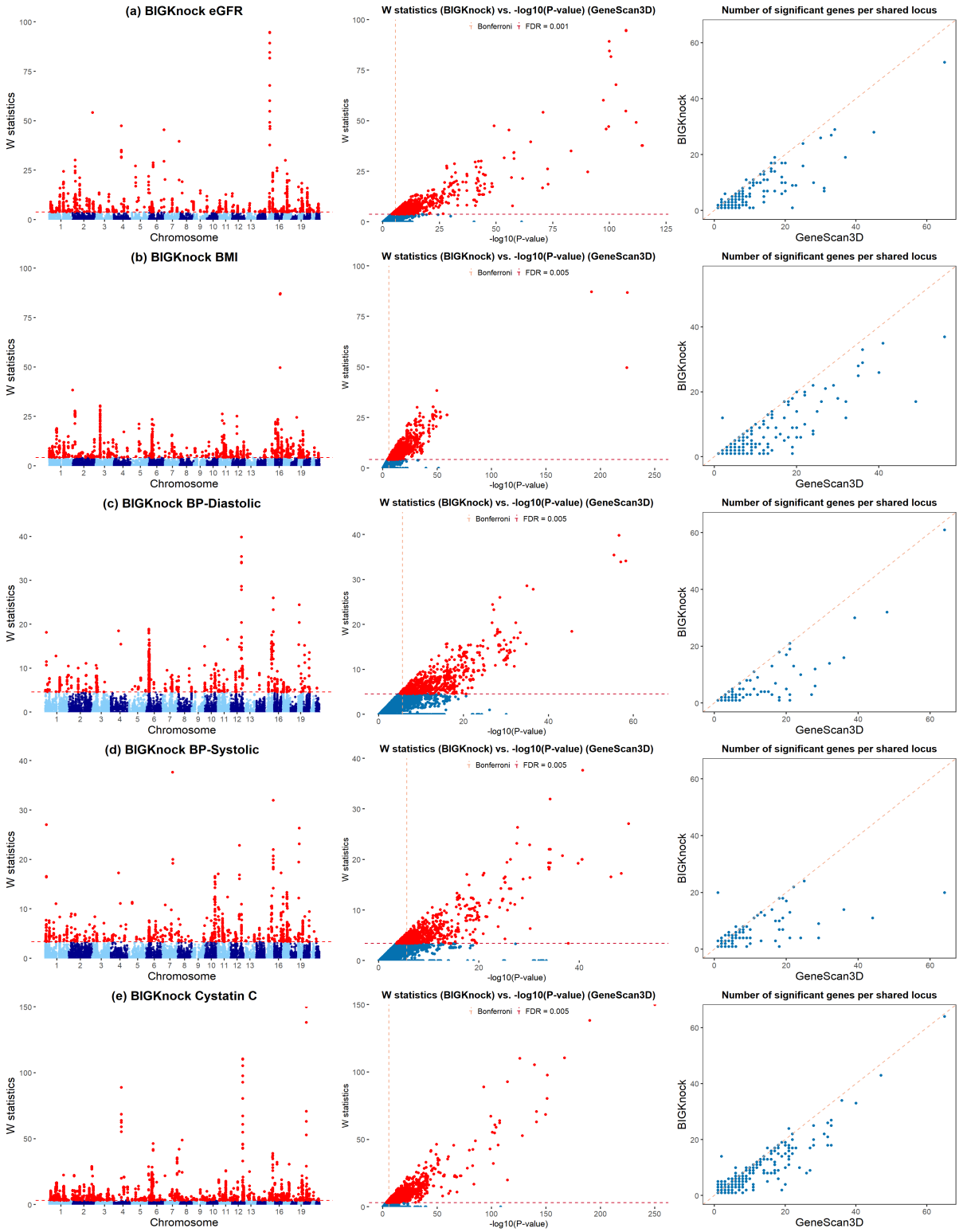

Figure S3: **Applications to five UK Biobank quantitative traits**. a-e, Manhattan plots for BIGKnock, Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D), and Scatter plot of the number of significant genes per locus between conventional GeneScan3D and BIGKnock are shown for (a) eGFR, (b) BMI, (c) BP-Diastolic, (d) BP-Systolic, and (e) Cystatin C. The dashed lines in the left and middle panels show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

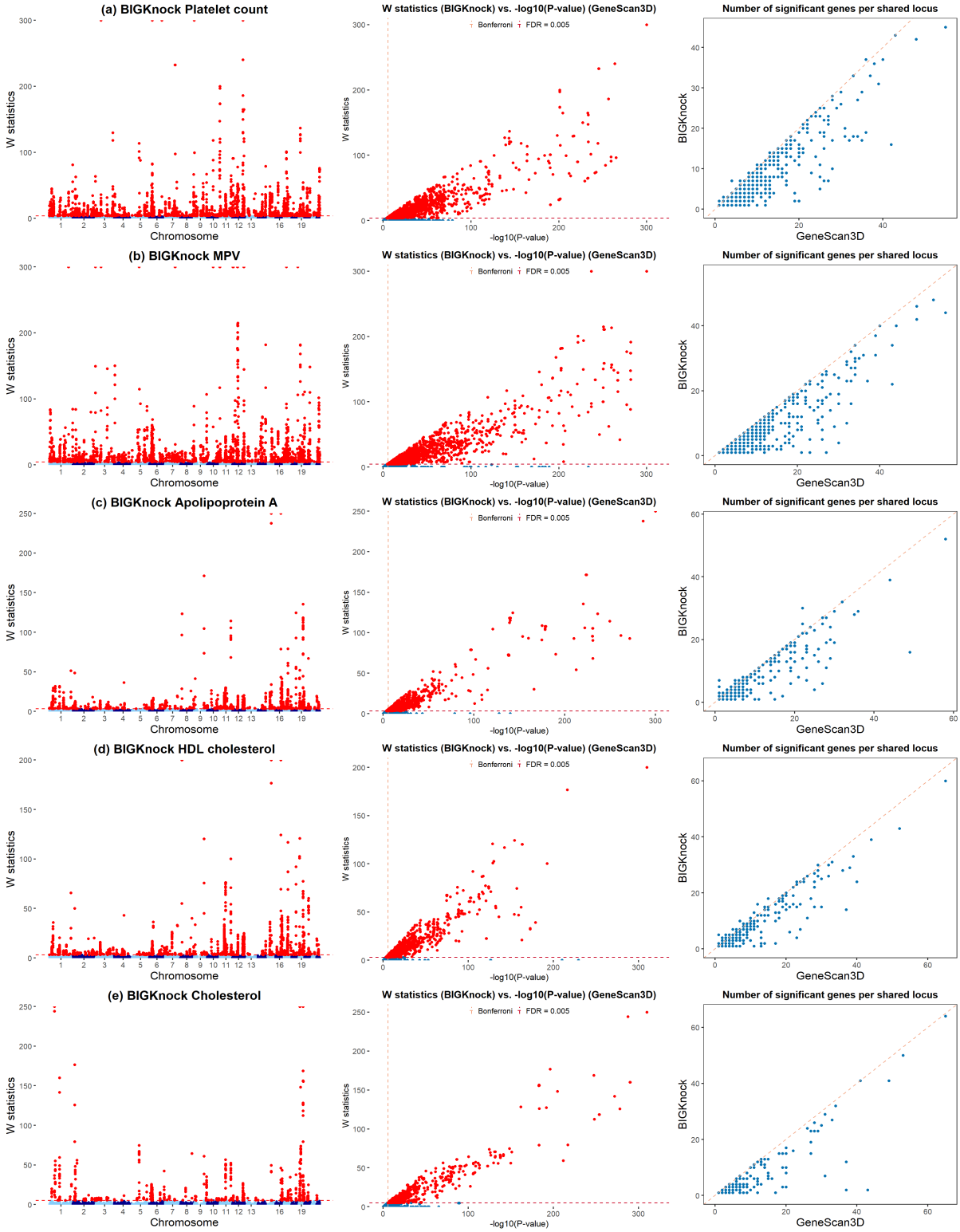

Figure S4: **Applications to five UK Biobank quantitative traits (2)**. a-e, Manhattan plots for BIGKnock, Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D), and Scatter plot of the number of significant genes per locus between conventional GeneScan3D and BIGKnock are shown for (a) Platelet count, (b) MPV, (c) Apolipoprotein A, (d) HDL cholesterol, and (e) Cholesterol. The dashed lines in the left and middle panels show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

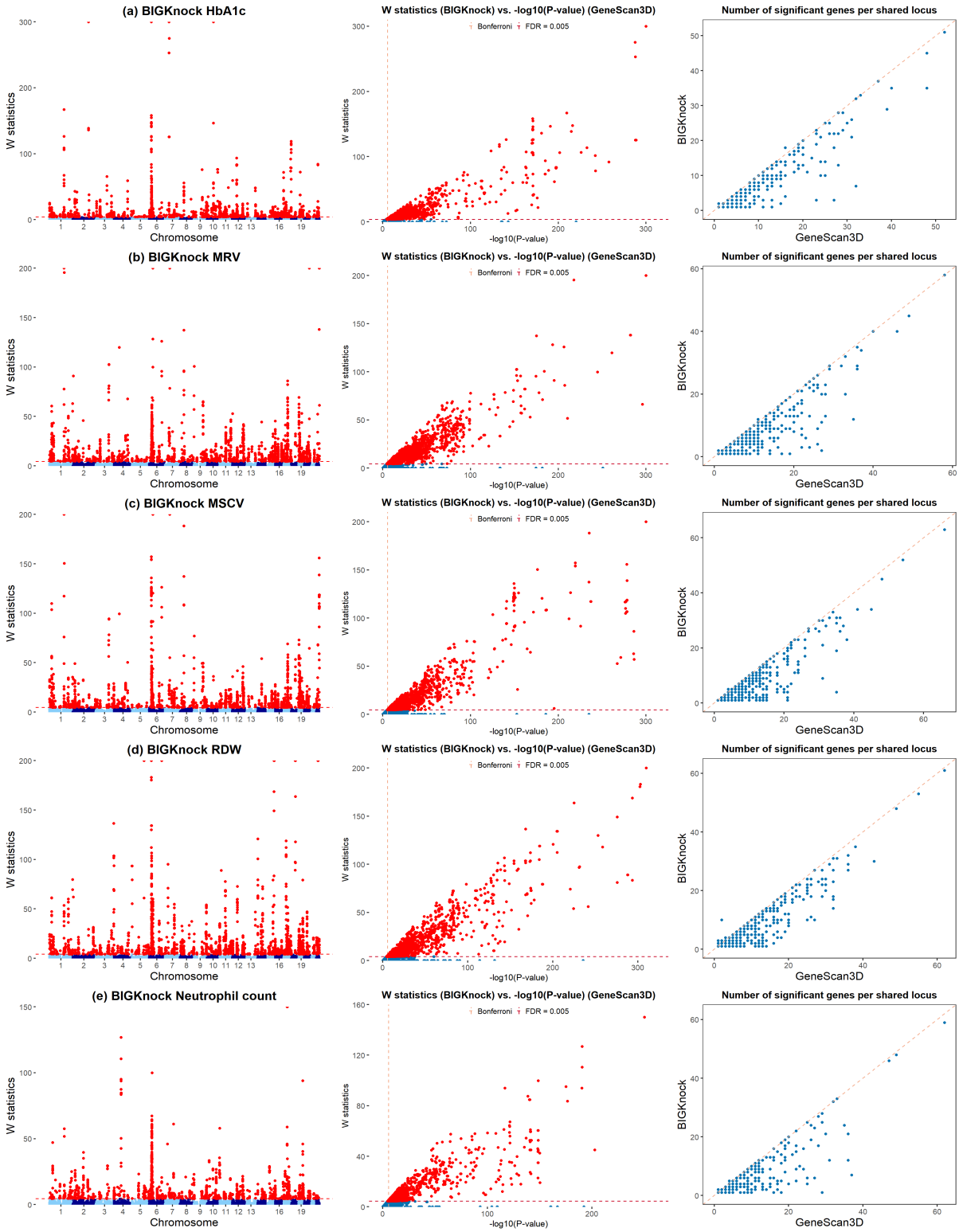

Figure S5: **Applications to five UK Biobank quantitative traits (3)**. a-e, Manhattan plots for BIGKnock, Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D), and Scatter plot of the number of significant genes per locus between conventional GeneScan3D and BIGKnock are shown for (a) HbA1c, (b) MRV, (c) MSCV, (d) RDW, and (e) Neutrophil count. The dashed lines in the left and middle panels show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

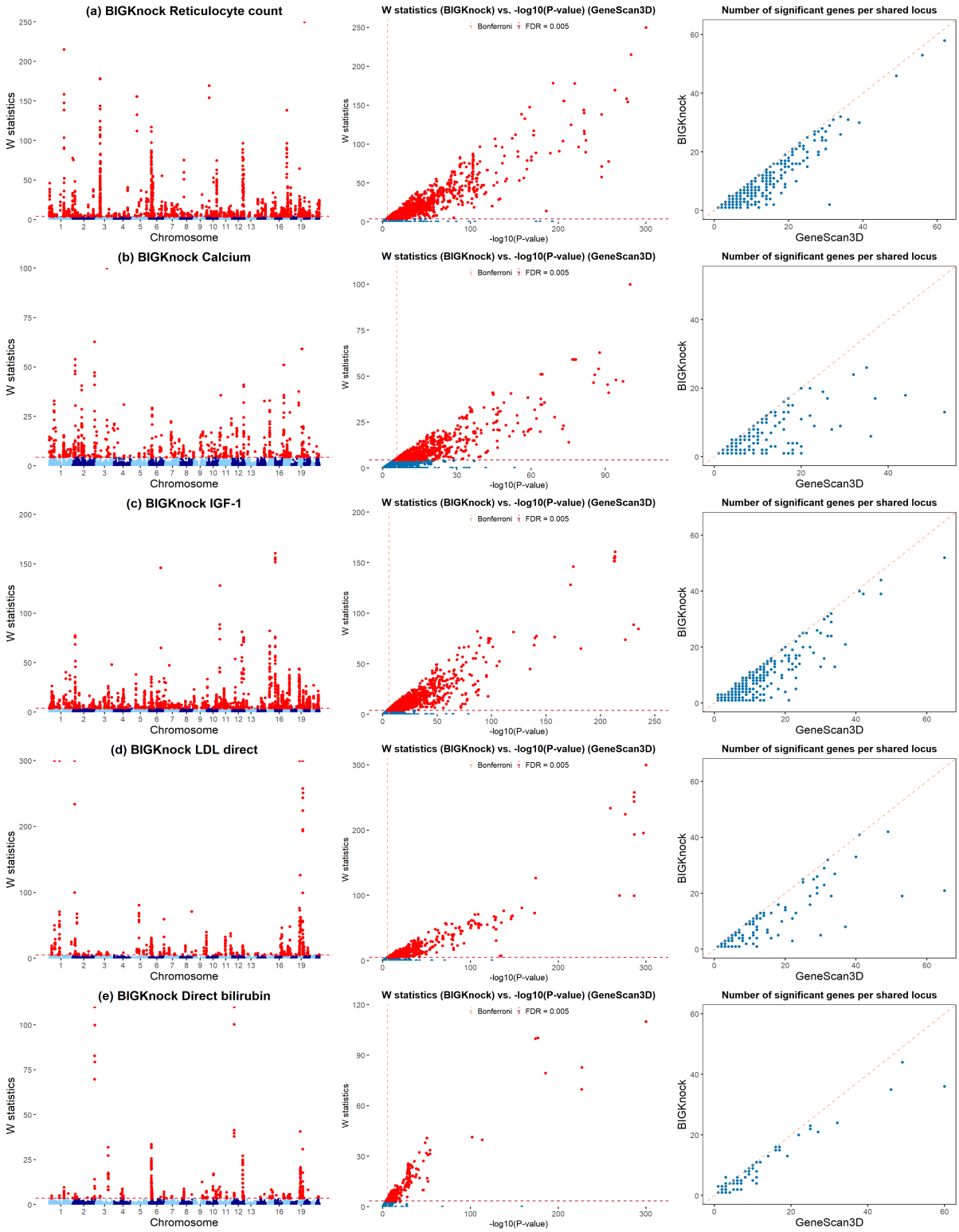

Figure S6: **Applications to five UK Biobank quantitative traits (4)**. a-e, Manhattan plots for BIGKnock, Scatter plot of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D), and Scatter plot of the number of significant genes per locus between conventional GeneScan3D and BIGKnock are shown for (a) Reticulocyte count, (b) Calcium, (c) IGF-1, (d) LDL direct, and (e) Direct bilirubin. The dashed lines in the left and middle panels show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

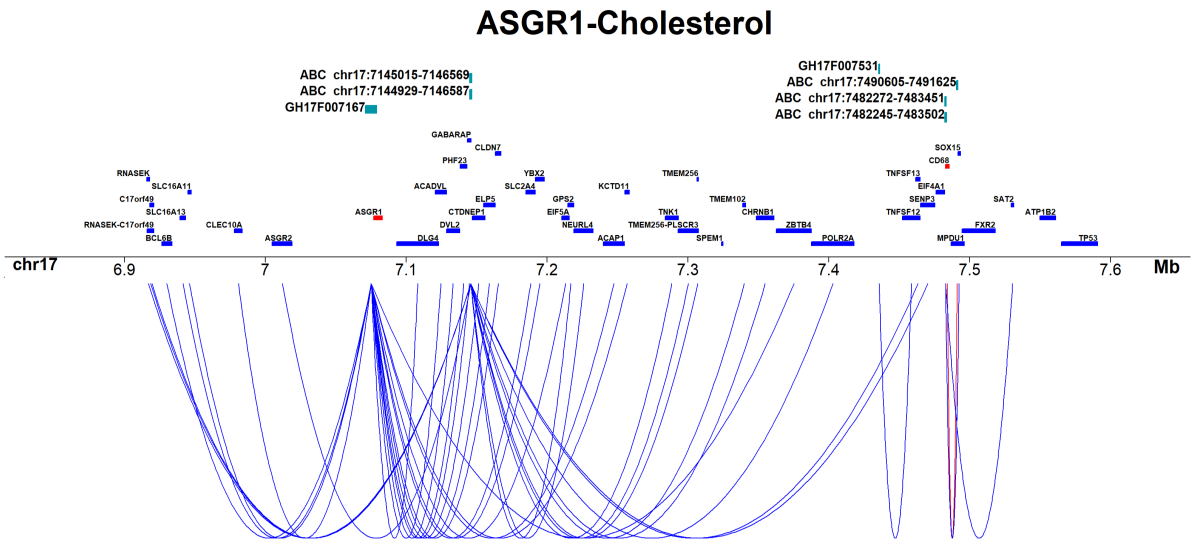

Figure S7: **Visualization of gene-enhancer interactions of significant genes at the ASGR1-Cholesterol locus.** Gene-enhancer interactions for 43 GeneScan3D significant genes at the ASGR1-Cholesterol locus, with the two BIGKnock significant genes (*ASGR1* and *CD68*) shown in red. The interaction between BIGKnock significant gene *CD68* and ABC enhancer chr17:7,490,605-7,491,625 is shown in red; other 35 gene-enhancer links are shown in blue (See Supplementary Table 26).

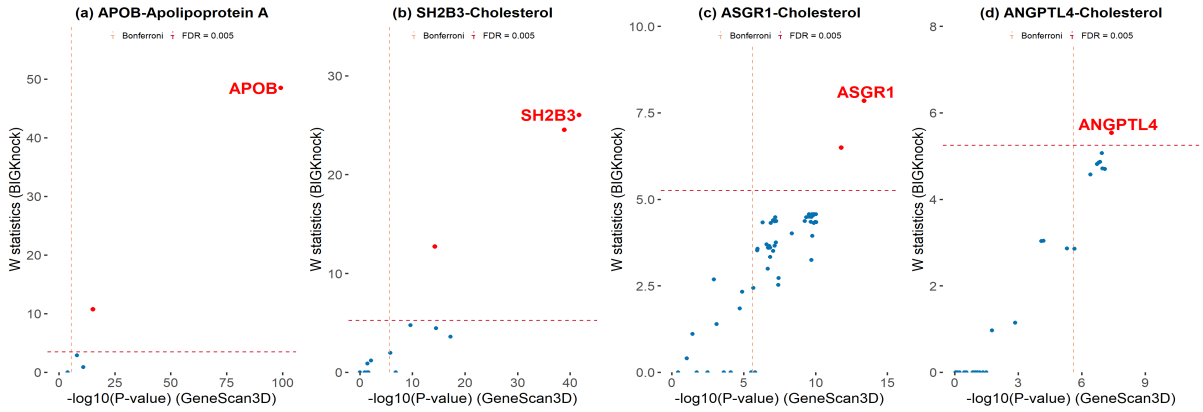

Figure S8: **Selected loci that pinpoint effector genes identified by Backman et al.[30]** Scatter plots of W knockoff statistics (BIGKnock) vs. $-\log_{10}$(p value) (GeneScan3D) for 4 selected loci that pinpoint effector genes identified by Backman et al.[30]. The effector genes are labeled in red. The dashed lines show the significance thresholds defined by Bonferroni correction (for p-values) and by false discovery rate (FDR; for W statistic).

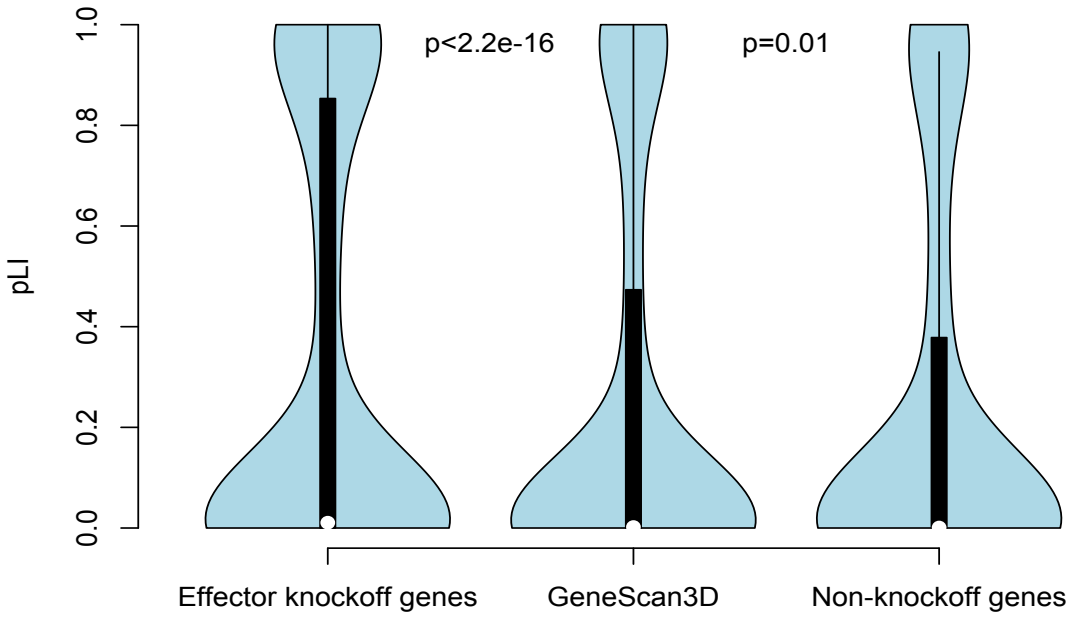

Figure S9: **pLI for effector BIGKnock genes, GeneScan3D significant genes and Non-knockoff genes.** Significant genes are aggregated across four binary traits and twenty quantitative traits. Non-knockoff genes are not significant under BIGKnock for any of the binary/quantitative traits considered. p-values are from a nonparametric Kolmogorov-Smirnov test comparing pLI scores for effector BIGKnock genes vs. GeneScan3D significant genes, and GeneScan3D significant genes vs. Non-knockoff genes.

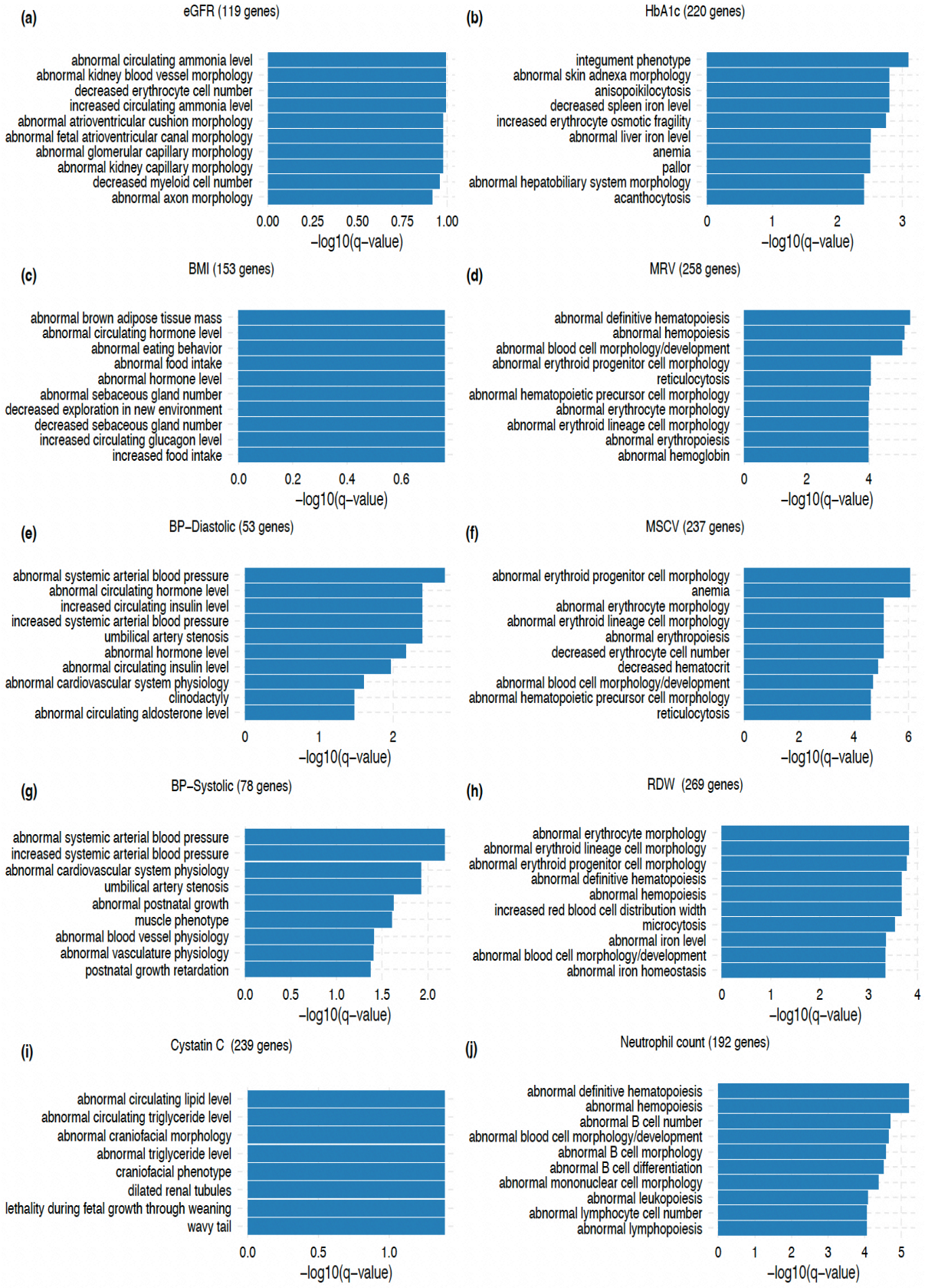

Figure S10: **Mouse phenotype enrichment analyses for 10 quantitative traits in ToppFun.** The top 10 mouse phenotypes in terms of q-value are shown for each trait. The number of effector BIGKnock genes used in these analyses are indicated for each trait.

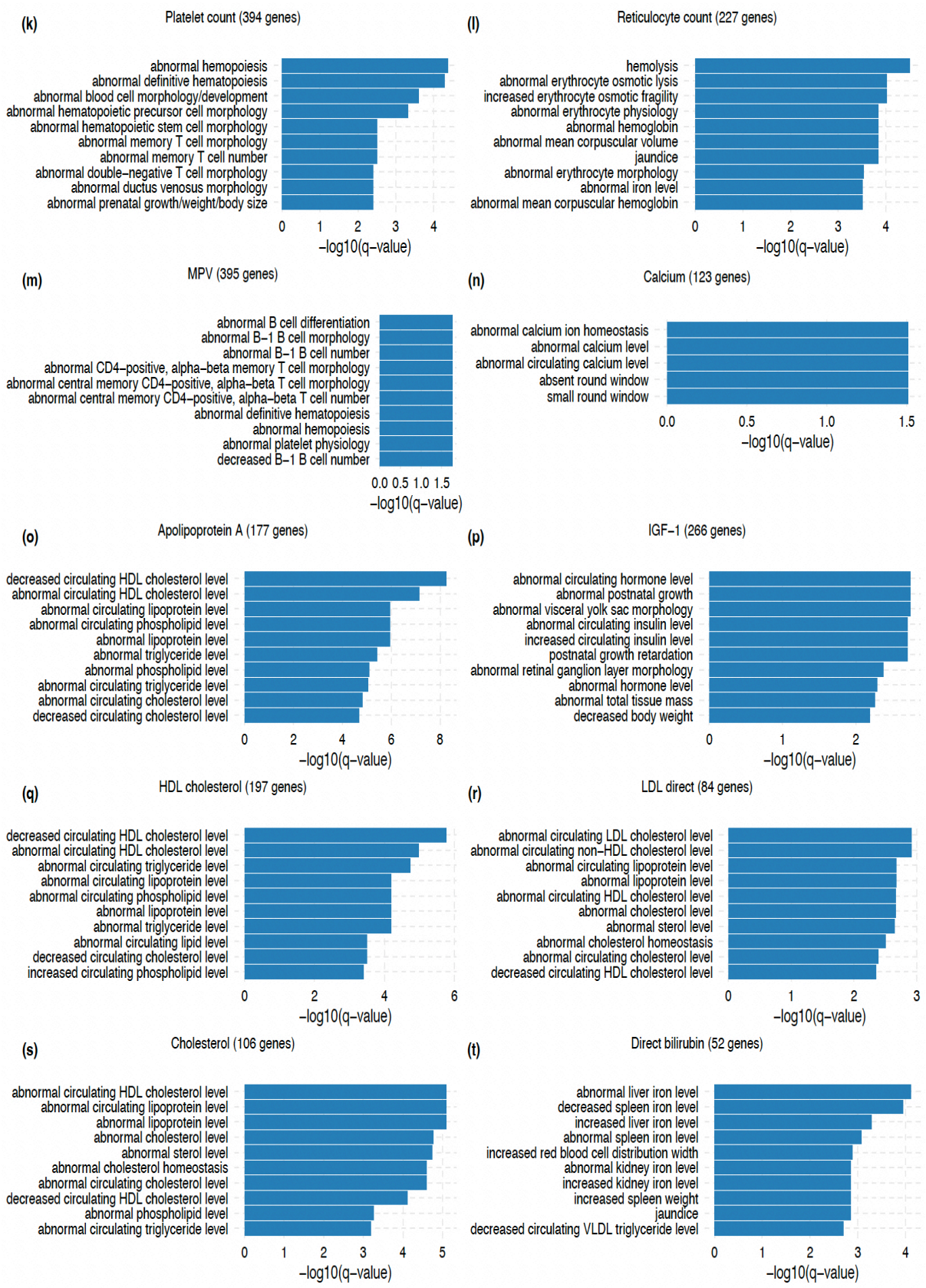

Figure S11: **Mouse phenotype enrichment analyses for 10 quantitative traits in ToppFun.** The top 10 mouse phenotypes in terms of q-value are shown for each trait. The number of effector BIGKnock genes used in these analyses are indicated for each trait.

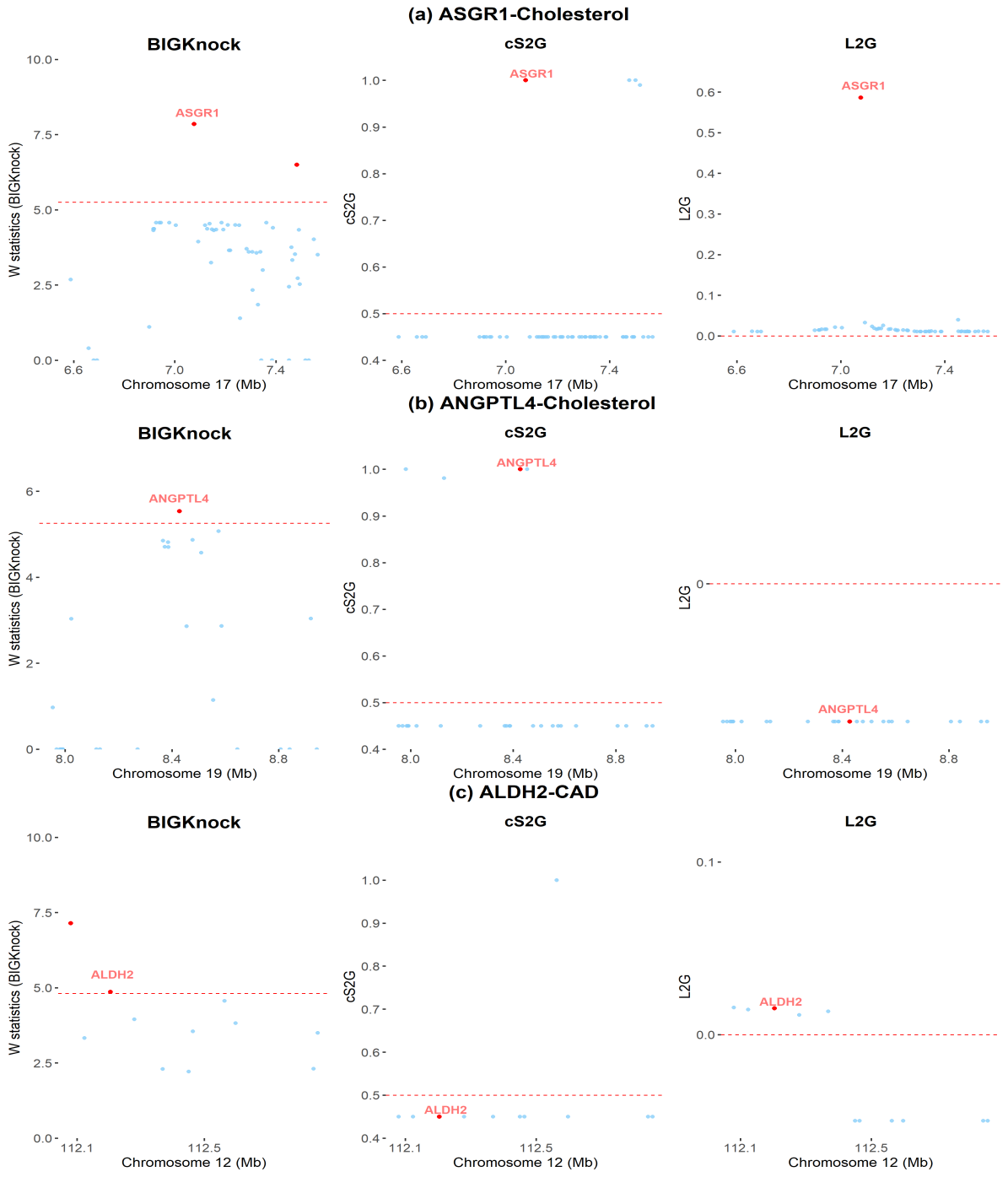

Figure S12: **BIGKnock, cS2G and L2G results at three loci containing known causal genes (ASGR1-Cholesterol, ANGPTL4-Cholesterol and ALDH2-CAD).** *W* knockoff statistics, cS2G scores, and L2G scores are shown for genes at the 1Mb loci containing known causal genes. The putative causal gene at each locus is labeled.

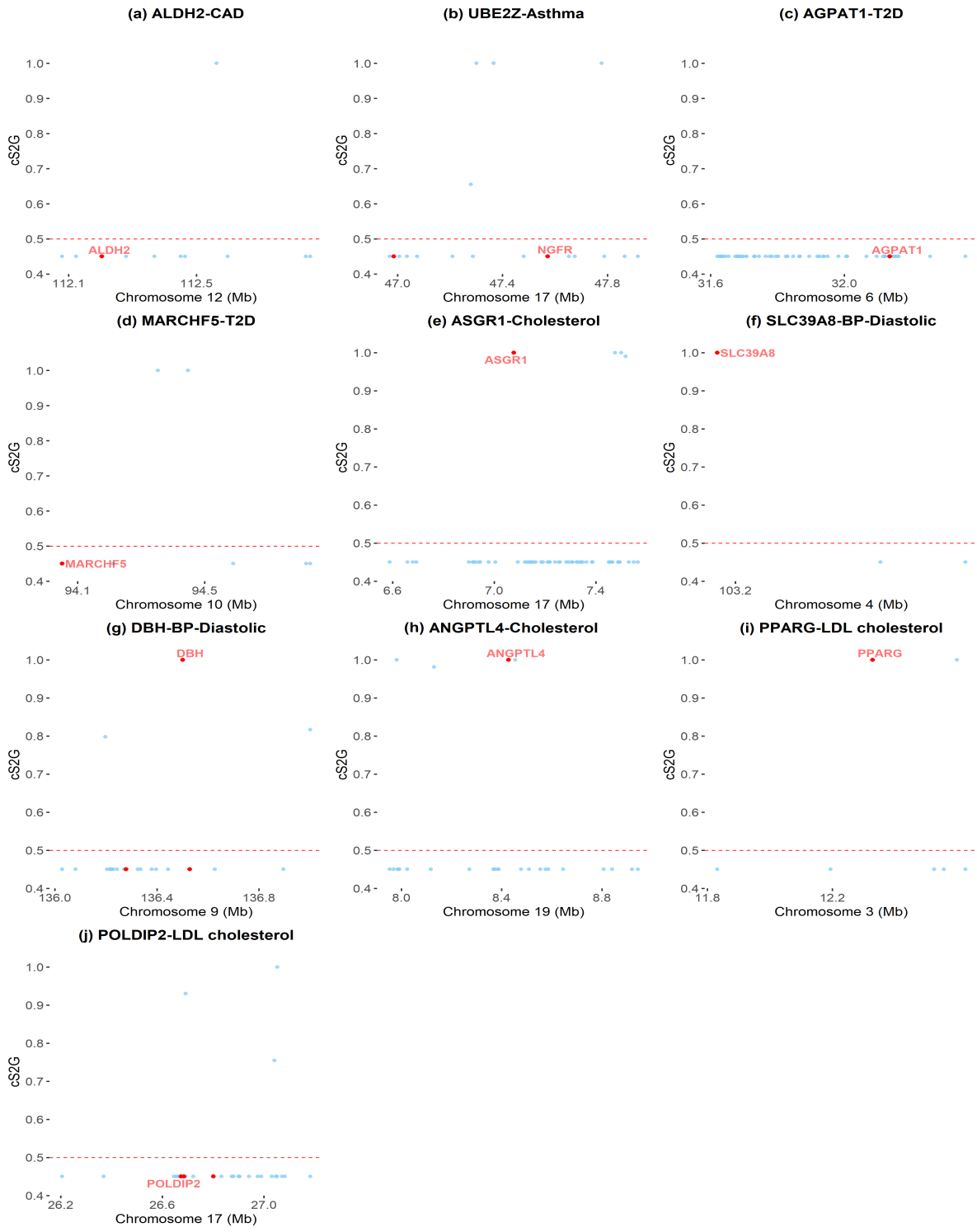

Figure S13: **cS2G scores of putative causal genes at selected loci for UK Biobank binary traits and quantitative traits**. cS2G scores for genes at selected loci for (a) CAD, (b) Asthma, (c-d) Type 2 diabetes (T2D), (e,h) Cholesterol, (f-g) BP-Diastolic and (i-j) LDL cholesterol. Loci are named according to the most significant gene in BIGKnock. The dashed line corresponds to the recommended threshold (0.5) for cS2G. Genes which do not have a cS2G score are shown just below the 0.5 threshold. The labeled gene corresponds to the putative causal gene at the locus as discussed in the Results section. Red dots correspond to genes that are significant under BIGKnock.

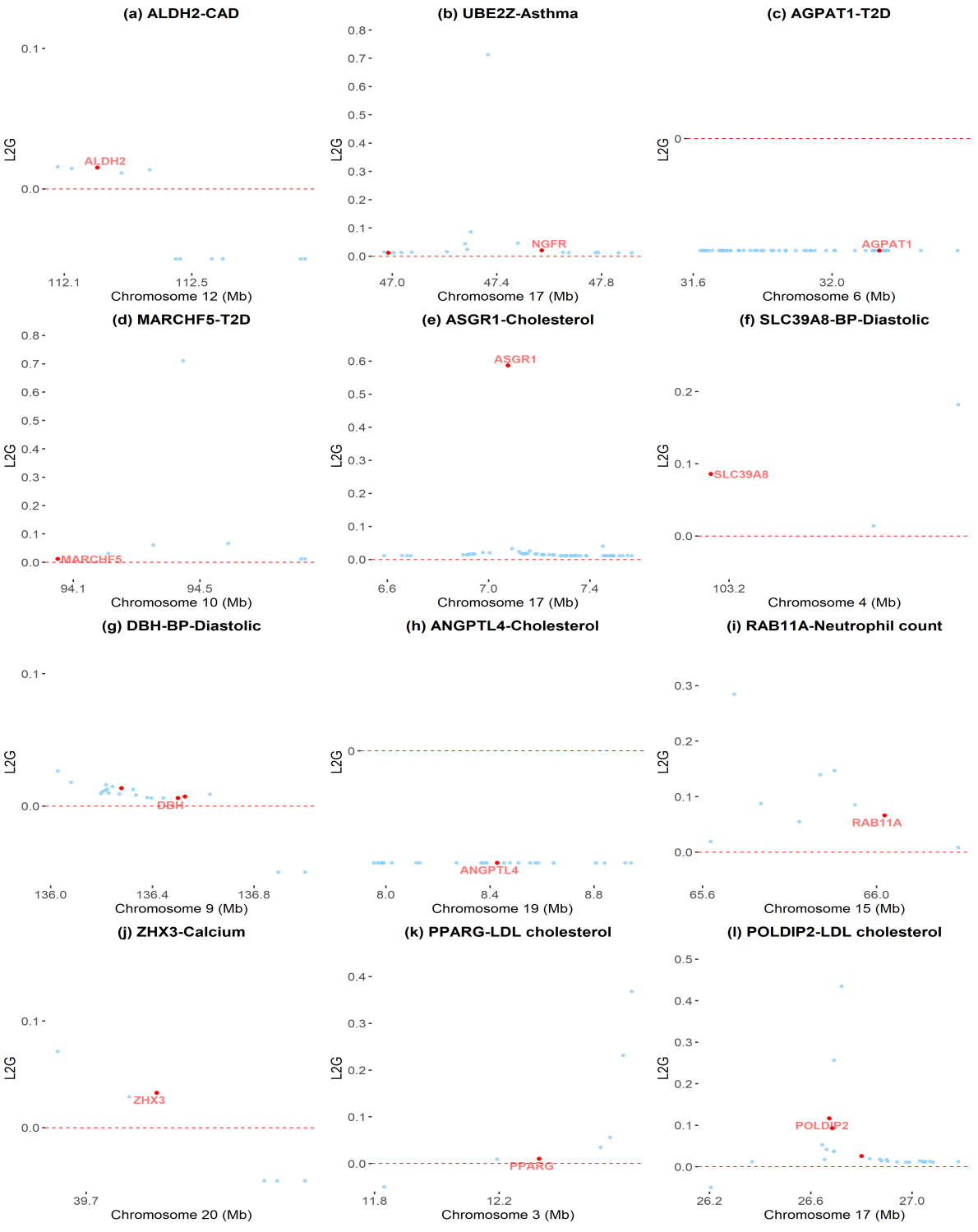

Figure S14: **L2G scores of putative causal genes at selected loci for UK Biobank binary traits and quantitative traits**. L2G scores for genes at selected loci for (a) CAD, (b) Asthma, (c-d) Type 2 diabetes (T2D), (e,h) Cholesterol, (f-g) BP-Diastolic, (i) Neutrophil count, (j) Calcium and (k-l) LDL cholesterol. Loci are named according to the most significant gene in BIGKnock. The dashed line corresponds to the recommended threshold (0) for L2G. Genes which do not have a L2G score are shown just below the 0 threshold. The labeled gene corresponds to the putative causal gene at the locus as discussed in the Results section. Red dots correspond to genes that are significant under BIGKnock.

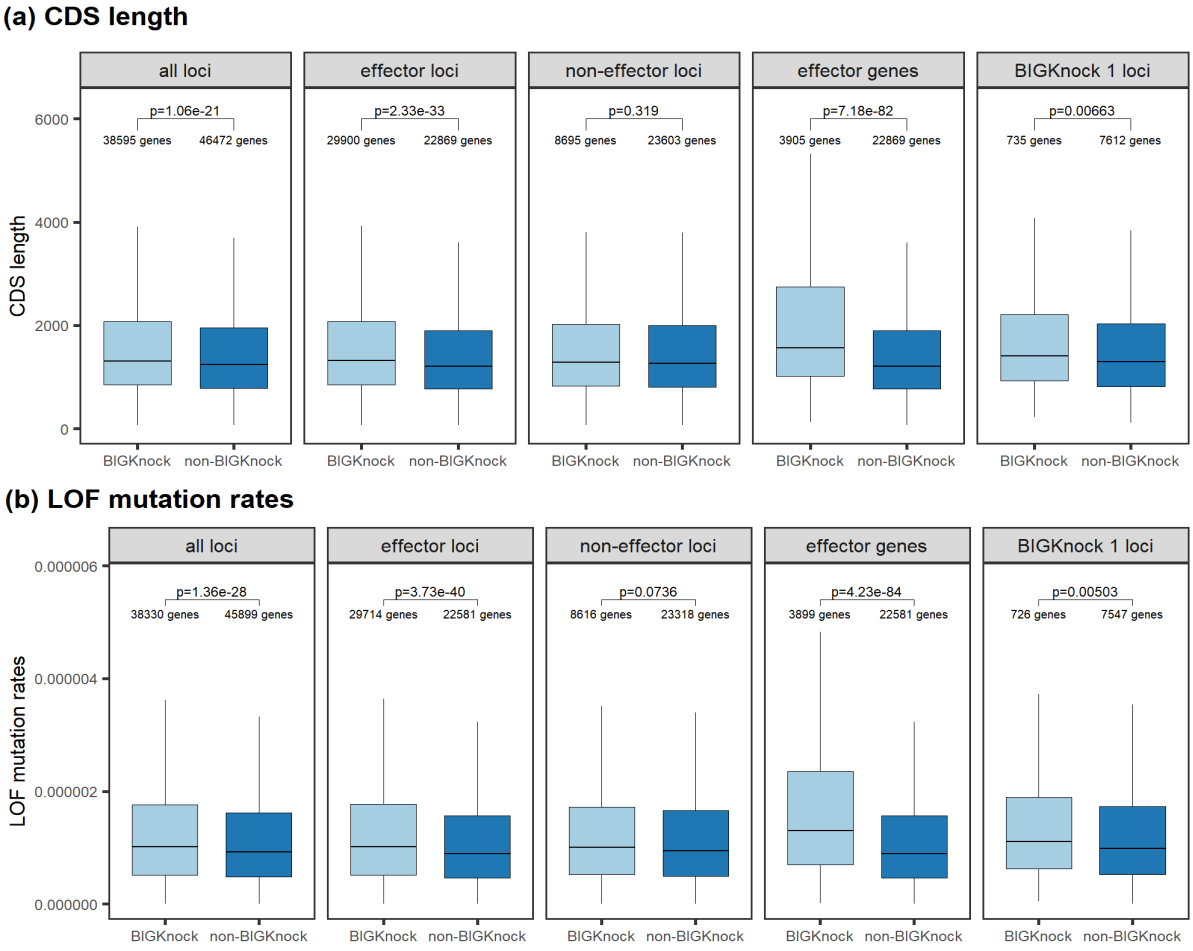

Figure S15: **Characteristics of prioritized genes.** (a) CDS length and (b) LOF mutation rates for BIGKnock significant genes and genes that are not significant using the BIGKnock test at all loci, effector BIGKnock loci (loci that have effector genes), non-effector loci, effector BIGKnock genes vs. genes that are not significant using the BIGKnock test, and BIGKnock 1 loci (loci where BIGKnock prioritizes only one gene). The number of genes in each group are shown, along with p-values from a Kolmogorov-Smirnov test.

Table S1: Summary of selected UK Biobank binary and quantitative traits.

Binary traits	*n*	# cases/# controls (case:control ratio)
Hypertension	404,265	90,714/313,551 (1:3.5)
Coronary artery disease	404,186	35,400/368,786 (1:10.4)
Asthma	395,745	29,934/365,811 (1:12)
Type 2 diabetes	403,490	21,964/381,526 (1:17)
Quantitative traits	*n*	
estimated glomerular filtration rate (eGFR)	386,253	
Body Mass Index (BMI)	403,979	
Diastolic Blood Pressure Automated Reading (BP-Diastolic)	378,531	
Systolic Blood Pressure Automated Reading (BP-Systolic)	378,522	
Cystatin C	386,398	
Platelet count	393,205	
Mean platelet volume (MPV)	393,200	
Apolipoprotein A	351,755	
HDL cholesterol	353,744	
Cholesterol	386,452	
Glycated haemoglobin (HbA1c)	386,300	
Mean reticulocyte volume (MRV)	386,844	
Mean sphered cell volume (MSCV)	386,844	
Red blood cell (erythrocyte) distribution width (RDW)	393,189	
Neutrophil count	392,512	
Reticulocyte count	386,843	
Calcium	353,787	
IGF-1	384,356	
LDL direct (LDL cholesterol)	385,728	
Direct bilirubin	328,800	

Table S2: The number of significant loci/genes associated with each binary trait for GeneScan3D and BIGKnock, and the number of loci shared between them.

Binary traits	**GeneScan3D**	**BIGKnock**	**Shared loci**	**FDR threshold**
Hypertension	80 loci/ 513 genes	66 loci/ 352 genes	62 loci	0.01
Coronary artery disease	25 loci/ 187 genes	15 loci/ 108 genes	15 loci	0.01
Asthma	27 loci/ 337 genes	26 loci/ 253 genes	25 loci	0.01
Type 2 diabetes	35 loci/ 172 genes	29 loci/ 88 genes	25 loci	0.05

Table S3: The number of significant loci/genes associated with each quantitative trait for GeneScan3D and BIGKnock, and the number of loci shared between them.

Quantitative traits	**GeneScan3D**	**BIGKnock**	**Shared loci**	**FDR threshold**
eGFR	270 loci/ 1,483 genes	232 loci/ 1,010 genes	218 loci	0.001
BMI	428 loci/ 2,767 genes	270 loci/ 1,300 genes	265 loci	0.005
BP-Diastolic	166 loci/ 1,498 genes	103 loci/ 677 genes	102 loci	0.005
BP-Systolic	151 loci/ 1,208 genes	146 loci/ 865 genes	131 loci	0.005
Cystatin C	405 loci/ 2,550 genes	427 loci/ 2,092 genes	366 loci	0.005
Platelet count	717 loci/ 5,242 genes	627 loci/ 3,768 genes	621 loci	0.005
MPV	765 loci/ 5,293 genes	640 loci/ 3,496 genes	638 loci	0.005
Apolipoprotein A	300 loci/ 2,362 genes	291 loci/ 1,815 genes	271 loci	0.005
HDL cholesterol	337 loci/ 2,412 genes	340 loci/ 1,974 genes	304 loci	0.005
Cholesterol	216 loci/ 1,812 genes	153 loci/ 998 genes	153 loci	0.005
HbA1c	329 loci/ 3,438 genes	293 loci/ 2,392 genes	289 loci	0.005
MRV	474 loci/ 4,139 genes	389 loci/ 2,661 genes	389 loci	0.005
MSCV	483 loci/ 4,104 genes	373 loci/ 2,479 genes	371 loci	0.005
RDW	469 loci/ 4,176 genes	405 loci/ 2,896 genes	399 loci	0.005
Neutrophil count	401 loci/ 3,273 genes	309 loci/ 1,865 genes	309 loci	0.005
Reticulocyte count	417 loci/ 3,447 genes	354 loci/ 2,367 genes	350 loci	0.005
Calcium	266 loci/ 1,940 genes	215 loci/ 963 genes	214 loci	0.005
IGF-1	520 loci/ 3,644 genes	422 loci/ 2,324 genes	416 loci	0.005
LDL direct	174 loci/ 1,601 genes	122 loci/ 902 genes	122 loci	0.005
Direct bilirubin	78 loci/ 654 genes	84 loci/ 547 genes	73 loci	0.005

Table S4: Selected loci that pinpoint effector genes identified by Backman et al.[30].

BIGKnock locus-trait	position (hg19)	# GeneScan3D	# BIGKnock	BIGKnock genes	Effector gene
APOB-Apolipoprotein A	2:20,731,524-21,731,524	4	2	*APOB,TDRD15*	*APOB*
SH2B3-Cholesterol	12:110,868,171-111,868,171	8	3	*FAM109A,PPTC7,SH2B3*	*SH2B3*
ASGR1-Cholesterol	17:6,569,412-7,569,412	43	2	*ASGR1,CD68*	*ASGR1*
ANGPTL4-Cholesterol	19:7,951,937-8,951,937	9	1	*ANGPTL4*	*ANGPTL4*