Bioinformatics, 2020, 1–3
doi: 10.1093/bioinformatics/btaa883
Advance Access Publication Date: 16 October 2020
Applications Note

Genome analysis

vSampler: fast and annotation-based matched variant sampling tool

Dandan Huang[1,2,†], Zhao Wang[2,†], Yao Zhou[2,†], Qian Liang[2], Pak Chung Sham[3], Hongcheng Yao[4,*] and Mulin Jun Li [1,2,*]

[1]Department of Epidemiology and Biostatistics, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin 300070, China, [2]Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China, [3]Department of Psychiatry and [4]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR 999077, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on March 17, 2020; revised on July 28, 2020; editorial decision on September 28, 2020; accepted on September 30, 2020

Abstract

Summary: Sampling of control variants having matched properties with input variants is widely used in enrichment analysis of genome-wide association studies/quantitative trait loci and negative data construction for pathogenic/regulatory variant prediction methods. Spurious enrichment results because of confounding factors, such as minor allele frequency and linkage disequilibrium pattern, can be avoided by calibration of statistical significance based on matched controls. Here, we presented vSampler which can generate sets of randomly drawn variants with comprehensive choices of matching properties, such as tissue/cell type-specific epigenomic features. Importantly, the development of a novel data structure and sampling algorithms for vSampler makes it significantly fast than existing tools.

Availability and implementation: vSampler web server and local program are available at http://mulinlab.org/vsampler.

Contact: hongchengyaonk@gmail.com or mulinli@connect.hku.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the past decade, genome-wide association studies (GWASs) and quantitative trait loci (QTLs) mapping have identified large amounts of genetic associations that help to elucidate the underlying mechanism of complex traits and diseases (Visscher *et al.*, 2017). Based on GWAS/QTL identified variants, enrichment analysis is often utilized to pinpoint related biological pathways (Jia *et al.*, 2010) or functional annotations (Farh *et al.*, 2015). Besides, pathogenic/regulatory variant prediction methods would also use these associated or fine-mapped variants for training/test data construction (Li *et al.*, 2017; Zhang *et al.*, 2019; Zhou and Troyanskaya, 2015). During these studies, control variants sampling which accounts for potential confounders is commonly employed. However, simple random sampling of control variants would often lead to spurious enrichment results or biased training/test dataset, because the GWAS/QTL identified variants and the random controls could differ in confounding factors such as allele frequency, pattern of linkage disequilibrium (LD) and other genomic features. Existing tool SNPsnap can only control for a few factors and cannot deal with large-scale inputs (Pers *et al.*, 2015), which limits its broader usage in the era of big genomic data.

Here, we presented a fast, scalable and versatile tool, vSampler, for sampling matched sets of variants both as a web server and as a local program (http://mulinlab.org/vsampler or https://github.com/mulinlab/vSampler). Given input variants, vSampler can randomly draw control variants with eight optional matching properties by a novel data structure and sampling algorithm. These matched random controls could be used to construct null distribution in enrichment/colocalization analysis to estimate the significance of statistical tests empirically or serve as negative training/test data for pathogenic/regulatory variant prediction models. Compared with existing methods, vSampler runs significantly faster (up to two orders of magnitude), supports both single-nucleotide polymorphisms (SNPs) and small insertions and deletions (Indels) and provides comprehensive context-specific functional annotations as matching properties.

2 Database construction and data structure

vSampler extracted biallelic variants from 1000 Genomes project and stored them in an annotation database and a sampling database. The annotation database contains all variants and is used for input variant annotation while the sampling database contains only

variants with minor allele frequency (MAF) > 0.01 and is used as the pool of control variants. For each variant, the following eight properties were computed (see Supplementary Notes for details): MAF, distance to closest transcript start site (DTCT), gene density (number of nearby genes), number of variants in LD, GC content, cell type-specific epigenomic marks, eQTL significance and coding/non-coding region. Since the selection of matching properties may sometime not be straightforward, a reference list of publications for selecting different matching properties was provided in Supplementary Table S1.

To save disk space and support quick data retrieval of the sampling database, we designed a novel data structure, called bin-wise chunk-indexed data structure (see Supplementary Notes for details) and a corresponding index system for the sampling database (Supplementary Fig. S1A). The novel data structure allows vSampler to read only variants of queried MAF bins and data chunks of queried properties, which saves disk reads and achieves great speed gain.

3 Sampling algorithm

vSampler requires users to supply a list of variants and number of control variants needed (sampling number), and to select matching properties together with its allowed deviations. The allowed deviations are necessary as it's impossible to sample enough variants with exactly matched properties, and the size of the control pool corresponding to different deviation values is shown in Supplementary Figure S2. The following processes were executed (Supplementary Fig. S1B):

1. All input variants will be annotated with selected matching properties and assigned to sorted MAF bins (MAF is a mandatory matching property due to the fact that it is more uniformly distributed compared with other properties (Supplementary Fig. S3), MAF bins of input variants are [0, 0.01), [0.01, 0.02), . . ., [0.49, 0.50]).
2. For each MAF bin of input variants, based on user-specified MAF deviation and other matching properties, vSampler will randomly access data of corresponding MAF bins of the sampling database and only read data chunks of selected properties. These data are stored as a queue in computer memory temporarily.
3. For each variant in one MAF bin of input variants, according to the user-specified property deviations and the property values of the variant, vSampler randomly samples sampling numbers of matched control variants from the data stored in the queue without replacement. If the number of qualified variants is less than sampling number, vSampler would sample with replacement.

During the sampling procedure, vSampler sequentially processes the MAF bins of input variants, and employed a read-store-delete-in-order algorithm to ensure that the whole sampling database would be read at most once, which minimize the disk reads and speed up the sampling process (Supplementary Fig. S1C, see Supplementary Notes for details).

4 Web server and standalone program

vSampler allows fast variant sampling by both web server and local program. The input variant format of vSampler can be Variant Call Format, dbSNP ID or tab-delimited chromosome coordinates. For matching properties, MAF and its allowed deviation are mandatory, while others are optional. vSampler also provides other options such as matching variant type (SNP/Indels). The output of vSampler is comprised of a main file of each input variant followed by all its controls per line, an annotation file of all variants and corresponding annotations, a configuration file and an excluding file of all input variants unfound in the annotation database. In addition, the web server provides visualization of the distribution of each matching property of input and matched control variants (see Supplementary Fig. S4 and web server documentation for more details).

5 Benchmark and usage examples

vSampler can process large number of queries much more efficiently than other tools. Benchmarking on the runtime of vSampler and that of SNPsnap demonstrated that vSampler is 3–439 times faster than SNPsnap depending on different query settings (Supplementary Table S2).

Sampling of matched control variants can help avoid spurious GWAS enrichment analysis results and we simulated a usage example similar to Pers *et al.* (2015). A GWAS was simulated with randomly distributed phenotypes without genetic basis, and index independent variants were selected by clumping. GWAS loci for index independent variants were defined and Fisher's exact test showed that a set of genes mapped to GO term 'negative regulation of transcription, DNA-templated' were significantly enriched (odds ratio=2.11, *P*-value=0.0026) in these loci (see Supplementary Notes and Supplementary Table S3 for details). 10 000 sets of matched control variants were sampled by vSampler for the GWAS independent variants to construct the null fold enrichment distribution and an insignificant empirical *P*-value of 0.3923 was got as we expected.

vSampler is also able to help assess the enrichment of target variants in functional regions (Schmidt *et al.*, 2015). eQTL enrichment in chromatin states of relevant cell types was evaluated using vSampler for illustration. We used the whole blood eQTLs and 15-state chromatin states data of 27 blood related cell types to perform the enrichment analysis. By calculation of enrichment scores based on the eQTL variants and the sampled 1000 sets of control variants by vSampler, we found that whole blood eQTLs are much more enriched in 8 active states than 7 repressed states in blood related cell types as expected (see Supplementary Fig. S5, Supplementary Table S4 and Supplementary Notes for more details).

The comprehensive annotations provided by vSampler as matching properties can also help to identify spurious results because of annotation correlation. Variants falling within DNase I hypersensitive sites (DHSs) of GM12878 lymphoblastoid cells were randomly sampled (denoted as DHS variants). We then evaluated the enrichment of these DHS variants in histone modification H3K4me3 in the same cell type by sampling control variants matching for only MAF and control variants matching for both MAF and DHS as a comparison. It turned out that when relying on control variants matching for only MAF, it yields an empirical *P*-value of <0.001 while when using control variants matching for both MAF and DHS, an insignificant *P*-value of 0.368 was obtained, which correctly identified that the enrichment of the DHS variants in H3K4me3 resulted from the correlation between DHS and H3K4me3, instead of an independent enrichment in H3K4me3 (see Supplementary Notes for more details).

In addition to the above examples of enrichment analysis, matched variant sampling is also used for construction of negative datasets for the training and testing of functional variant prediction models such as GWAVA (Ritchie *et al.*, 2014) and cepip (Li *et al.*, 2017). More specifically, in GWAVA, three models were trained based on corresponding type of control SNVs including random SNVs, SNVs matched for DTCT with positive data and SNV matched for genomic region with positive data, while most test data of GWAVA was established by sampling SNVs match for DTCT. In cepip, four cell type-specific models were established using different controls and matched epigenomic features. These models built with different variant sampling strategies demonstrated distinct performance and facilitated the analysis of feature importance (see Supplementary Notes for more details).

Acknowledgements

The authors thank software engineer Mr. Sifa Wang for technical consultant.

Funding

This work was supported by grants from the National Natural Science Foundation of China 31871327, 32070675 (M.J.L.).

Conflict of Interest:
The authors declare that there are no conflict of interests.

References

Farh,K.K. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.

Jia,P. *et al.* (2010) Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophr Res.*, **122**, 38–42.

Li,M.J. *et al.* (2017) cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol.*, **18**, 52.

Pers,T.H. *et al.* (2015) SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics*, **31**, 418–420.

Ritchie,G.R. *et al.* (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.

Schmidt,E.M. *et al.* (2015) GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, **31**, 2601–2606.

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Zhang,S. *et al.* (2019) regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. *Nucleic Acids Res.*, 47, e134.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.

Supplementary Notes

Database construction

Annotation database

vSampler used publicly available genotype call sets of AFR, AMR, EAS, EUR, and SAS super populations from 1000 Genomes Project phase 3 release 20130502 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/; release date 05/02/2013; sample size: AFR: 645, AMR:346, EAS: 497, EUR: 495, SAS:486) and processed these call sets using bcftools. We first split multi-allelic variants into multiple bi-allelic variants and then left-aligned and normalized reference and alternative alleles of all variants. Duplicate variants that map to the same position with identical reference and alternative alleles were removed. Finally, the annotation database contained 81,647,035 variants for genome build GRCh37 (hg19) and 78,122,255 variants for genome build GRCh38 (hg38).

Sampling database

The number of variants in the annotation database was too large to be feasible for the sampling process. We kept only variants with MAF > 0.01 of the annotation database to construct the sampling database (GRCh37 (hg19): 16,750,259 variants for AFR population, 11,184,049 variants for AMR population, 8,668,864 variants for EAS population, 9,808,459 variants for EUR population and 10,264,032 variants for SAS population; GRCh38 (hg38): 14,399,202 variants for AFR population, 9,464,290 variants for AMR population, 7,892,804 variants for EAS population, 8,873,459 variants for EUR population and 9,124,174 variants for SAS population).

Annotation of variant properties

Detailed annotation process of variant properties is described below. Distributions of MAF, DTCT, gene density, number of variants in LD and GC content are shown in Supplementary Figure S3. It should be noted that variants in annotation database and sampling database of different populations are different, and the choice of population would also affect the value of MAF, gene density (when using LD to define variant loci) and number of variants in LD.

Minor allele frequency

Variants' MAF of EUR, EAS and AFR population were computed based on allele frequency information from 1000 Genomes Project phase 3 release as described in database construction section.

Distance to closest transcription start site

All 5' transcription start sites were defined according to GENCODE v32 and then we calculated variants' distance to the closest 5' transcription start sites.

Gene density

Gene density refers to number of genes overlapping with variant loci. Genes were extracted from GENCODE v32, and variant loci were defined by LD thresholds (r^2 > (0.1, 0.2, …, 0.9)) or physical distance (window size of 100, 200, ..., 1000 kb).

Number of variants in LD

Number of variants in LD were calculated using LD thresholds (r^2 > (0.1, 0.2, …, 0.9)).

GC content

GC content of variants were computed with various window sizes (100bp, 200bp,...,500bp) based on 5 base GC content file from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gc5Base/hg19.gc5Base.txt.gz for hg19, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.gc5Base.wigVarStep.gz for hg38).

Cell type specific epigenomic marks

Annotation of cell type-specific epigenomic marks is binary indicator of whether variants fall within selected cell type-specific epigenomic marks. There are 6 cell type specific epigenomic marks including DNase I hypersensitive sites (DHSs) and histone modifications H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3, and each covering 127 tissues/cell types. All cell type specific epigenomic mark data were downloaded from the NIH Roadmap Epigenomics Project (https://egg2.wustl.edu/roadmap/web_portal/), and more details of the 127 tissues/cell types can be found in Consolidated_EpigenomeIDs_summary_Table (https://docs.google.com/spreadsheets/d/1yikGx4MsO9Ei36b64yOy9Vb6oPC5IBGlFbYEt-N6gOM/edit#gid=15) in the website (https://egg2.wustl.edu/roadmap/web_portal/meta.html) of NIH Roadmap project. 16 out of 127 epigenomes (E114-E129) in the RoadMap Project were directly borrowed from ENCODE project.

eQTL

Annotation of eQTL significance is binary indicator of whether variants are significant eQTL variants. Significant eQTL variants data of 49 tissues/cell types were downloaded from GTEx project v8 (https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL.tar). Significance of eQTL variants were determined based on permutation by GTEx project. Genome coordinates of GTEx eQTLs were converted from hg38 to hg19 by PyLiftover 0.4, a python implementation of UCSC LiftOver tool. The number of eQTLs we lost in this process for each tissue was summarized in Supplementary Table S5.

Coding/noncoding region

We first identified variant effects using Jannovar, and then variants effects were classified as coding, non-coding and others as listed in Supplementary Table S6.

Data structure and index system

The bin-wise chunk-indexed data structure means that all variants in the sampling database are first sorted by MAF and then assigned to MAF bins ([0.01, 0.02), [0.02, 0.03), …, [0.49, 0.50]). For variants belonging to each MAF bin, they have many properties and the data of each property of variants are treated as one data chunk (property data chunk). Every property data chunk of variants in each MAF bin are indexed and stored in a linear structure. Fitting in with the linear data structure, the index system recorded the address of every property data chunk within each MAF bin and can quickly retrieve any single property data without loading all of them. In other words, based on the index system, vSampler can randomly access only necessary MAF bins and property data chunks. For example, if there 50 input variants in MAF bin [0.34, 0.35) and the matching properties and their allowed deviations are MAF +- 0.01, distance to closest transcription start site (DTCT) +- 1000bp, then only variants of sampling database in MAF bins [0.33, 0.34), [0.34, 0.35), [0.35, 0.36) would be extracted and only property data chunks of MAF and DTCT would be retrieved, thus significantly reduces disk reads.

Searching strategy

We used a read-store-delete-in-order algorithm which means that, for each MAF bin of input variants, corresponding MAF bins of sampling database are read and stored as a queue in computer memory. Since MAF bins of input variant are processed in order, when processing the next MAF bin of input variants, vSampler would compare the required MAF bins of

sampling database with the data in the queue and delete unwanted MAF bins in the queue. For example, if there are 50 input variants in MAF bin [0.34, 0.35), 50 input variants in MAF bin [0.35, 0.36) and the matching properties and their allowed deviations are MAF +- 0.01, variants of sampling database in MAF bins [0.33, 0.34), [0.34, 0.35), [0.35, 0.36) would first be read and stored in queue. Then for MAF bin [0.35, 0.36) of input variants, variants of sampling database in MAF bins [0.34, 0.35), [0.35, 0.36), [0.36, 0.37) are required. The queue would delete MAF bin [0.33, 0.34) and read variants of MAF bin [0.36, 0.37) from sampling database. This strategy makes sure that vSampler wouldn't read the same MAF bins of sampling database repeatedly.

GWAS simulation

We used 1000 Genome phase 3 genotype data of EUR population (minor allele frequency > 0.01) and simulated random phenotype following standard normal distribution (N(0,1)) without genetic basis. PLINK was used to perform association test and clumping (PLINK--clump-p1 1e-4 --clump-kb 500 --clump-r2 0.01). Finally, we got 248 independent variants.

Usage example processing

Example 1: GWAS-associated gene enrichment analysis

All genes mapped to Gene Ontology (GO) terms were downloaded using Ensembl BioMart, which resulted in 23,393 genes mapped to 14,224 GO terms. We further extracted 17,747 genes overlapped with variants of vSampler sampling database, and 465 genes of them were mapped to negative regulation of transcription (DNA-templated) GO term (GO: 0045892). To perform enrichment analysis, overlapped genes of GWAS variants were defined as the union of: (1) nearest gene to each GWAS variant; (2) genes overlapping with each GWAS locus (locus was defined by flank physical distance of 50 kb). We then retrieved 380 overlapped genes of 248 independent GWAS variants. To test the significance of enrichment of negative regulation of transcription (DNA-templated) genes in simulated GWAS loci, a contingency table (Supplementary Table S3) was constructed to calculate odds ratio and to calculate *P*-value by one-side Fisher's exact test, which gave an odds ratio of 2.11 and corresponding *P*-value of 0.0025803.

vSampler was used to sample 10,000 set of matched control variants for 248 GWAS independent variants using default parameters (MAF deviation: 0.05, DTCT deviation: 5000,

Exclude Input: True, GC deviation: 0.05), and odds ratio could be calculated for each set as described above. This resulted in a null distribution of 10,000 odds ratios and 3932 of them were larger than 2.11, which indicated an empirical P-value of 0.3923.

Example 2: eQTL-associated functional annotation enrichment analysis

We used a set of significant whole blood eQTLs identified by FastQTL from GTEx v8, and only retained eQTLs overlapped with vSampler sampling database. 10000 sets of control variants were then sampled by vSampler using default parameters (MAF deviation: 0.05, DTCT deviation: 5000, Exclude Input: True, GC deviation: 0.05). Cell type-specific 15-state chromatin states data were downloaded from NIH Roadmap Epigenomics Project (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz) which includes 27 blood related cell types. For each chromatin state of each cell type, number of eQTL variants/control variants falling within the chromatin state region was calculated and denoted as overlapping eQTL number and overlapping control number respectively, and the overlapping control numbers of 10000 sets of control variants was used to construct null distributions. Then for each chromatin state of each cell type, the enrichment score was calculated as (overlapping eQTL number) / (mean overlapping control number of 10000 sets), and the empirical *P*-value was computed based on the null distributions. The computed *P*-values were corrected for multiple testing by Bonferroni correction for 15*27. Enrichment scores and *P*-values are shown in Supplementary Figure S5 and Supplementary Table S4 respectively. The heatmap of enrichment scores indicates that blood eQTLs are more enriched in 8 active states (1_TssA, 2_TssAFlnk, 3_TxFlnk, 4_Tx, 5_TxWk, 6_EnhG, 7_Enh, 8_ZNF-Rpts) than 7 repressed states (9_Het, 10_TssBiv, 11_BivFlnk, 12_EnhBiv, 13_ReprPC, 14_ReprPCWk, 15_Quies) in blood-related cell types as expected.

Example 3: Identification of spurious enrichment resulting from annotation correlation

We first randomly sampled 1000 variants falling within DHS regions (defined by narrowPeak file) of GM12878 lymphoblastoid cells (E116) as the positive dataset (denoted as DHS variants). Then 1000 sets of control variants matching for only MAF (MAF deviation: 0.05) were sampled by vSampler each with the same sample size as DHS variants (denoted as random controls). Similarly, 1000 sets of control variants matching for both MAF (MAF deviation: 0.05) and DHS regions of the same cell type were generated with vSampler (denoted as DHS-matched controls). Then number of DHS variants residing in histone modification H3K4me3

regions (defined by narrowPeak file) of GM12878 lymphoblastoid cells (E116) was computed to be 357 (denoted as positive overlapping number). Numbers of variants from each set of random controls and DHS-matched controls falling in H3K4me3 regions of the same cell type were calculated (denoted as random overlapping number and DHS-matched overlapping number) and used to construct random null distribution and DHS-matched null distribution respectively. For the random null distribution, all 1000 random overlapping numbers were smaller than 357 and thus resulted in an empirical *P*-value of < 0.001. In comparison, for the matched-null distribution, 368 DHS-matched overlapping numbers were larger than the positive overlapping number 357 and led to an insignificant empirical *P*-value of 0.368, which was consistent with the fact that these DHS variants were enriched in H3K4me3 regions as result of the correlation between DHS and H3K4me3, rather than an independent enrichment.

Introduction of sampling strategies in GWAVA and cepip

In GWAVA (Ritchie, et al., 2014), when constructing the training dataset, 'regulatory mutations' from the Human Gene Mutation database (HGMD) were used as the disease-implicated set, while three control sets were generated using the idea of matched control sampling based on common SNVs from 1000 Genomic Projects. The first control set is established by randomly selecting SNVs from the whole genome. The second control set is established by randomly selecting SNVs matching for DTCT with variants of the disease-implicated set. The third control set is comprised of all variants within 1000 bp around each variants of the disease-implicated set. Three classifiers were built on the basis of the disease-implicated set and each control set respectively, and they showed distinct performance by cross validation (AUC of 1st control set: 0.97, AUC of 2nd control set: 0.88, AUC of 3rd control set: 0.71). Furthermore, feature importance of three classifiers with different control sets provides interesting biological insights. For example, DTCT is the most important feature for classifier based on the 1st control set, and it's still within the top 3 most important features in other two classifiers even if the second control set is generated by matching for DTCT. DNase1 footprints is only highly ranked for the third classifier, which indicates that when disease-implicated and controls are more physically close and similar to each other, this kind of specific annotation becomes more discriminating. In addition, although there is only minor difference between the average conservation scores of the disease-implicated set and any control set, they are consistently highly ranked across three classifiers irrespective of the control sets, which

indicates the importance of conservation score for functional variants prediction even when conditioning on other features.

In cepip (Li, et al., 2017), fine-mapped eQTL SNPs were selected as the positive data, and two kinds of control sets were constructed with the idea of matched control sampling, a random control set generated by random sampling around the same TSS (within 10kb) and matching for MAF with the positive variants; a strict control set generated by random sampling of variants matching for MAF, DTCT, GC content while excluding any variants in high LD with positive variants. With the two types of control sets, the author identified consistently important features, "selected chromatin features", across 11 eQTL datasets and the "selected chromatin features" were then employed to construct four generalized models. More specifically, the four generalized models were established by using (1) random control; (2) strict control; (3) random control without DHS-related features and (4) strict control without DHS-related features. Test on an independent eQTL dataset demonstrated that the generalized model with random control has the highest partial AUC (0.626) than others (0.619, 0.624, 0.609).

Software, web server and code availability

vSampler local program, web server and codes are available at http://mulinlab.org/vsampler or https://github.com/mulinlab/vSampler.

Supplementary Tables

Supplementary Table S1 – List of publications utilizing matched variant sampling with different matching properties

Task	Purpose	Matching properties	Reference
GWAS significant noncoding SNPs from the NHGRI-EBI GWAS Catalog	Enrichment in DHS	MAF, distance to closest TSS, genomic feature localization	(Maurano, et al., 2012)
Genome-wide significant SNPs associated with platelet and erythrocyte phenotypes	Enrichment in nucleosome-depleted regions (NDR)	MAF, distance to a TSS, number of proxy SNPs	(Paul, et al., 2013)
66128 common disease-associated SNPs	Enrichment in NFκB Binding Regions	MAF, distance to a TSS	(Karczewski, et al., 2013)
Genome-wide significant variants associated with migraine	Enrichment in DHS	MAF, distance to closest TSS, GC content	(Anttila, et al., 2013)
Independent lead SNPs associated with myocardial infarction or coronary artery disease	Enrichment in histone marks H3K27ac, H3K4me3, H3K9ac	MAF, number of SNPs in LD, distance to the nearest gene, gene density	(Won, et al., 2015)
methylation QTLs	Enrichment in eQTLs	MAF, number of SNPs in LD, distance to the nearest gene, gene density	(Zaghlool, et al., 2016)
Independent SNPs associated with 33 GRASP GWAS phenotypes	Enrichment in eQTLs	MAF, number of SNPs in LD, distance to the nearest gene	(DeBoever, et al., 2017)
eQTLs of iPS cell	Enrichment in chromatin states, transcription factor binding sites, NHGRI–EBI GWAS catalogue variants,	MAF, number of SNPs in LD, distance to the nearest gene, gene density	(Kilpinen, et al., 2017)

Genome-wide significant variants associated with Alzheimer's disease	Enrichment in open chromatin regions, H3K4me3, H3K4me1, H3K27ac	MAF, number of SNPs in LD, distance to the nearest gene, gene density	(Tansey, et al., 2018)
eQTLs of failing and nonfailing human heart tissue	Enrichment in NHGRI-EBI GWAS Catalog variants	Number of variants in LD, gene density	(Cordero, et al., 2019)
cis-eQTL of liver tissue	Enrichment in H3K4me3 and H3K27ac peaks of liver tissue	MAF, gene density, distance to the nearest gene, number of variants in LD	(Caliskan, et al., 2019)
Sentinel SNPs associated with coronary artery disease and blood pressure	Comparison of distance from test SNPs to closest TAD boundary	MAF, gene density, gene proximity, number of variants in LD	(Lalonde, et al., 2019)
SNPs associated with obsessive–compulsive disorder	Enrichment in immune or brain eQTLs	MAF, gene density, distance to the nearest gene, and number of variants in LD	(Khramtsova, et al., 2019)
PICS identified candidate causal SNPs	Enrichment in DeepFIGV high absolute z-score (absolute z-score above a given cutoff)	MAF, gene density, distance to the nearest gene, number of SNPs in LD	(Hoffman, et al., 2019)
SNVs associated with schizophrenia and autism spectrum disorders (ASDs)	Enrichment in looping class	size of the LD block, MAF, distance to the nearest gene, gene density	(Beagan, et al., 2020)
Independent SNPs from the fine-mapped m6A QTLs	Enrichment in m6A consensus motif (RRACH), binding sites of RNA-binding proteins (RBPs), riboSNitches (genetic variants changing RNA secondary structure), predicted miRNA binding sites	MAF, number of SNPs in LD, distances to the nearest gene, gene density, SNP locations relative to genes	(Zhang, et al., 2020)

Supplementary Table S2 – Runtime comparison between vSampler and SNPSnap. Elapsed CPU times (in seconds) for vSampler (running in 1 thread of Intel Xeon E5 2650 V4 CPU) and SNPSnap with variable number of queries and sampling number

Method	**Sampling number**	**Number of queries**						
		10	100	1000	10000	50000	100000	500000
vSampler	1	10.6	13.6	60.0	516.4	2572.5	4867.4	24996.4
SNPSnap	1	60	480	198	15120	NA[a]	NA	NA
vSampler	10	9.8	13.2	60.6	500.8	2629.0	5000.4	25650.4
SNPSnap	10	60	540	3960	28680	NA	NA	NA
vSampler	100	10.0	13.6	61.6	505.2	2688.0	5104.4	27191.6
SNPSnap	100	120	480	3360	33780	NA	NA	NA
vSampler	1000	10.4	14.2	63.8	535.0	2849.5	5597.8	30727.8
SNPSnap	1000	140	600	28020	54780	NA	NA	NA
vSampler	10000	10.4	16.0	84.6	754.4	4083.5	8229.6	41149.0
SNPSnap	10000	180	915	11712	NA	NA	NA	NA

[a] When number of queries is larger than 10000, or when the number of queries is 10000 and sampling number is 10000, SNPSnap will collapse and thus the runtime is marked as NA

Supplementary Table S3 - Contingency table for enrichment analysis. Values used in Fisher's exact test to compute enrichment of negative regulation of transcription (DNA-templated) genes (GO: 0045892) in simulated GWAS loci.

	Mapped to GO term	**Not mapped to GO term**	**Total**
Overlapped genes	20	360	380
Non-overlapped genes	445	16922	17367
Total	465	17282	17747

Supplementary Table S4 – P-values for enrichment scores of blood eQTLs in 15 chromatin states of 27 blood related cell types

Cell Types	1_TssA	2_TssAFlnk	3_TxFlnk	4_Tx	5_TxWk	6_EnhG	7_Enh	8_ZNF-Rpts	9_Het	10_TssBiv	11_BivFlnk	12_EnhBiv	13_ReprPC	14_ReprPCWk	15_Quies
BLD.CD14.MONO	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1	1
BLD.CD14.PC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	0.0405	1	1	1	1
BLD.CD15.PC	0.0405	0.0405	0.4455	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	0.5265	0.0405	1	1	1
BLD.CD19.CPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.9315	1	0.0405	1	1	1
BLD.CD19.PPC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.486	0.0405	1	1	1	1
BLD.CD3.CPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD3.PPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1
BLD.CD34.CC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD34.PC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD4.CD25.CD127M.TREGPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1
BLD.CD4.CD25I.CD127.TMEMPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.567	0.324	1	1	1	1
BLD.CD4.CD25M.CD45RA.NPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.891	1	1	1	1	1
BLD.CD4.CD25M.CD45RO.MPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD4.CD25M.IL17M.PL.TPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.972	0.081	1	1	1	1
BLD.CD4.CD25M.IL17P.PL.TPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD4.CD25M.TPC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.405	1	1	1	1
BLD.CD4.MPC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD4.NPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.9315	1	1	1	1	1
BLD.CD56.PC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	0.0405	0.2835	1	1	1
BLD.CD8.MPC	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.CD8.NPC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.DND41.CNCR	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1	1
BLD.GM12878	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	0.8505	1	1	1	1
BLD.K562.CNCR	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.MOB.CD34.PC.F	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.MOB.CD34.PC.M	0.0405	0.0405	0.405	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	1	1	1	1	1	1
BLD.PER.MONUC.PC	0.0405	0.0405	1	0.0405	0.0405	0.0405	0.0405	0.0405	0.0405	0.2835	0.0405	1	1	1	1

Supplementary Table S5 - Number of eQTLs lost during liftover from hg38 to hg19

Tissue	LiftOver Failed eQTLs Number/%
Adipose Subcutaneous	8617/3.197984
Adipose Visceral Omentum	6706/3.7559159
Adrenal Gland	3725/5.2933737
Artery Aorta	5661/3.2706474
Artery Coronary	3444/6.6499324
Artery Tibial	7022/2.5419373
Brain Amygdala	1381/7.0448401
Brain Anterior cingulate cortex BA24	1868/5.835494
Brain Caudate basal ganglia	3006/4.7238155
Brain Cerebellar Hemisphere	4042/5.3177214
Brain Cerebellum	4530/4.2302053
Brain Cortex	3396/4.5245011
Brain Frontal Cortex BA9	2288/4.5819565
Brain Hippocampus	1971/5.6958733
Brain Hypothalamus	2700/7.4474541
Brain Nucleus accumbens basal ganglia	3258/5.0812564
Brain Putamen basal ganglia	2794/6.182647
Brain Spinal cord cervical c-1	763/3.2207683
Brain Substantia nigra	789/5.0325297
Breast Mammary Tissue	5933/4.3644255
Cells Cultured fibroblasts	7335/2.8231083
Cells EBV-transformed lymphocytes	1017/3.595164
Colon Sigmoid	5367/4.4798544
Colon Transverse	6240/4.4853686
Esophagus Gastroesophageal Junction	5020/3.9685991
Esophagus Mucosa	7490/3.195747
Esophagus Muscularis	7533/3.3727334
Heart Atrial Appendage	5375/3.8904169
Heart Left Ventricle	4628/3.8623314
Kidney Cortex	241/5.1016088

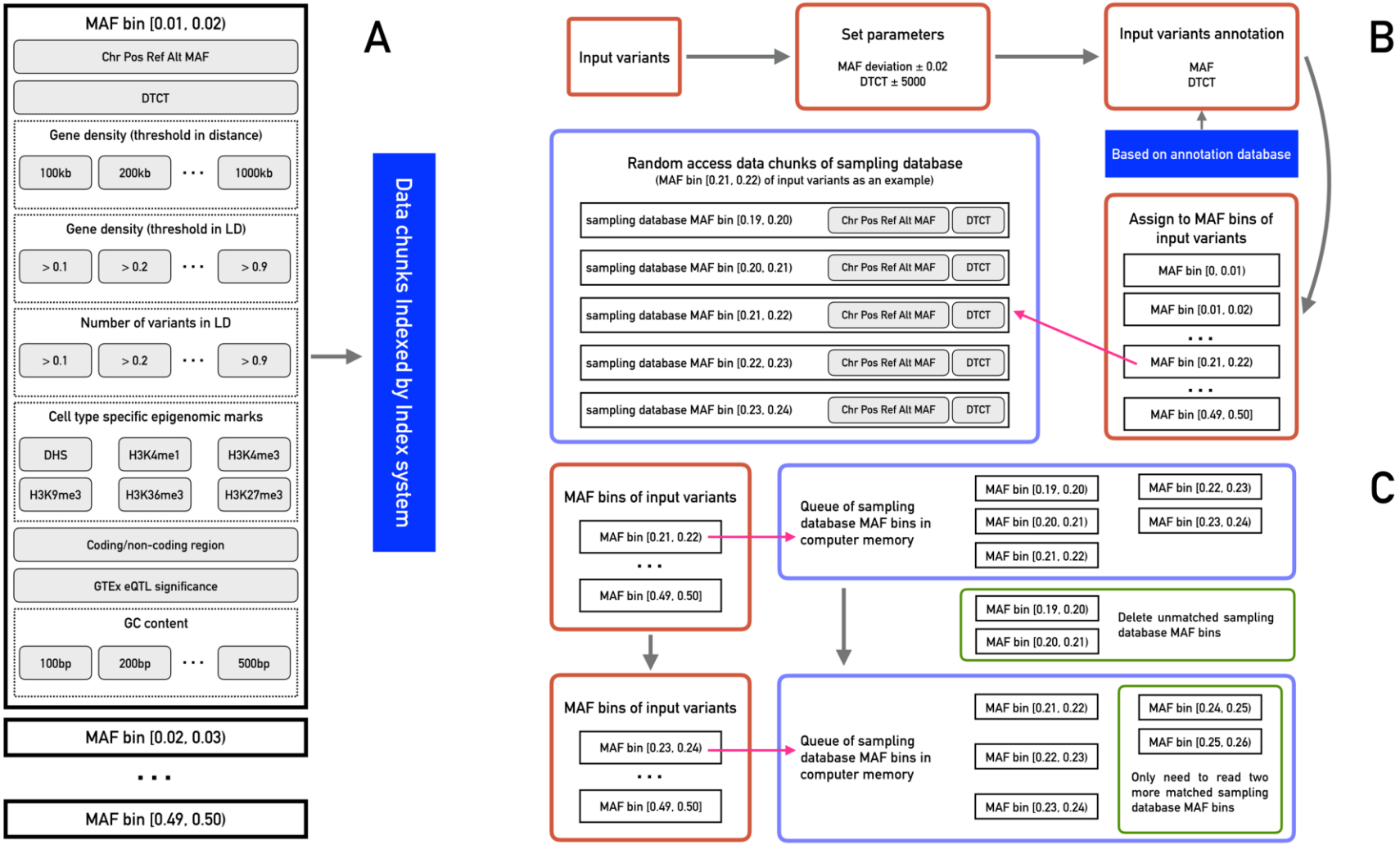
Liver	2110/4.7516101
Lung	7477/3.498339
Minor Salivary Gland	2047/7.8973765
Muscle Skeletal	6443/2.593675
Nerve Tibial	7983/2.53675
Ovary	2524/6.8427045
Pancreas	4799/4.3970643
Pituitary	4543/5.1560549
Prostate	3716/6.3085699
Skin Not Sun Exposed Suprapubic	7950/3.2383815
Skin Sun Exposed Lower leg	9146/3.0570021
Small Intestine Terminal Ileum	3027/7.030379
Spleen	5402/5.7231852
Stomach	4817/5.0244075
Testis	9547/4.112445
Thyroid	9349/2.7472899
Uterus	883/4.7740052
Vagina	2110/10.9400114
Whole Blood	6602/2.9090364

Supplementary Table S6 – The variant classification according to Jannovar variant effect

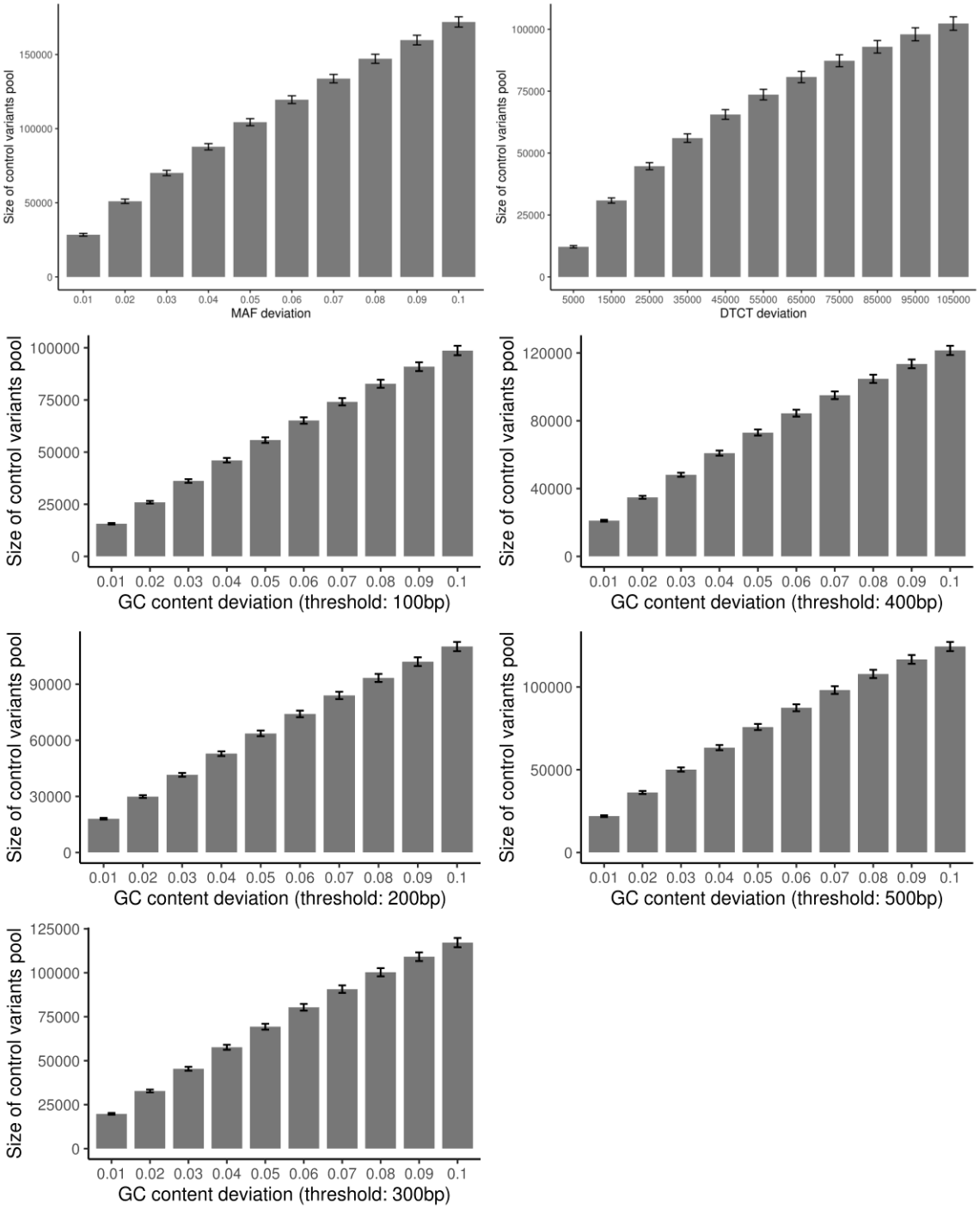
Coding	Noncoding	Others
TRANSCRIPT_VARIANT	NON_CODING_TRANSCRIPT_VARIANT	MOBILE_ELEMENT_DELETION
EXON_VARIANT	REGULATORY_REGION_ABLATION	SEQUENCE_VARIANT
GENE_VARIANT	INTERGENIC_VARIANT	DIRECT_TANDEM_DUPLICATION
INTRON_VARIANT	REGULATORY_REGION_AMPLIFICATION	STRUCTURAL_VARIANT
CODING_SEQUENCE_VARIANT	MIRNA	CUSTOM
CONSERVED_INTRON_VARIANT	TFBS_AMPLIFICATION	MOBILE_ELEMENT_INSERTION
SPLICING_VARIANT	CONSERVED_INTERGENIC_VARIANT	CHROMOSOME
CODING_TRANSCRIPT_VARIANT	TFBS_ABLATION	_SMALLEST_LOW_IMPACT
INTRAGENIC_VARIANT	TF_BINDING_SITE_VARIANT	_SMALLEST_MODERATE_IMPACT
FIVE_PRIME_UTR_PREMATURE_START_CODON_GAIN_VARIANT	INTERGENIC_REGION	INSERTION
THREE_PRIME_UTR_EXON_VARIANT	DOWNSTREAM_GENE_VARIANT	INVERSION
THREE_PRIME_UTR_INTRON_VARIANT	UPSTREAM_GENE_VARIANT	TRANSLOCATION
FIVE_PRIME_UTR_INTRON_VARIANT	REGULATORY_REGION_VARIANT	COPY_NUMBER_CHANGE
STOP_RETAINED_VARIANT	NON_CODING_TRANSCRIPT_INTRON_VARIANT	MNV
SYNONYMOUS_VARIANT	NON_CODING_TRANSCRIPT_EXON_VARIANT	COMPLEX_SUBSTITUTION
INITIATOR_CODON_VARIANT		FEATURE_TRUNCATION
SPLICE_REGION_VARIANT		CHROMOSOME_NUMBER_VARIATION
CODING_TRANSCRIPT_INTRON_VARIANT		_SMALLEST_HIGH_IMPACT
FIVE_PRIME_UTR_EXON_VARIANT		INTERNAL_FEATURE_ELONGATION
DISRUPTIVE_INFRAME_DELETION		
DISRUPTIVE_INFRAME_INSERTION		
INFRAME_DELETION		
MISSENSE_VARIANT		
FIVE_PRIME_UTR_TRUNCATION		
INFRAME_INSERTION		
THREE_PRIME_UTR_TRUNCATION		
FRAMESHIFT_ELONGATION		
EXON_LOSS_VARIANT		
SPLICE_DONOR_VARIANT		
TRANSCRIPT_AMPLIFICATION		
FRAMESHIFT_VARIANT		
STOP_GAINED		
SPLICE_ACCEPTOR_VARIANT		
TRANSCRIPT_ABLATION		
STOP_LOST		
RARE_AMINO_ACID_VARIANT		
FRAMESHIFT_TRUNCATION		
START_LOST		

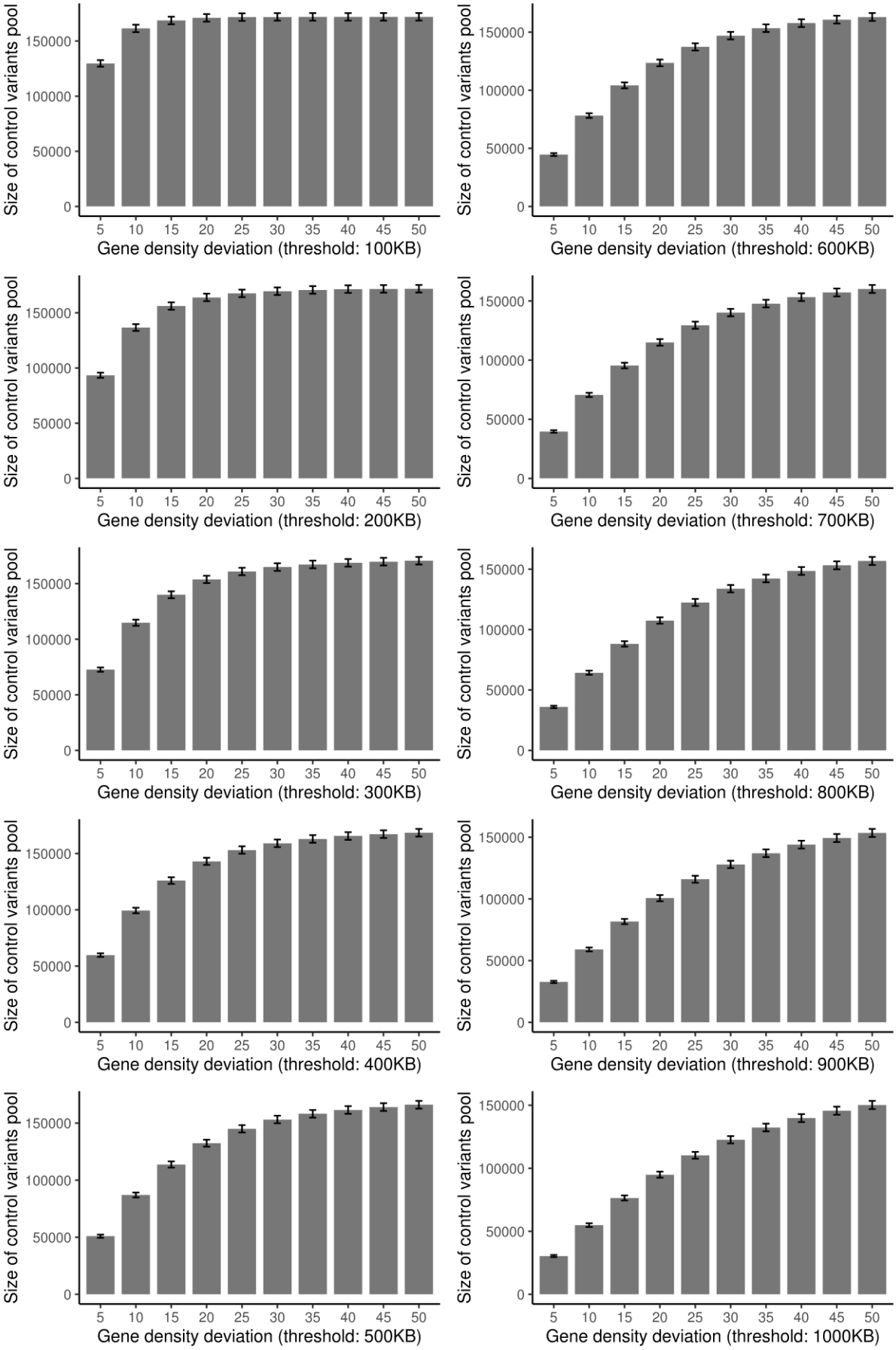
Supplementary Figures

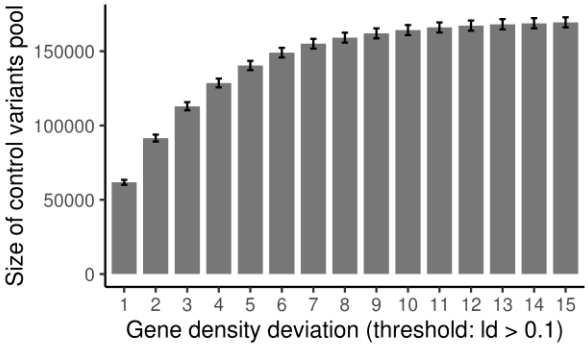
Supplementary Figure S1 – **A. Data structure of sampling database.** Each gray solid box represents one data chunk indexed by the index system. DTCT: distance to closest transcription start site; **B. vSampler pipeline.** We set MAF deviation to ± 0.02 and DTCT deviation to ± 5,000 for illustration. It's worth noting that MAF bins of input variants are different from MAF bins of sampling database; **C. Read-store-delete-in-order algorithm.** When vSampler processes from MAF bin [0.21, 0.22) of input variants to MAF bin [0.23, 0.24) of input variants, the change of queue of sampling database MAF bins in computer memory is shown. MAF deviation is set to ± 0.02 (see Supplementary Notes for more details).

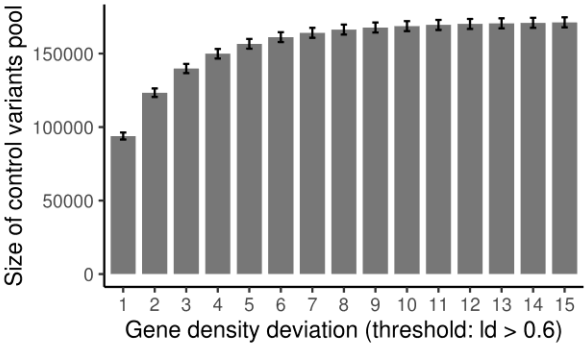

A
MAF bin [0.01, 0.02)
Chr Pos Ref Alt MAF
DTCT
Gene density (threshold in distance)
100kb
200kb
1000kb
Gene density (threshold in LD)
> 0.1
> 0.2
> 0.9
Number of variants in LD
> 0.1
> 0.2
> 0.9
Cell type specific epigenomic marks
DHS
H3K4me1
H3K4me3
H3K9me3
H3K36me3
H3K27me3
Coding/non-coding region
GTEx eQTL significance
GC content
100bp
200bp
500bp
MAF bin [0.02, 0.03)
MAF bin [0.49, 0.50)
Data chunks Indexed by Index system
B
Input variants
Set parameters
MAF deviation ± 0.02
DTCT ± 5000
Input variants annotation
MAF
DTCT
Based on annotation database
Random access data chunks of sampling database
(MAF bin [0.21, 0.22) of input variants as an example)
sampling database MAF bin [0.19, 0.20)
sampling database MAF bin [0.20, 0.21)
sampling database MAF bin [0.21, 0.22)
sampling database MAF bin [0.22, 0.23)
sampling database MAF bin [0.23, 0.24)
Chr Pos Ref Alt MAF
DTCT
Assign to MAF bins of input variants
MAF bin [0, 0.01)
MAF bin [0.01, 0.02)
MAF bin [0.21, 0.22)
MAF bin [0.49, 0.50]
C
MAF bins of input variants
MAF bin [0.21, 0.22)
MAF bin [0.49, 0.50]
Queue of sampling database MAF bins in computer memory
MAF bin [0.19, 0.20)
MAF bin [0.20, 0.21)
MAF bin [0.21, 0.22)
MAF bin [0.22, 0.23)
MAF bin [0.23, 0.24)
MAF bin [0.19, 0.20)
MAF bin [0.20, 0.21)
Delete unmatched sampling database MAF bins
MAF bins of input variants
MAF bin [0.23, 0.24)
MAF bin [0.49, 0.50]
Queue of sampling database MAF bins in computer memory
MAF bin [0.21, 0.22)
MAF bin [0.22, 0.23)
MAF bin [0.23, 0.24)
MAF bin [0.24, 0.25)
MAF bin [0.25, 0.26)
Only need to read two more matched sampling database MAF bins

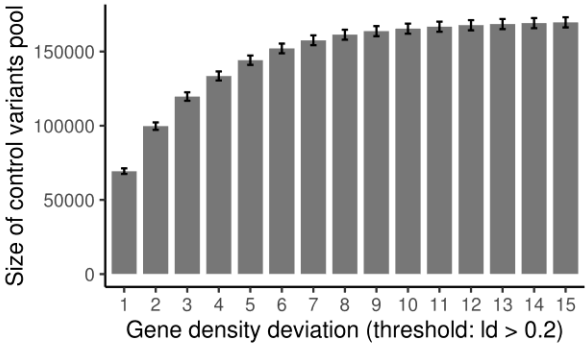
Supplementary Figure S2 – Size of control pool corresponding to varying values of matching property deviations for sampling database of EUR population. Size of control pool correspond to varying values of deviation of MAF, DTCT, GC content (window size: 100bp, 200bp, 300bp, 400bp, 500bp), Gene density (Physical distance window size: 100kb, 200kb, …, 1000kb, LD threshold: 0.1, 0.2, …, 0.9) and Number of variants in LD (LD threshold: 0.1, 0.2, …, 0.9). The bar plots show the mean and standard error of size of control pool, which was estimated by using 1000 random variants as the input data to query the sampling database of EUR population with varying matching property deviations. It's noteworthy that since the MAF is a mandatory matching property and its largest deviation is 0.1, MAF deviation is always set to 0.1 when we estimate the size of control pool with other matching properties.

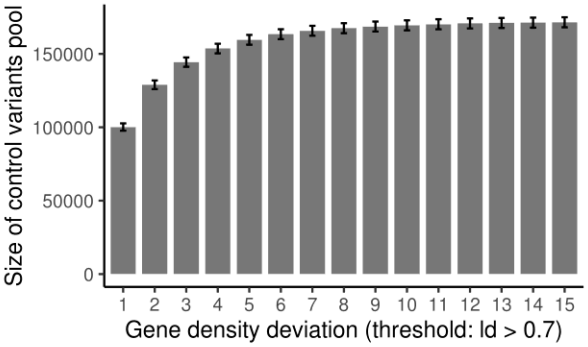


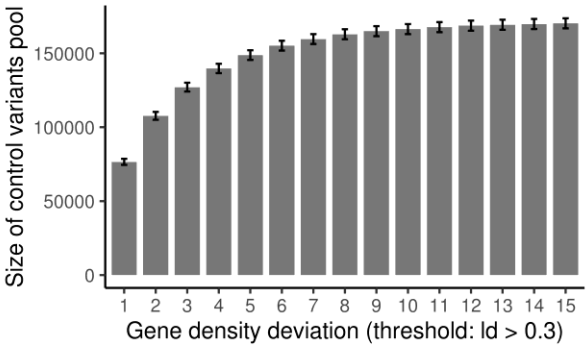

Size of control variants pool
0
50000
100000
150000
5 10 15 20 25 30 35 40 45 50
Gene density deviation (threshold: 100KB)
Gene density deviation (threshold: 200KB)
Gene density deviation (threshold: 300KB)
Gene density deviation (threshold: 400KB)
Gene density deviation (threshold: 500KB)
Gene density deviation (threshold: 600KB)
Gene density deviation (threshold: 700KB)
Gene density deviation (threshold: 800KB)
Gene density deviation (threshold: 900KB)
Gene density deviation (threshold: 1000KB)

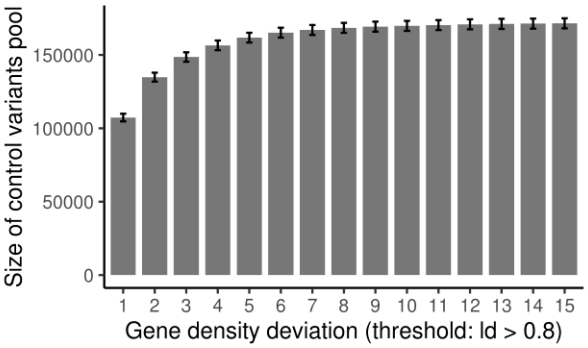

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.1)

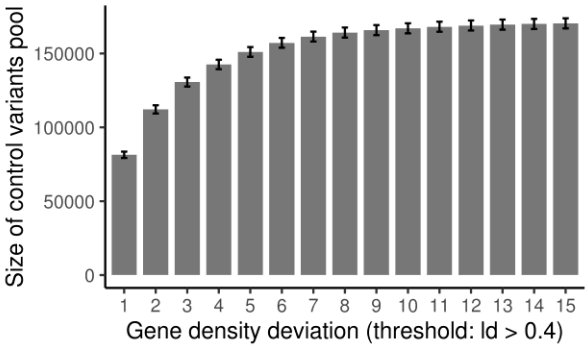

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.6)

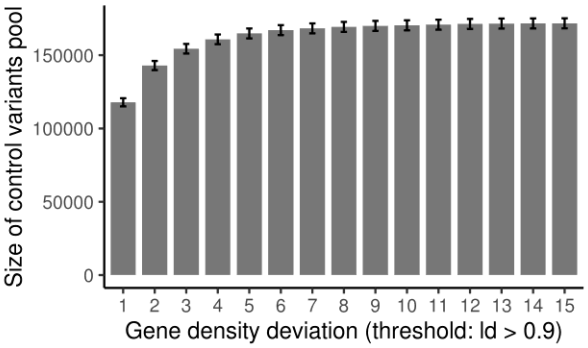

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.2)

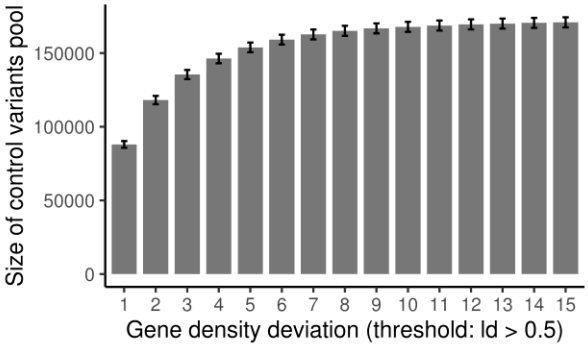

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.7)

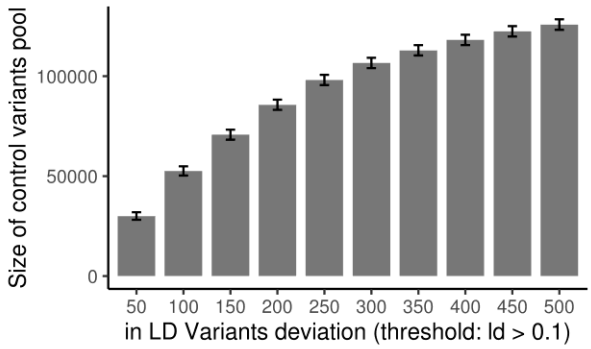

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.3)

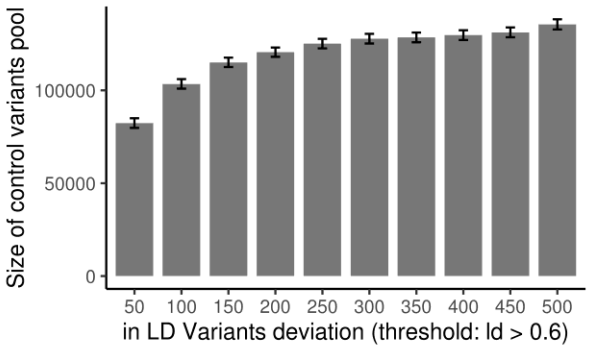

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.8)

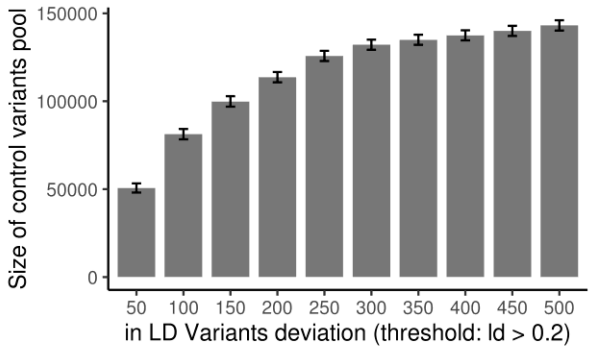

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.4)

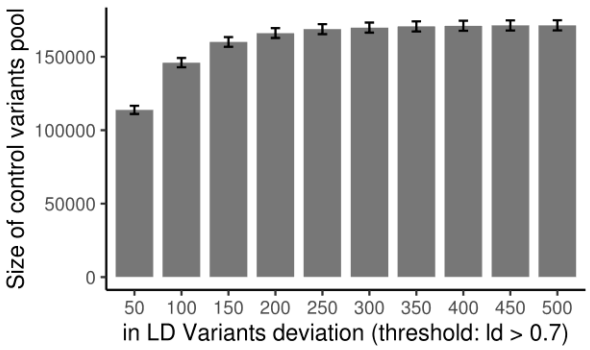

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.9)

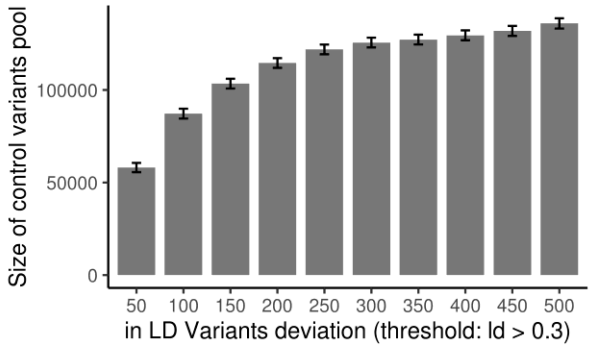

Size of control variants pool
0
50000
100000
150000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Gene density deviation (threshold: ld > 0.5)

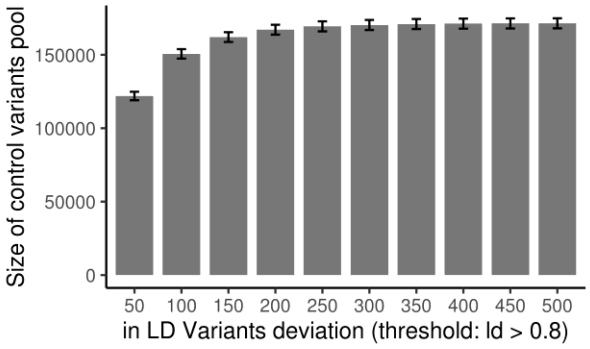

Size of control variants pool
0
50000
100000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.1)

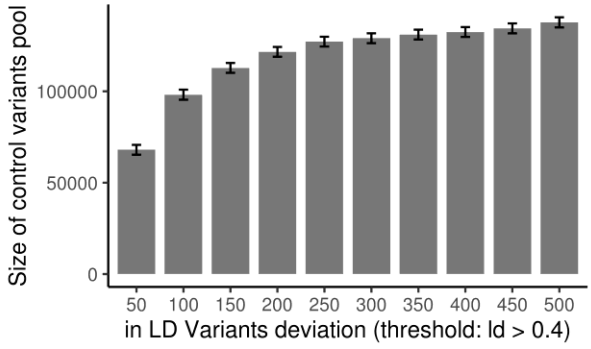

Size of control variants pool
0
50000
100000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.6)

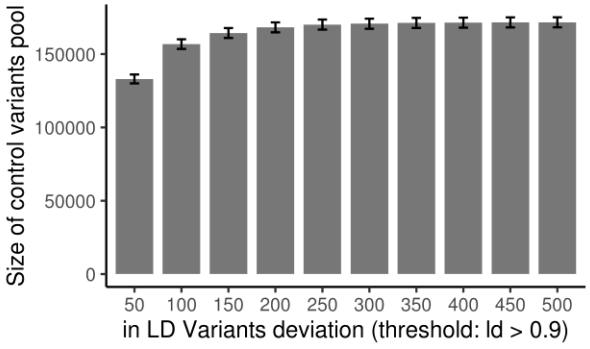

Size of control variants pool
0
50000
100000
150000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.2)

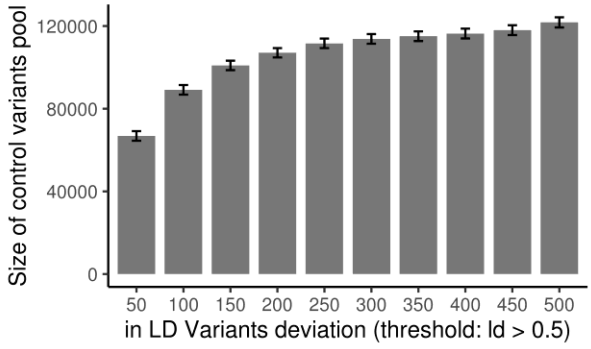

Size of control variants pool
0
50000
100000
150000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.7)


Size of control variants pool
0
50000
100000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.3)

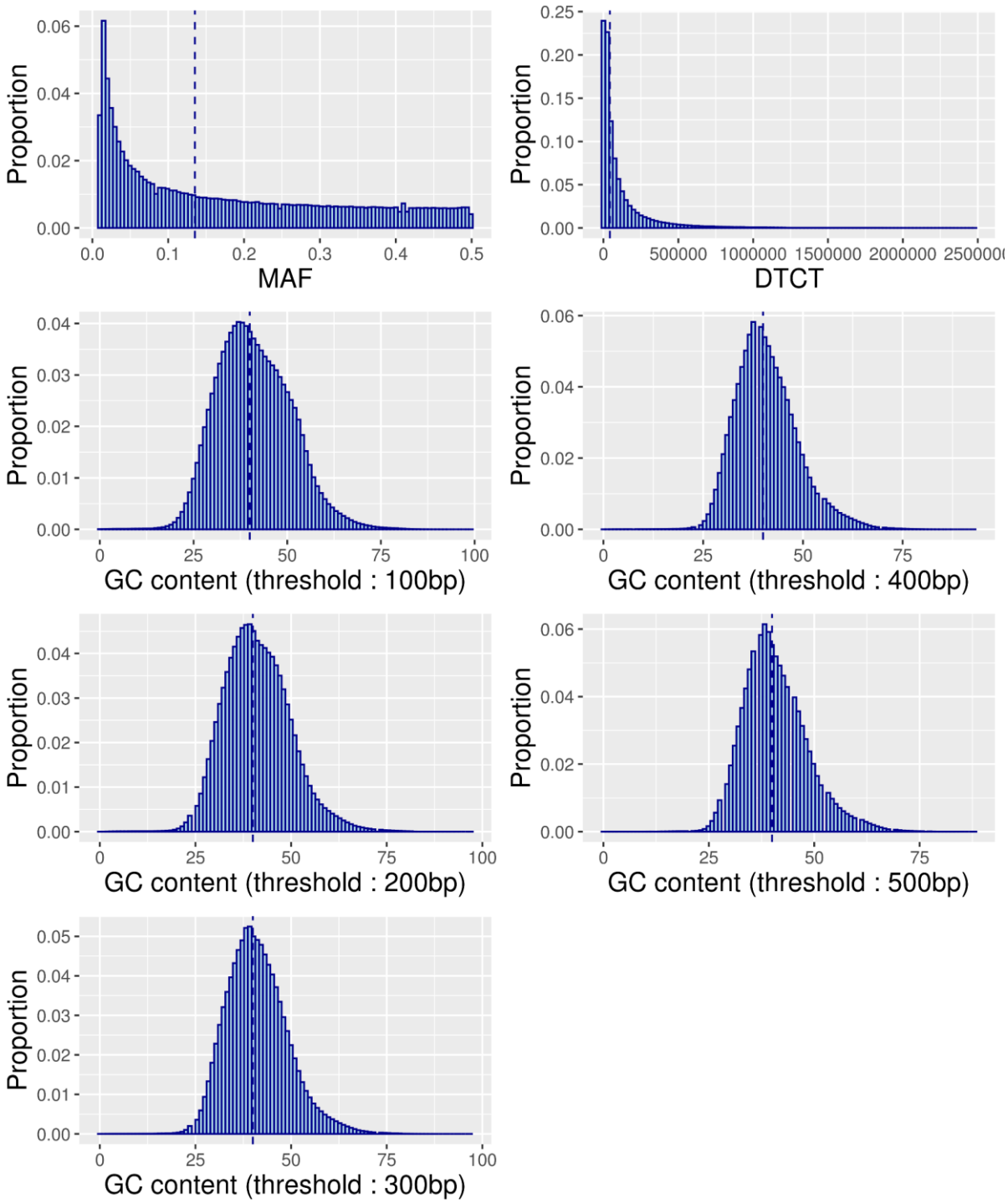

Size of control variants pool
0
50000
100000
150000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.8)

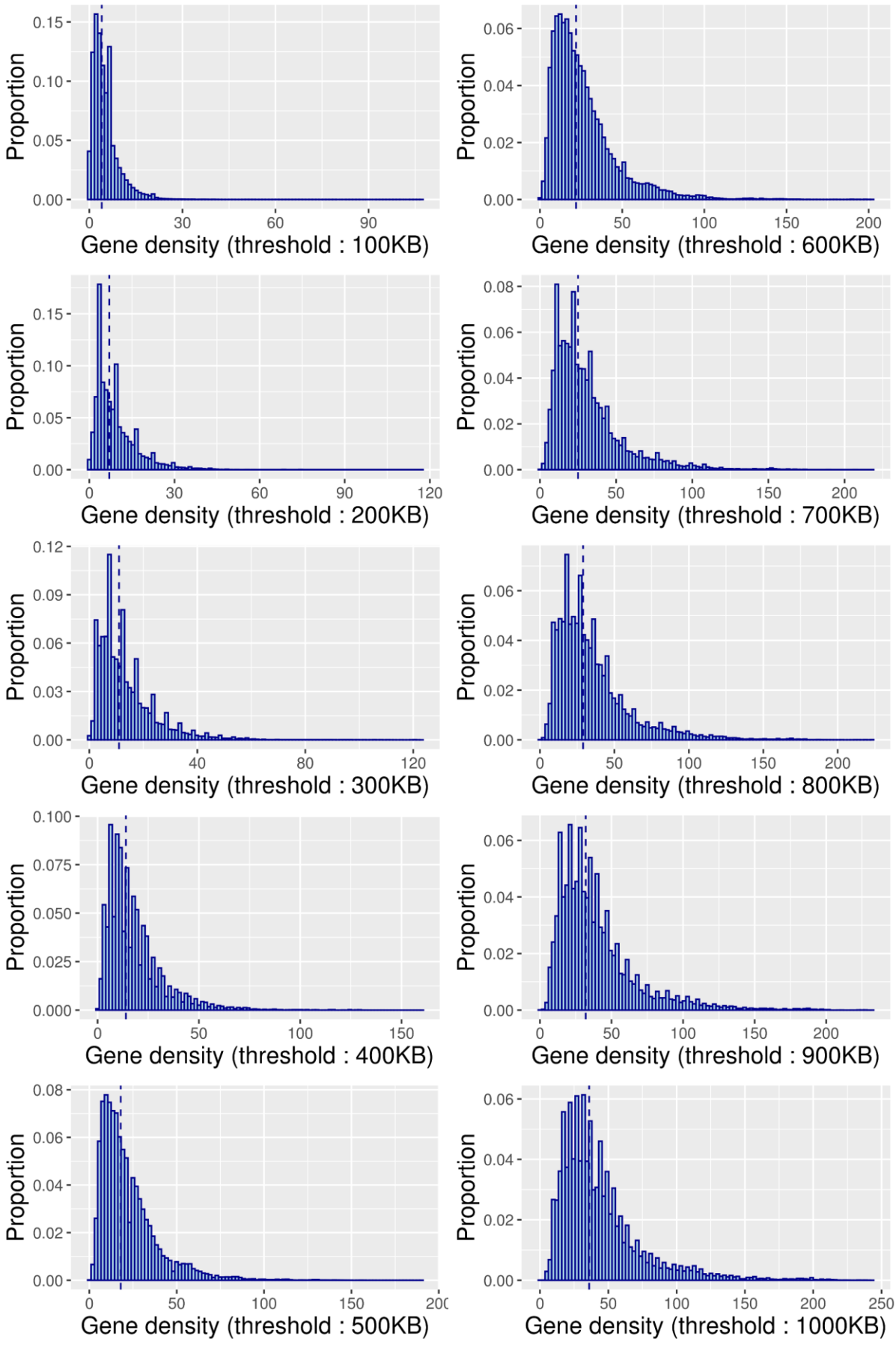

Size of control variants pool
0
50000
100000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.4)

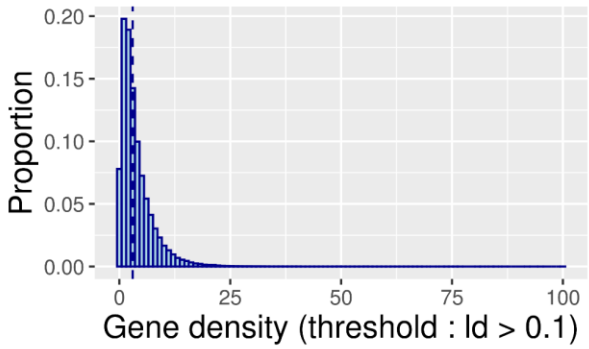

Size of control variants pool
0
50000
100000
150000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.9)

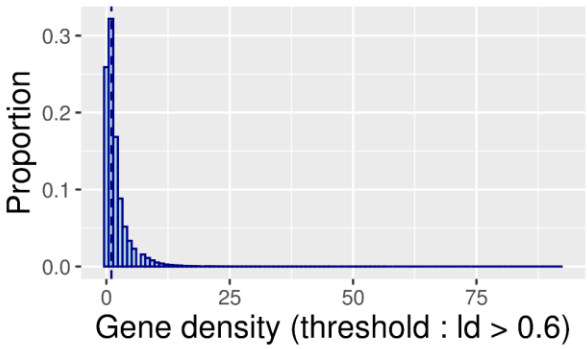

Size of control variants pool
0
40000
80000
120000
50 100 150 200 250 300 350 400 450 500
in LD Variants deviation (threshold: ld > 0.5)

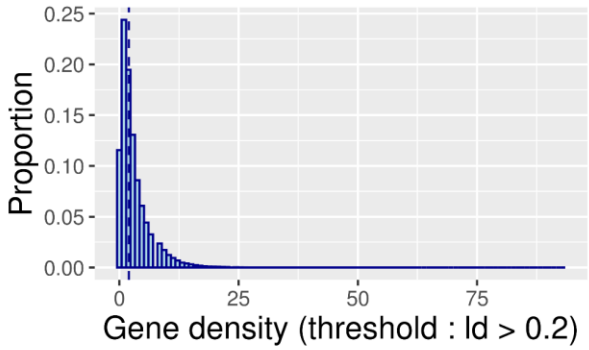
Supplementary Figure S3 - Histograms of distributions of variant properties for sampling database of EUR population. Distributions of MAF, DTCT, GC content (window size: 100bp, 200bp, 300bp, 400bp, 500bp), Gene density (Physical distance window size: 100kb, 200kb, …, 1000kb, LD threshold: 0.1, 0.2, …, 0.9) and Number of variants in LD (LD threshold: 0.1, 0.2, …, 0.9) were plotted for all variants in sampling database of EUR population.

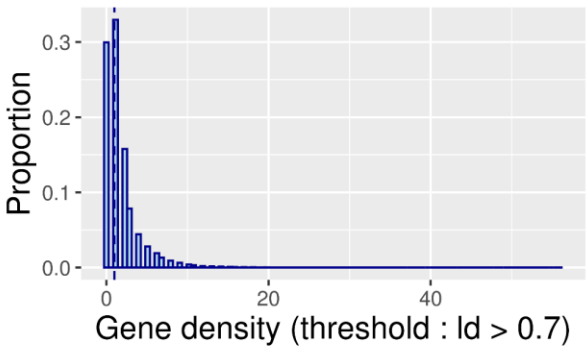


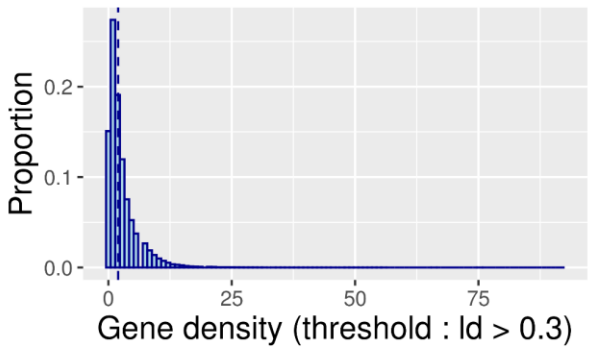

Proportion
Gene density (threshold : 100KB)
Gene density (threshold : 600KB)
Gene density (threshold : 200KB)
Gene density (threshold : 700KB)
Gene density (threshold : 300KB)
Gene density (threshold : 800KB)
Gene density (threshold : 400KB)
Gene density (threshold : 900KB)
Gene density (threshold : 500KB)
Gene density (threshold : 1000KB)

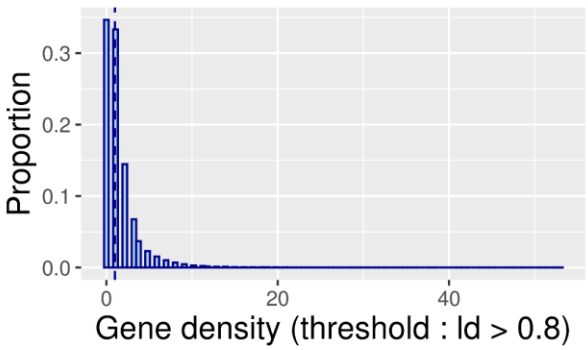

Proportion
0.20
0.15
0.10
0.05
0.00
0
25
50
75
100
Gene density (threshold : ld > 0.1)

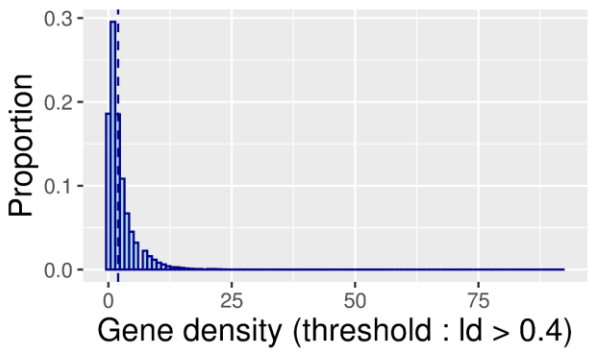

Proportion
0.3
0.2
0.1
0.0
0
25
50
75
Gene density (threshold : ld > 0.6)

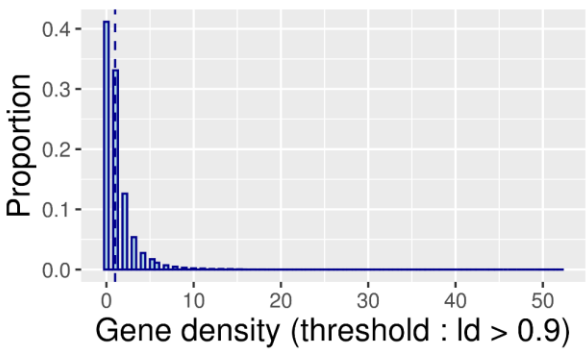

Proportion
0.25
0.20
0.15
0.10
0.05
0.00
0
25
50
75
Gene density (threshold : ld > 0.2)

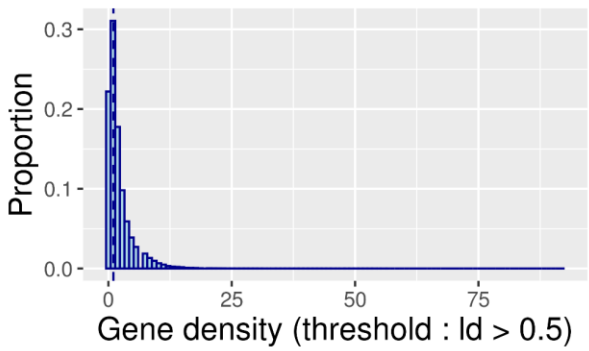

Proportion
0.3
0.2
0.1
0.0
0
20
40
Gene density (threshold : ld > 0.7)

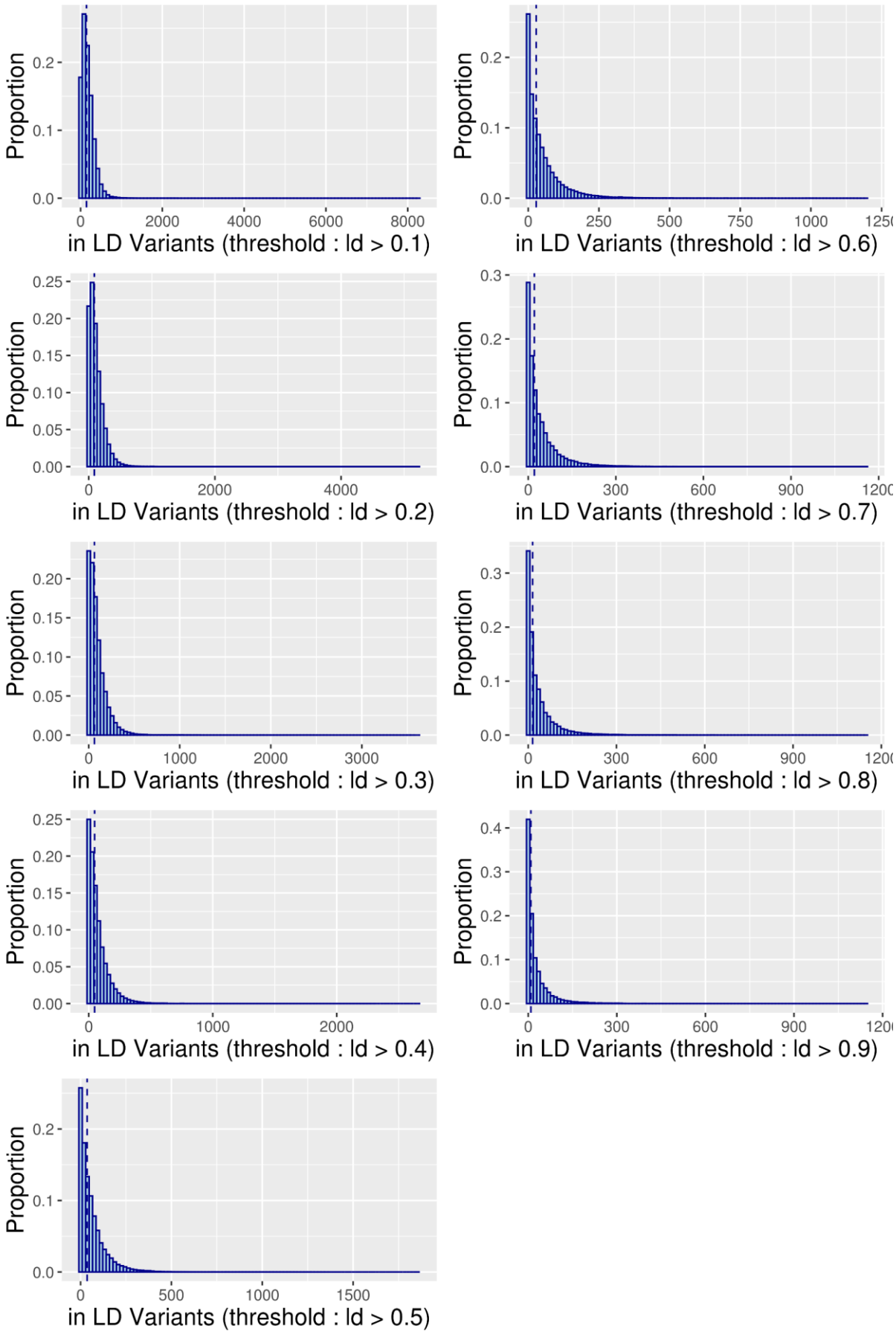

Proportion
0.2
0.1
0.0
0
25
50
75
Gene density (threshold : ld > 0.3)


Proportion
0.3
0.2
0.1
0.0
0
20
40
Gene density (threshold : ld > 0.8)

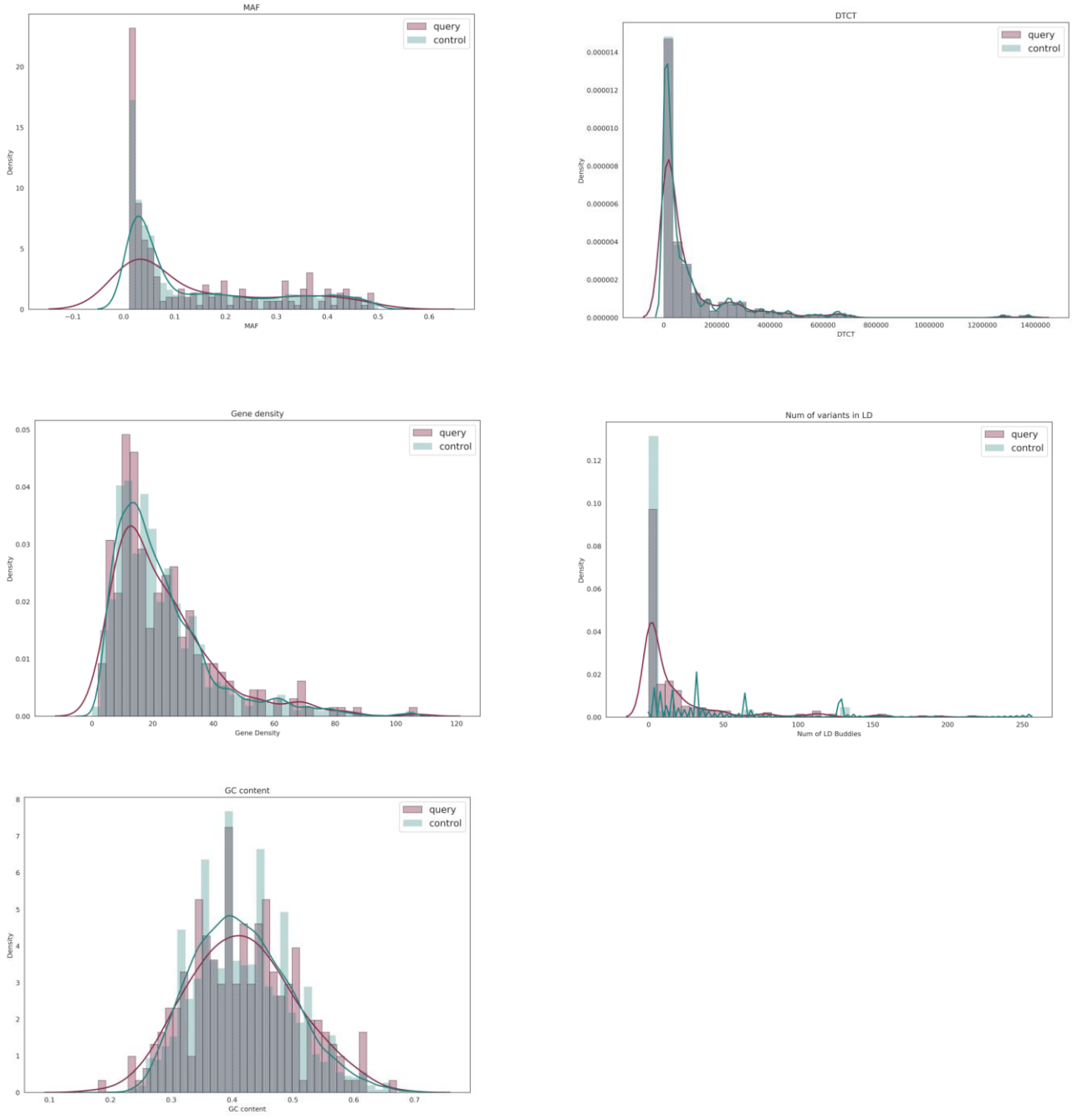

Proportion
0.3
0.2
0.1
0.0
0
25
50
75
Gene density (threshold : ld > 0.4)


Proportion
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Gene density (threshold : ld > 0.9)

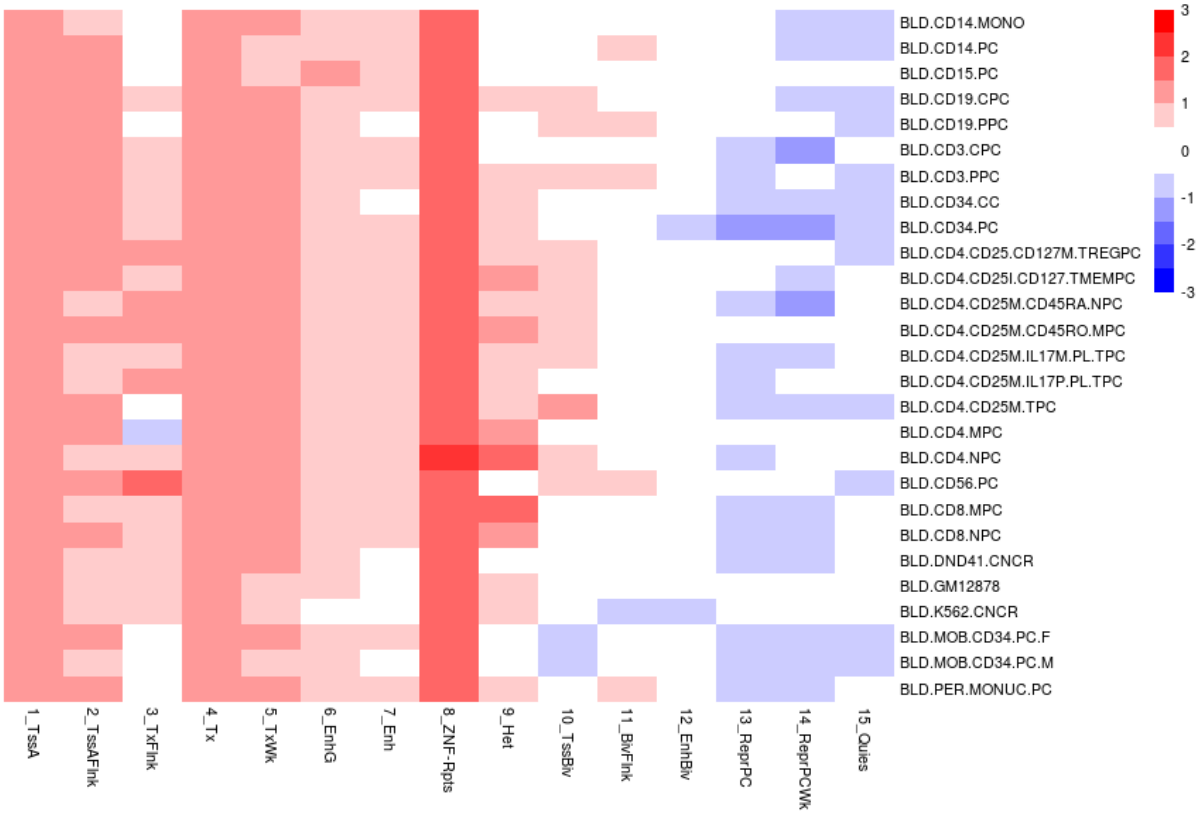

Proportion
0.3
0.2
0.1
0.0
0
25
50
75
Gene density (threshold : ld > 0.5)


Proportion
in LD Variants (threshold : ld > 0.1)
in LD Variants (threshold : ld > 0.6)
in LD Variants (threshold : ld > 0.2)
in LD Variants (threshold : ld > 0.7)
in LD Variants (threshold : ld > 0.3)
in LD Variants (threshold : ld > 0.8)
in LD Variants (threshold : ld > 0.4)
in LD Variants (threshold : ld > 0.9)
in LD Variants (threshold : ld > 0.5)

Supplementary Figure S4 - Example of variant sampler visualization function. 1000 random variants were used as query variants and matching property deviations were set as following, MAF deviation: 0.05, DTCT deviation: 5000, Gene density deviation: 5 (physical distance window size: 100KB), Number of variants in LD deviation: 50 (LD threshold: 0.1), GC content deviation: 0.01 (window size: 100bp). The distribution of MAF, DTCT, Gene density, Number of variants in LD and GC content for both query variants and sampled controls were shown.



Supplementary Figure S5 - Enrichment scores of blood eQTLs in 15 chromatin states of 27 blood related cell types. The scores in the heatmap are in log2-transformed scale.



Anttila, V.*, et al.* Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nat Genet* 2013;45(8):912-917.
Beagan, J.A.*, et al.* Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression. *Nat Neurosci* 2020;23(6):707-717.
Caliskan, M.*, et al.* Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *Am J Hum Genet* 2019;105(1):89-107.
Cordero, P.*, et al.* Pathologic gene network rewiring implicates PPP1R3A as a central regulator in pressure overload heart failure. *Nat Commun* 2019;10(1):2760.
DeBoever, C.*, et al.* Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* 2017;20(4):533-546 e537.
Hoffman, G.E.*, et al.* Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Res* 2019;47(20):10597-10611.
Karczewski, K.J.*, et al.* Systematic functional regulatory assessment of disease-associated variants. *Proc Natl Acad Sci U S A* 2013;110(23):9607-9612.
Khramtsova, E.A.*, et al.* Sex differences in the genetic architecture of obsessive-compulsive disorder. *Am J Med Genet B Neuropsychiatr Genet* 2019;180(6):351-364.
Kilpinen, H.*, et al.* Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* 2017;546(7658):370-375.
Lalonde, S.*, et al.* Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. *Genome Biol* 2019;20(1):133.
Li, M.J.*, et al.* cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol* 2017;18(1):52.
Maurano, M.T.*, et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;337(6099):1190-1195.
Paul, D.S.*, et al.* Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res* 2013;23(7):1130-1141.
Ritchie, G.R.*, et al.* Functional annotation of noncoding sequence variants. *Nat Methods* 2014;11(3):294-296.
Tansey, K.E., Cameron, D. and Hill, M.J. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. *Genome Med* 2018;10(1):14.
Won, H.H.*, et al.* Disproportionate Contributions of Select Genomic Compartments and Cell Types to Genetic Risk for Coronary Artery Disease. *PLoS Genet* 2015;11(10):e1005622.
Zaghlool, S.B.*, et al.* Mendelian inheritance of trimodal CpG methylation sites suggests distal cis-acting genetic effects. *Clin Epigenetics* 2016;8:124.
Zhang, Z.*, et al.* Genetic analyses support the contribution of mRNA N(6)-methyladenosine (m(6)A) modification to human disease heritability. *Nat Genet* 2020.