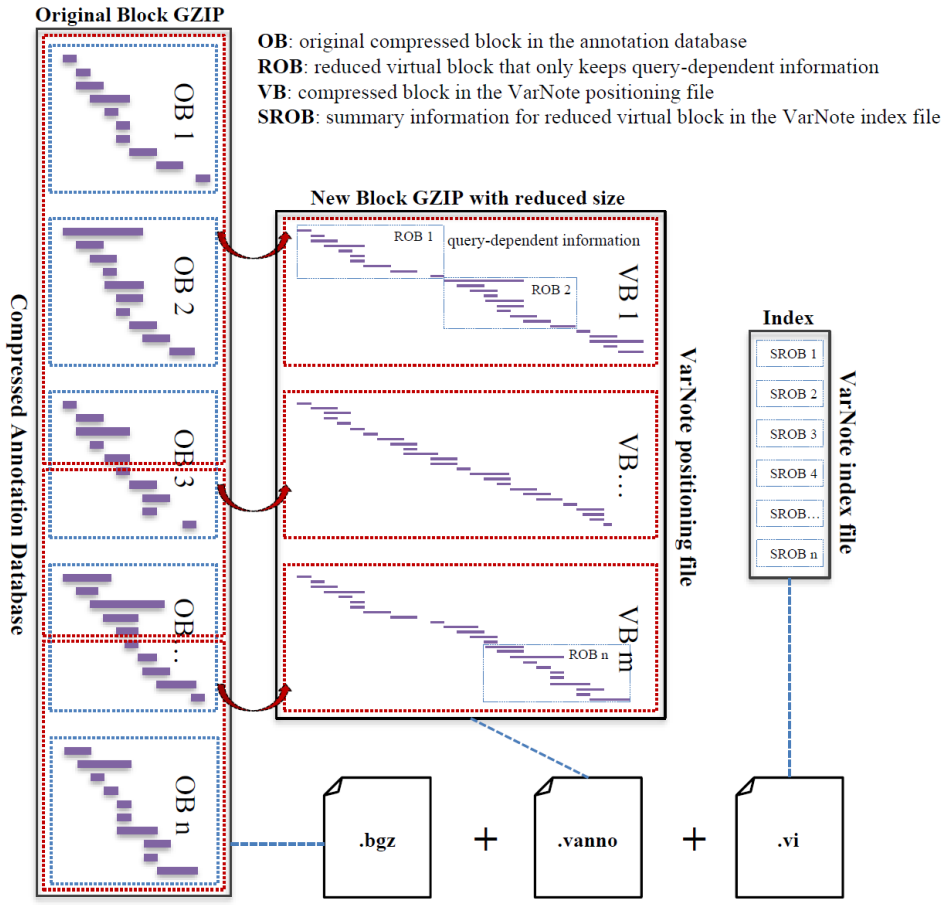
Supplemental Material

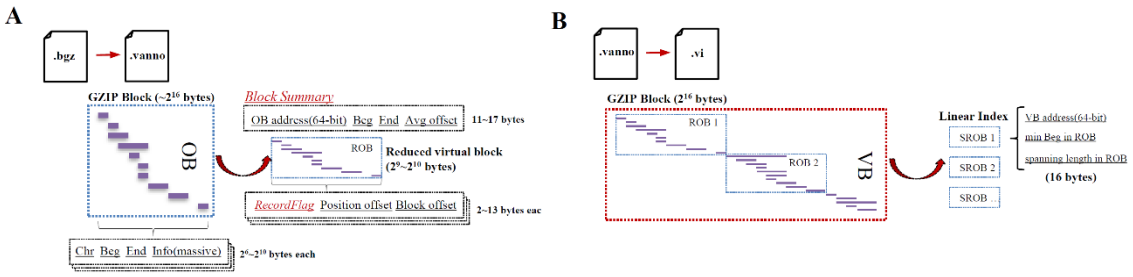
Ultrafast and scalable variant annotation and prioritization with big functional genomics data

This file includes: Supplemental Figures S1-S11, Supplemental Tables S1-S6 and the description of Supplemental Codes.

Supplemental Figures



Supplemental Figure S1. Architecture of VarNote index system. Bgzip-compressed annotation database (.bgz file) will be firstly converted to VarNote positioning file (.vanno file). The system tailors and encodes information of each original compressed block in the annotation database (OB) to generate reduced virtual block that only keeps query-dependent information (ROB). The bgzip-compressed VarNote positioning file contains concatenate compressed block that stores ROB bytes (VB). Then, summary information of reduced virtual block (SROB) is linearly indexed to generate VarNote index file (.vi file).

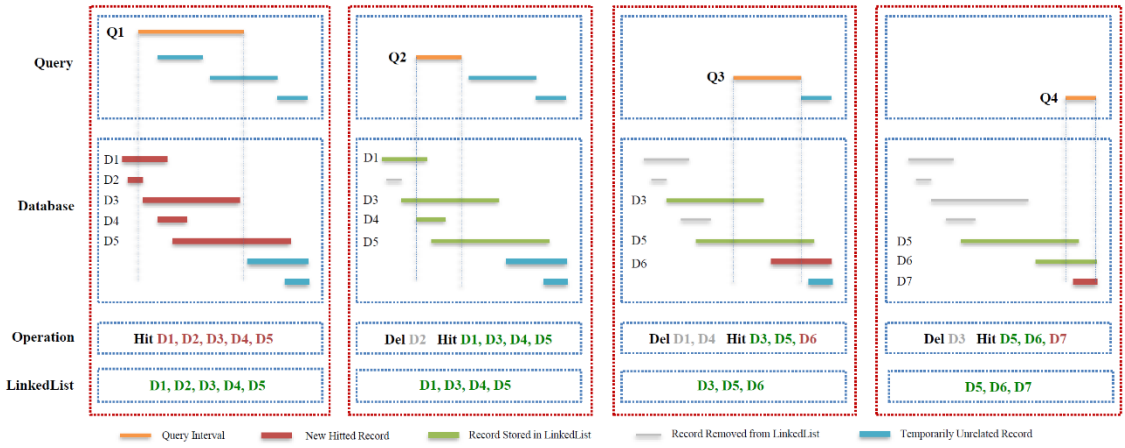


Supplemental Figure S2. The construction of VarNote positioning file and index file. (A) ROB transformation. A normal OB usually stores 2^{16} bytes and may contains hundreds to thousands annotation records (2^{6}~2^{10} bytes each). Since the most space-consuming information is annotation metadata that is useless for interval query, the system only need to remember query-dependent information, called ROB. ROB comprises a block summary (11~17 bytes in total, including a unique 64-bit bgzip OB address, position information of first record and average block offset of all records in the current OB) and hundreds to thousands encoded annotation records (2~13 bytes each, including a record flag, position offset and block offset to average). (B) Linear index of ROB. The system indexes the VarNote positioning file by a series of ROB summary information (16 bytes each, including a unique 64-bit bgzip address of ROB start, initial beginning position of ROB and spanning length of all records in ROB).

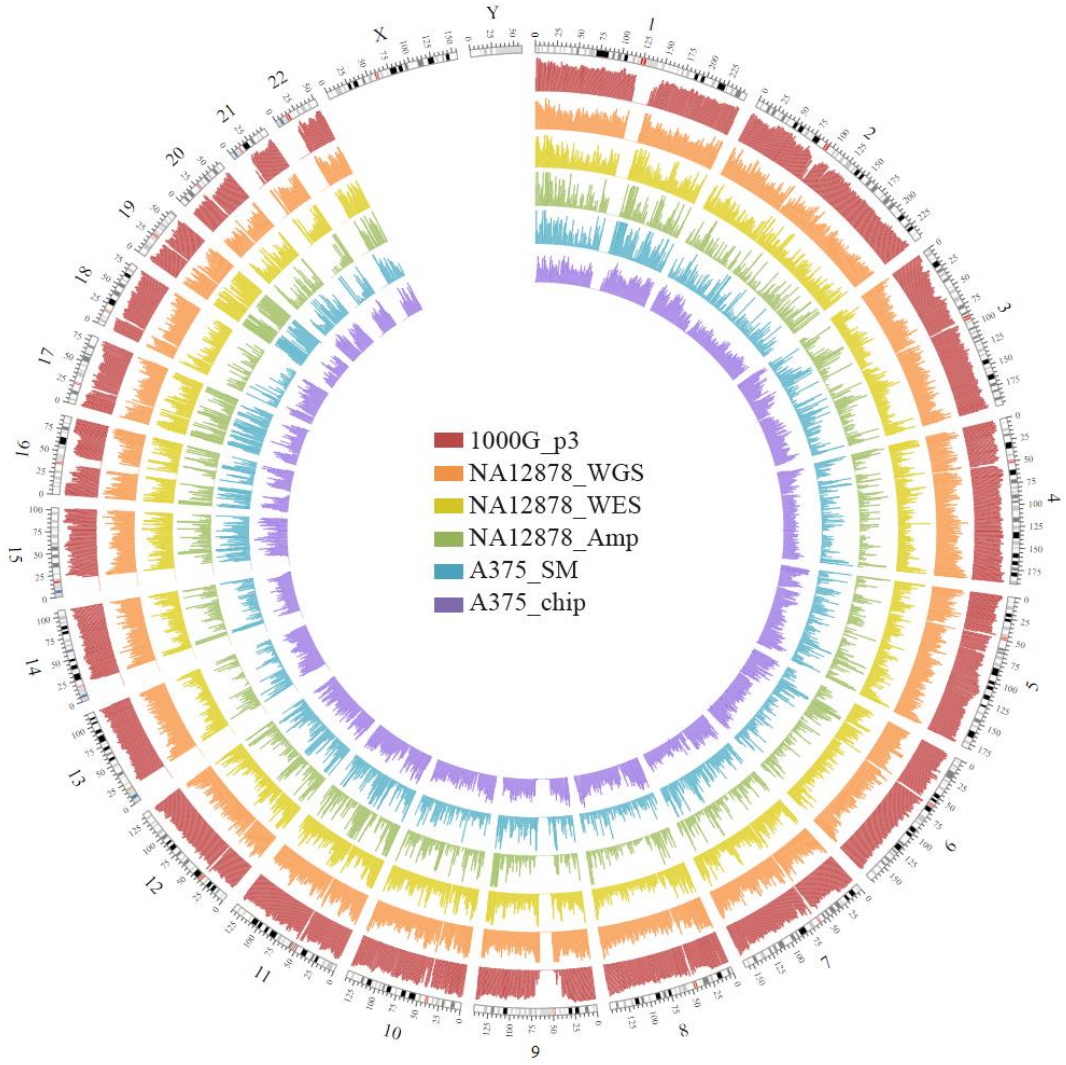
Field	Description	Type
Block Start Flag	-1 means block start and first record distance is 1; -2 means block start and first record distance is a unsigned byte value; -3 means block start and first record distance is a unsigned short value; -4 means block start and first record distance is a int value.	byte
First Record in Block		
Block Address	64-bit bgzip address of record in database file.	long
Record Distance	Record Distance (End minus Beg), optional (If record distance is greater than one, 1,2 or 4 bytes are allocated to store the unsigned byte, unsigned short and int distance value respectively).	0, unsigned byte, unsigned short, int
Average Offset	Average block offset of all records in this block.	short, int
Other Records in Block		
RecordFlag		byte
Record Sign	The 1st bit represents the start of a record flag.	bit[0]
Beg Distance Sign	The 2nd-4th bits encodes the storage size of distance between the current beg and previous (additional bytes are allocated to store the difference).	bit[1-3]
Record Distance Sign	The 5th-6th bits encodes the storage size of record distance (additional bytes are allocated to store the difference).	bit[4-5]
Direction Sign	The 7th bits encodes direction sign indicates whether the difference of block offset to average is a positive or negative value.	bit[6]
Storage Size Sign	The 8th bits encodes a sign indicates the storage size of block offset to average (additional bytes are allocated to store the difference).	bit[7]
Beg Distance	Optional, if the distance between current beg and previous is greater than four, 1,2 or 4 bytes are allocated to store the unsigned byte, unsigned short and int distance value respectively.	0, unsigned byte, unsigned short, int
Record Distance	Optional, if record distance is greater than one, 1,2 or 4 bytes are allocated to store the unsigned byte, unsigned short and int distance value respectively.	0, unsigned byte, unsigned short, int
Difference between Block Offset to Average Offset	Additional 1-4 bytes are allocated to store the difference, allocating 1,2 or 4 for unsigned byte, short, int value of difference respectively.	unsigned byte, short, int

Field	Description	Type
Block Count	Total block in current chromosome	int
Summary of Each Block		
Block Address	Address of this block in the VarNote file.	long
Start Position	The start position (Beg) of the first record in this block.	int
Block Distance	The max distance of all the records in this block.	int

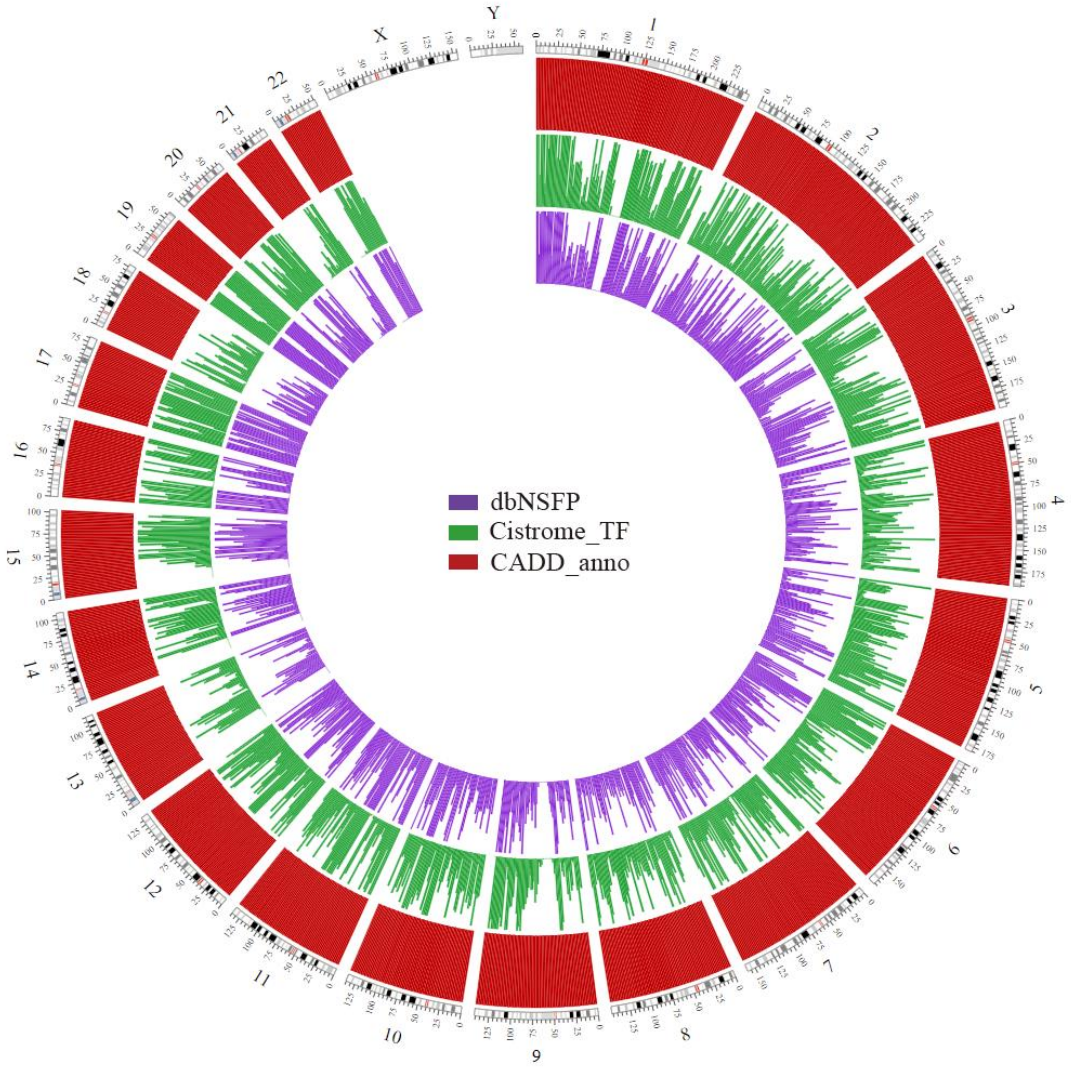
Supplemental Figure S3. The VarNote index format (including .vanno and .vi files). The original database file should be bgzip-compressed.



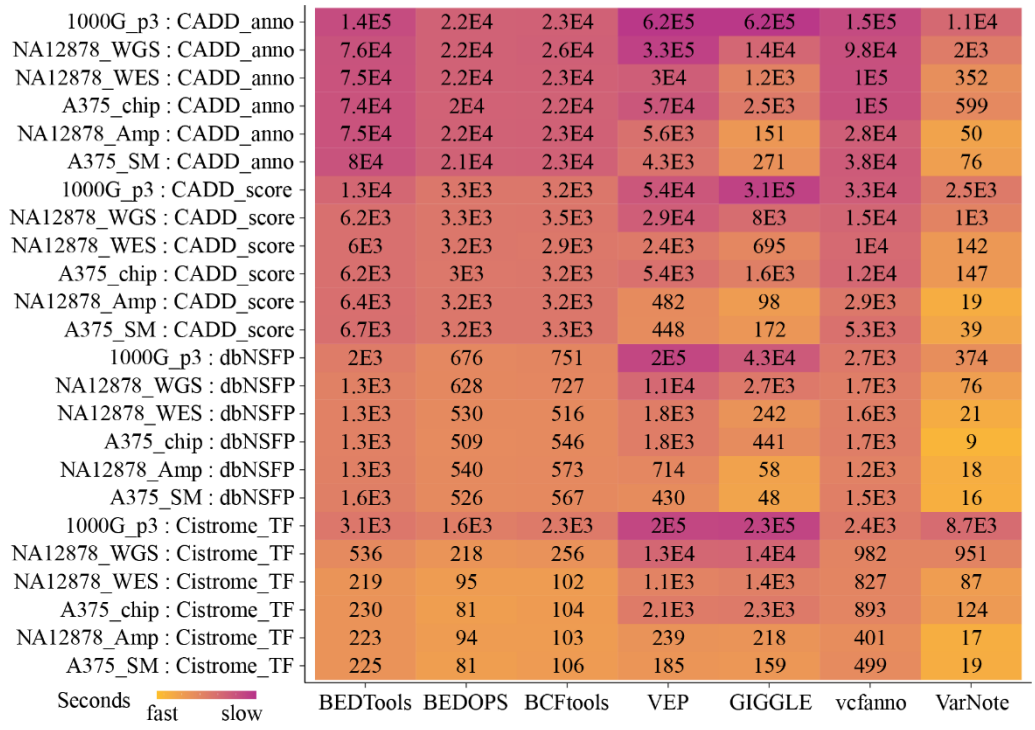
Supplemental Figure S4. Process of chromosome sweeping. Suppose there are four query intervals (Q1-Q4, some of them overlap each other) and seven annotation records in the database (D1-D7, some of them overlap each other), VarNote sequentially intersects Q1 with annotation records to identify all record hits (Hit D1, D2, D3, D4, D5). A global linked list is used to cache these intersected records in case next query interval encounters. Then, VarNote processes Q2 by firstly inspecting whether there are any overlaps in the global linked list (Hit D1, D3, D4, D5), and then dropping useless cached records from the global linked list (Del D2). The algorithm continues to sweep and cache the following annotation records for Q2 (no hit). Iteratively, VarNote accurately finds all intersections for Q3 (hit D3, D5, D6), deletes never used records from the global linked list (Del D1, D4) until the last query interval ends.



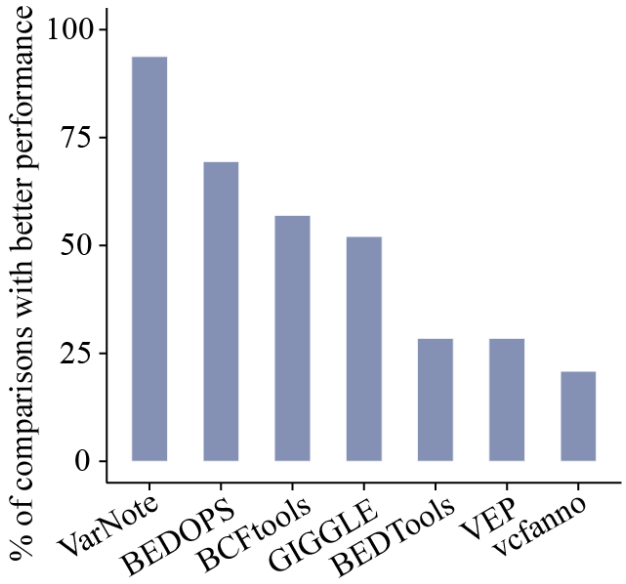
Supplemental Figure S5. The genomic distribution of six query variant datasets across the human reference genome, including variant call result of 1000 Genomes project phase3 (1000G_p3), variant call result of 10X Genomics Chromium whole genome sequencing for NA12878 (NA12878_WGS), variant call result of Nextera Rapid Capture Exome and Expanded Exome whole exome sequencing for NA12878 (NA12878_WES), variant call result of Ion Ampliseq Exome capture sequencing data for NA12878 (NA12878_Amp), genotype call result of Affymetrix Genome-Wide Human SNP Array 6.0 data for A375 cell line (A375_chip) and somatic mutation call result of whole exome sequencing data for A375 cell line (A375_SM). These datasets span the highly unbalanced and sparse queries to the highly balanced and dense queries.



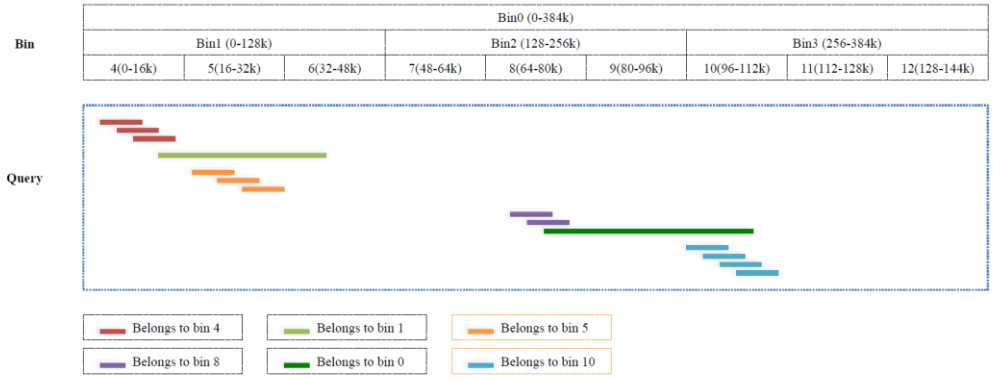
Supplemental Figure S6. The genomic distribution of three annotation databases across human reference genome, including functional prediction and annotation of all potential non-synonymous SNVs (dbNSFP), Cistrome aggregated ChIP-seq peak calling result of human transcription factors (Cistrome_TF), CADD deleteriousness score and related annotation of all possible SNVs (CADD_anno).



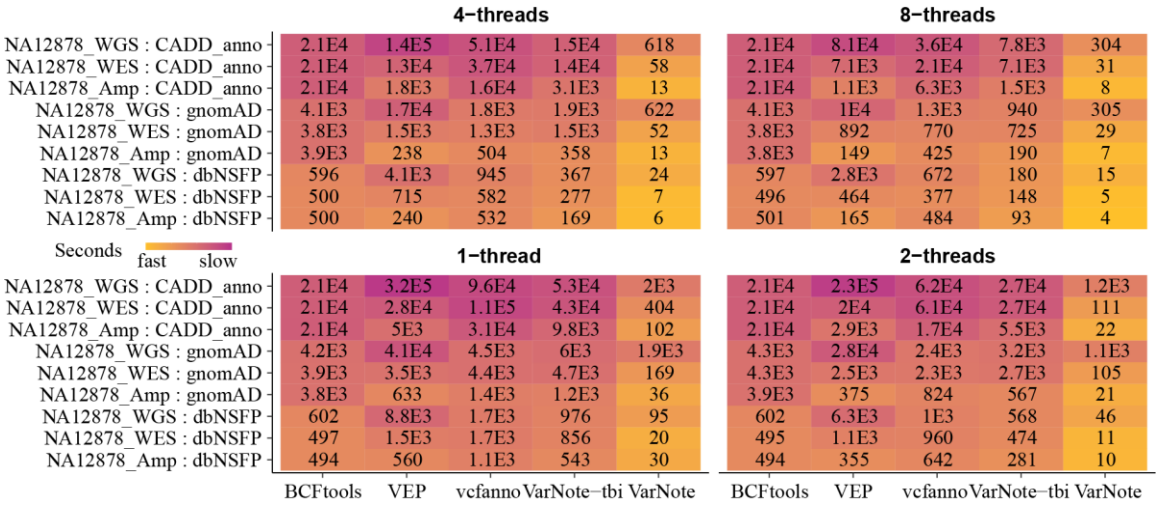
Supplemental Figure S7. The runtime comparisons of interval-level annotations among six methods. Six query variant datasets (A375_SM, NA12878_Amp, A375_chip, NA12878_WES, NA12878_WGS and 1000G_p3) and four genome-wide annotation databases (Cistrome_TF, dbNSFP, CADD_score and CADD_anno) were used.



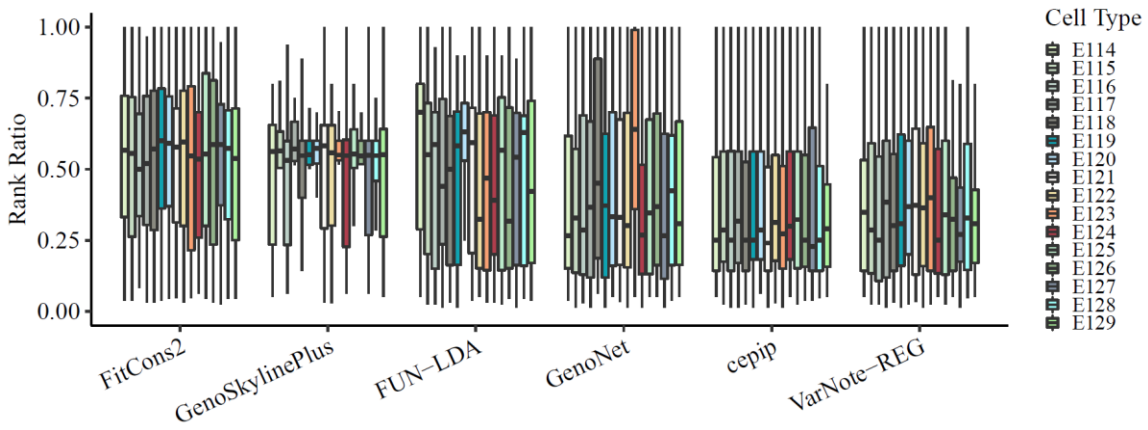
Supplemental Figure S8. The runtime comparisons between VarNote and other tools for 144 benchmarks. VarNote outperforms other method in 93.75% comparisons, with only nine evaluations are slightly behindhand in seconds for small queries or databases.



Supplemental Figure S9. Random-sweep searching algorithm based on Tabix index implemented by VarNote. Each chromosome is partitioned into bins span 128kb intervals. Query intervals belonging to separate bins will be grouped together. Tabix algorithm is used to identify and merge associated bins. VarNote chromosome-sweep algorithm is used to intersect annotation records across associated bins.



Supplemental Figure S10. The runtime comparisons of variant annotation using multi-threading. Three variant call results as query datasets (NA12878_Amp, NA12878_WES and NA12878_WGS) and three genome-wide annotation databases (dbNSFP, gnomAD and CADD_anno) were used. VarNote-tbi is the new multi-threading algorithm only based on Tabix index.



Supplemental Figure S11. The boxplot of rank ratio for each PICS causal variant in the GWAS signals across 16 ENCODE cell types and six different prioritization methods, the rank ratio is measured by the rank of observed variant/total number of investigated variants (including extended highly-linked variant in LD) in each GWAS signal.

Supplemental Tables

Supplemental Table S1. The query datasets and annotation databases used in the performance evaluation

	Dataset	Source
Query Datasets	1000G_p3	Variant call result of 1000 Genomes project phase3: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz
	NA12878_WGS	Variant call result of 10X Genomics Chromium whole genome sequencing for NA12878 from GIAB: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.1_09302016/NA12878_hg19/NA12878_hg19_phased_variants.vcf.gz
	NA12878_WES	Variant call result of Nextera Rapid Capture Exome and Expanded Exome whole exome sequencing for NA12878 from GIAB: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/project.NIST.hc.snps.indels.vcf
	NA12878_Amp	Variant call result of Ion Ampliseq Exome capture sequencing data for NA12878 from GIAB: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/IonTorrent_TVC_04302015/TSVC_variants.vcf
	A375_chip	Genotype call result of Affymetrix Genome-Wide Human SNP Array 6.0 data for A375 cell line from GDSC: https://www.cancerrxgene.org/
	A375_SM	Somatic mutation call result of whole exome sequencing data for A375 cell line from GDSC: https://www.cancerrxgene.org/
Annotation Databases	CADD_score	CADD v1.3 raw score and PHRED score for all possible SNVs (80G): https://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs.tsv.gz
	CADD_anno	CADD v1.3 all annotations for all possible SNVs (344G): https://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs_inclAnno.tsv.gz
	dbNSFP	dbNSFP v3.4a all functional predictions (16G): ftp://dbnsfp:dbnsfp@dbnsfp.softgenetics.com/dbNSFPv3.4a.zip
	gnomAD	gnomAD r2.0.2 Genomes all known variants (86G): https://gnomad.broadinstitute.org/
	Cistrome_TF	CistromeDB 20181120 aggregated human transcription regulator ChIP-seq peaks (12G): http://cistrome.org/db

Supplemental Table S2. The software information and running parameters used in the evaluation of interval-level overlap annotation

Tool	Version	Command line
BCFtools	1.6 (using htslib 1.6)	bcftools annotate -a {DATABASE FILE} -c CHROM,FROM,TO,ANN -h {HEADER FILE} --threads 1 {QUERY FILE}
BEDOPS	2.4.37 (megarow)	bedops -i {QUERY FILE} {DATABASE FILE}
BEDTools	v2.27.1	bedtools intersect -wa -wb -a {QUERY FILE} -b {DATABASE File} -sorted
GIGGLE	v0.6.3	giggle search -i {DATABASE INDEX DIRECTORY} -q {QUERY FILE} -o -v
vcfanno	0.2.8 (built with go1.8)	vcfanno -p 1 {DATABASE CONFIGURE FILE} {QUERY FILE}
VEP	91.1	vep --dir {VEP DATABASE DIRECTORY} --assembly GRCh37 --vcf --format vcf -i {QUERY FILE} -custom {DATABASE FILE},ANN,bed,overlap,1 --offline --fork 1
VarNote	v1.1	java -jar VarNote.jar Intersect -Q {QUERY FILE} -D {DATABASE FILE} -T 1

Supplemental Table S3. Summary information of 144 runtime comparisons between VarNote and other methods

Comparison	Test datasets	VarNote runtime (seconds)	Compared tool runtime (seconds)	Speedup
VarNote vs. BCFtools	1000G_p3 vs. CADD_anno	11356.00	22665.00	2.00
VarNote vs. BEDOPS	1000G_p3 vs. CADD_anno	11356.00	21895.00	1.93
VarNote vs. BEDTools	1000G_p3 vs. CADD_anno	11356.00	138318.00	12.18
VarNote vs. GIGGLE	1000G_p3 vs. CADD_anno	11356.00	624065.00	54.95
VarNote vs. VEP	1000G_p3 vs. CADD_anno	11356.00	618069.52	54.43
VarNote vs. vcfanno	1000G_p3 vs. CADD_anno	11356.00	150495.00	13.25
VarNote vs. BCFtools	1000G_p3 vs. CADD_score	2492.52	3196.39	1.28
VarNote vs. BEDOPS	1000G_p3 vs. CADD_score	2492.52	3316.72	1.33
VarNote vs. BEDTools	1000G_p3 vs. CADD_score	2492.52	13386.00	5.37
VarNote vs. GIGGLE	1000G_p3 vs. CADD_score	2492.52	307813.00	123.49
VarNote vs. VEP	1000G_p3 vs. CADD_score	2492.52	53617.49	21.51
VarNote vs. vcfanno	1000G_p3 vs. CADD_score	2492.52	33408.05	13.40
VarNote vs. BCFtools	1000G_p3 vs. Cistrome_TF	8681.00	2343.35	0.27
VarNote vs. BEDOPS	1000G_p3 vs. Cistrome_TF	8681.00	1613.43	0.19
VarNote vs. BEDTools	1000G_p3 vs. Cistrome_TF	8681.00	3138.06	0.36
VarNote vs. GIGGLE	1000G_p3 vs. Cistrome_TF	8681.00	234685.00	27.03
VarNote vs. VEP	1000G_p3 vs. Cistrome_TF	8681.00	202897.00	23.37
VarNote vs. vcfanno	1000G_p3 vs. Cistrome_TF	8681.00	2438.10	0.28
VarNote vs. BCFtools	1000G_p3 vs. dbNSFP	374.20	750.69	2.01
VarNote vs. BEDOPS	1000G_p3 vs. dbNSFP	374.20	675.86	1.81
VarNote vs. BEDTools	1000G_p3 vs. dbNSFP	374.20	2007.84	5.37
VarNote vs. GIGGLE	1000G_p3 vs. dbNSFP	374.20	42901.00	114.65
VarNote vs. VEP	1000G_p3 vs. dbNSFP	374.20	202757.00	541.84
VarNote vs. vcfanno	1000G_p3 vs. dbNSFP	374.20	2661.92	7.11
VarNote vs. BCFtools	A375_SM vs. CADD_anno	75.99	22865.00	300.89
VarNote vs. BEDOPS	A375_SM vs. CADD_anno	75.99	21054.00	277.06
VarNote vs. BEDTools	A375_SM vs. CADD_anno	75.99	79781.00	1049.89
VarNote vs. GIGGLE	A375_SM vs. CADD_anno	75.99	270.60	3.56
VarNote vs. VEP	A375_SM vs. CADD_anno	75.99	4338.00	57.09
VarNote vs. vcfanno	A375_SM vs. CADD_anno	75.99	38313.00	504.18
VarNote vs. BCFtools	A375_SM vs. CADD_score	38.99	3321.12	85.18
VarNote vs. BEDOPS	A375_SM vs. CADD_score	38.99	3215.11	82.46
VarNote vs. BEDTools	A375_SM vs. CADD_score	38.99	6656.00	170.71
VarNote vs. GIGGLE	A375_SM vs. CADD_score	38.99	171.71	4.40
VarNote vs. VEP	A375_SM vs. CADD_score	38.99	447.71	11.48
VarNote vs. vcfanno	A375_SM vs. CADD_score	38.99	5344.13	137.06
VarNote vs. BCFtools	A375_SM vs. Cistrome_TF	19.04	106.42	5.59
VarNote vs. BEDOPS	A375_SM vs. Cistrome_TF	19.04	80.52	4.23
VarNote vs. BEDTools	A375_SM vs. Cistrome_TF	19.04	224.79	11.81
VarNote vs. GIGGLE	A375_SM vs. Cistrome_TF	19.04	158.60	8.33
VarNote vs. VEP	A375_SM vs. Cistrome_TF	19.04	185.04	9.72
VarNote vs. vcfanno	A375_SM vs. Cistrome_TF	19.04	499.15	26.22
VarNote vs. BCFtools	A375_SM vs. dbNSFP	15.65	567.43	36.26
VarNote vs. BEDOPS	A375_SM vs. dbNSFP	15.65	526.01	33.61
VarNote vs. BEDTools	A375_SM vs. dbNSFP	15.65	1570.74	100.37
VarNote vs. GIGGLE	A375_SM vs. dbNSFP	15.65	47.72	3.05
VarNote vs. VEP	A375_SM vs. dbNSFP	15.65	429.66	27.45
VarNote vs. vcfanno	A375_SM vs. dbNSFP	15.65	1528.87	97.69
VarNote vs. BCFtools	A375_chip vs. CADD_anno	599.43	21648.00	36.11
VarNote vs. BEDOPS	A375_chip vs. CADD_anno	599.43	20461.00	34.13
VarNote vs. BEDTools	A375_chip vs. CADD_anno	599.43	74452.00	124.20
VarNote vs. GIGGLE	A375_chip vs. CADD_anno	599.43	2500.85	4.17
VarNote vs. VEP	A375_chip vs. CADD_anno	599.43	57313.00	95.61
VarNote vs. vcfanno	A375_chip vs. CADD_anno	599.43	100672.00	167.95
VarNote vs. BCFtools	A375_chip vs. CADD_score	146.93	3207.64	21.83
VarNote vs. BEDOPS	A375_chip vs. CADD_score	146.93	3020.57	20.56
VarNote vs. BEDTools	A375_chip vs. CADD_score	146.93	6161.00	41.93
VarNote vs. GIGGLE	A375_chip vs. CADD_score	146.93	1570.83	10.69
VarNote vs. VEP	A375_chip vs. CADD_score	146.93	5398.00	36.74
VarNote vs. vcfanno	A375_chip vs. CADD_score	146.93	12372.27	84.21
VarNote vs. BCFtools	A375_chip vs. Cistrome_TF	124.04	103.91	0.84
VarNote vs. BEDOPS	A375_chip vs. Cistrome_TF	124.04	80.70	0.65
VarNote vs. BEDTools	A375_chip vs. Cistrome_TF	124.04	230.20	1.86

VarNote vs. GIGGLE	A375_chip vs. Cistrome_TF	124.04	2345.14	18.91
VarNote vs. VEP	A375_chip vs. Cistrome_TF	124.04	2123.99	17.12
VarNote vs. vcfanno	A375_chip vs. Cistrome_TF	124.04	892.76	7.20
VarNote vs. BCFtools	A375_chip vs. dbNSFP	8.71	546.31	62.72
VarNote vs. BEDOPS	A375_chip vs. dbNSFP	8.71	509.24	58.47
VarNote vs. BEDTools	A375_chip vs. dbNSFP	8.71	1268.80	145.67
VarNote vs. GIGGLE	A375_chip vs. dbNSFP	8.71	441.38	50.68
VarNote vs. VEP	A375_chip vs. dbNSFP	8.71	1799.69	206.62
VarNote vs. vcfanno	A375_chip vs. dbNSFP	8.71	1715.83	197.00
VarNote vs. BCFtools	NA12878_Amp vs. CADD_anno	49.66	22689.00	456.89
VarNote vs. BEDOPS	NA12878_Amp vs. CADD_anno	49.66	21680.00	436.57
VarNote vs. BEDTools	NA12878_Amp vs. CADD_anno	49.66	75198.00	1514.26
VarNote vs. GIGGLE	NA12878_Amp vs. CADD_anno	49.66	151.46	3.05
VarNote vs. VEP	NA12878_Amp vs. CADD_anno	49.66	5637.00	113.51
VarNote vs. vcfanno	NA12878_Amp vs. CADD_anno	49.66	27697.00	557.73
VarNote vs. BCFtools	NA12878_Amp vs. CADD_score	19.18	3232.38	168.53
VarNote vs. BEDOPS	NA12878_Amp vs. CADD_score	19.18	3200.88	166.89
VarNote vs. BEDTools	NA12878_Amp vs. CADD_score	19.18	6362.00	331.70
VarNote vs. GIGGLE	NA12878_Amp vs. CADD_score	19.18	97.80	5.10
VarNote vs. VEP	NA12878_Amp vs. CADD_score	19.18	481.52	25.11
VarNote vs. vcfanno	NA12878_Amp vs. CADD_score	19.18	2871.12	149.69
VarNote vs. BCFtools	NA12878_Amp vs. Cistrome_TF	17.42	102.88	5.91
VarNote vs. BEDOPS	NA12878_Amp vs. Cistrome_TF	17.42	94.41	5.42
VarNote vs. BEDTools	NA12878_Amp vs. Cistrome_TF	17.42	223.01	12.80
VarNote vs. GIGGLE	NA12878_Amp vs. Cistrome_TF	17.42	218.17	12.52
VarNote vs. VEP	NA12878_Amp vs. Cistrome_TF	17.42	238.72	13.70
VarNote vs. vcfanno	NA12878_Amp vs. Cistrome_TF	17.42	400.80	23.01
VarNote vs. BCFtools	NA12878_Amp vs. dbNSFP	17.62	573.18	32.53
VarNote vs. BEDOPS	NA12878_Amp vs. dbNSFP	17.62	539.58	30.62
VarNote vs. BEDTools	NA12878_Amp vs. dbNSFP	17.62	1340.92	76.10
VarNote vs. GIGGLE	NA12878_Amp vs. dbNSFP	17.62	58.25	3.31
VarNote vs. VEP	NA12878_Amp vs. dbNSFP	17.62	714.00	40.52
VarNote vs. vcfanno	NA12878_Amp vs. dbNSFP	17.62	1192.21	67.66
VarNote vs. BCFtools	NA12878_WES vs. CADD_anno	352.47	22544.00	63.96
VarNote vs. BEDOPS	NA12878_WES vs. CADD_anno	352.47	21747.00	61.70
VarNote vs. BEDTools	NA12878_WES vs. CADD_anno	352.47	74965.00	212.68
VarNote vs. GIGGLE	NA12878_WES vs. CADD_anno	352.47	1153.61	3.27
VarNote vs. VEP	NA12878_WES vs. CADD_anno	352.47	29828.00	84.63
VarNote vs. vcfanno	NA12878_WES vs. CADD_anno	352.47	102118.00	289.72
VarNote vs. BCFtools	NA12878_WES vs. CADD_score	142.37	2917.42	20.49
VarNote vs. BEDOPS	NA12878_WES vs. CADD_score	142.37	3212.17	22.56
VarNote vs. BEDTools	NA12878_WES vs. CADD_score	142.37	6048.00	42.48
VarNote vs. GIGGLE	NA12878_WES vs. CADD_score	142.37	694.71	4.88
VarNote vs. VEP	NA12878_WES vs. CADD_score	142.37	2400.67	16.86
VarNote vs. vcfanno	NA12878_WES vs. CADD_score	142.37	10201.37	71.65
VarNote vs. BCFtools	NA12878_WES vs. Cistrome_TF	87.26	102.22	1.17
VarNote vs. BEDOPS	NA12878_WES vs. Cistrome_TF	87.26	95.09	1.09
VarNote vs. BEDTools	NA12878_WES vs. Cistrome_TF	87.26	218.88	2.51
VarNote vs. GIGGLE	NA12878_WES vs. Cistrome_TF	87.26	1383.11	15.85
VarNote vs. VEP	NA12878_WES vs. Cistrome_TF	87.26	1066.25	12.22
VarNote vs. vcfanno	NA12878_WES vs. Cistrome_TF	87.26	827.16	9.48
VarNote vs. BCFtools	NA12878_WES vs. dbNSFP	21.38	515.86	24.13
VarNote vs. BEDOPS	NA12878_WES vs. dbNSFP	21.38	530.14	24.80
VarNote vs. BEDTools	NA12878_WES vs. dbNSFP	21.38	1255.68	58.73
VarNote vs. GIGGLE	NA12878_WES vs. dbNSFP	21.38	242.23	11.33
VarNote vs. VEP	NA12878_WES vs. dbNSFP	21.38	1807.31	84.53
VarNote vs. vcfanno	NA12878_WES vs. dbNSFP	21.38	1562.93	73.10
VarNote vs. BCFtools	NA12878_WGS vs. CADD_anno	1960.29	25960.75	13.24
VarNote vs. BEDOPS	NA12878_WGS vs. CADD_anno	1960.29	21744.00	11.09
VarNote vs. BEDTools	NA12878_WGS vs. CADD_anno	1960.29	75880.00	38.71
VarNote vs. GIGGLE	NA12878_WGS vs. CADD_anno	1960.29	13855.00	7.07
VarNote vs. VEP	NA12878_WGS vs. CADD_anno	1960.29	331010.00	168.86
VarNote vs. vcfanno	NA12878_WGS vs. CADD_anno	1960.29	98069.00	50.03
VarNote vs. BCFtools	NA12878_WGS vs. CADD_score	1004.61	3510.78	3.49
VarNote vs. BEDOPS	NA12878_WGS vs. CADD_score	1004.61	3257.67	3.24
VarNote vs. BEDTools	NA12878_WGS vs. CADD_score	1004.61	6191.00	6.16
VarNote vs. GIGGLE	NA12878_WGS vs. CADD_score	1004.61	7950.00	7.91
VarNote vs. VEP	NA12878_WGS vs. CADD_score	1004.61	28902.00	28.77
VarNote vs. vcfanno	NA12878_WGS vs. CADD_score	1004.61	15361.17	15.29

VarNote vs. BCFtools	NA12878_WGS vs. Cistrome_TF	951.34	255.86	0.27
VarNote vs. BEDOPS	NA12878_WGS vs. Cistrome_TF	951.34	218.20	0.23
VarNote vs. BEDTools	NA12878_WGS vs. Cistrome_TF	951.34	535.84	0.56
VarNote vs. GIGGLE	NA12878_WGS vs. Cistrome_TF	951.34	14432.00	15.17
VarNote vs. VEP	NA12878_WGS vs. Cistrome_TF	951.34	13163.00	13.84
VarNote vs. vcfanno	NA12878_WGS vs. Cistrome_TF	951.34	981.81	1.03
VarNote vs. BCFtools	NA12878_WGS vs. dbNSFP	76.03	727.30	9.57
VarNote vs. BEDOPS	NA12878_WGS vs. dbNSFP	76.03	627.57	8.25
VarNote vs. BEDTools	NA12878_WGS vs. dbNSFP	76.03	1310.52	17.24
VarNote vs. GIGGLE	NA12878_WGS vs. dbNSFP	76.03	2652.47	34.89
VarNote vs. VEP	NA12878_WGS vs. dbNSFP	76.03	11439.00	150.45
VarNote vs. vcfanno	NA12878_WGS vs. dbNSFP	76.03	1704.22	22.42

Supplemental Table S4. The running parameters used in the evaluation of variant-level annotation

Tool	Database dataset	Database format	Extracted Features	Command line
BCFtools	dbNSFP	BED	SIFT_score, Polyphen2_HDIV_score, MutationAssessor_score, M-CAP_score, REVEL_score	bcftools annotate -a {DATABASE FILE} -c {BED FIELDS} -h {HEADER LINES} --threads {THREADS} {QUERY FILE}
	gnomAD	VCF	AF, AN, GC, CSQ	bcftools annotate -a {DATABASE FILE} -c {FIELDS} --threads {THREADS} {QUERY FILE}
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	bcftools annotate -a {DATABASE FILE} -c {BED FIELDS} -h {HEADER LINES} --threads {THREADS} {QUERY FILE}
VarNote	dbNSFP	BED	SIFT_score, Polyphen2_HDIV_score, MutationAssessor_score, M-CAP_score, REVEL_score	java -jar VarNote.jar Annotation -Q {QUERY FILE} -D:db,mode=1 {DATABASE FILE} -T {THREADS} -A {FIELDS}
	gnomAD	VCF	AF, AN, GC, CSQ	
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	
vcfanno	dbNSFP	BED	SIFT_score, Polyphen2_HDIV_score, MutationAssessor_score, M-CAP_score, REVEL_score	vcfanno -p {THREADS} {CONFIGURE FILE} {QUERY FILE}
	gnomAD	VCF	AF, AN, GC, CSQ	
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	
VEP	dbNSFP	BED	SIFT_score, Polyphen2_HDIV_score, MutationAssessor_score, M-CAP_score, REVEL_score	vep --dir {VEP DATABASE} --assembly GRCh37 --vcf --format vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},bed,exact,0 --offline --fork {THREADS}
	gnomAD	VCF	AF, AN, GC	vep --dir {VEP DATABASE} --assembly GRCh37 --vcf --format vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},vcf,exact,0,{DATABASE FIELDS} --offline --fork {THREADS}
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	vep --dir {VEP DATABASE} --assembly GRCh37 --vcf --format vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},bed,exact,0 --offline --fork {THREADS}

Supplemental Table S5. The curation list of rare pathogenic regulatory variants and experimental validation evidence

CHR	POS	rsID	REF	ALT	Disease	Functional Assays	Molecular Effect	Source	PMID
1	11905784	rs12565	T	G	Familial atrial fibrillation	EMSA, luciferase reporter assay (HEK293)	decreases binding affinity for REST to prompt NPPA expression	HGMD	22496669
1	101184877	rs3783605	A	G	Thromboembolic diseases, asthma, and multiple myeloma	ChIP (EBV-transformed lymphoblast), luciferase reporter assay (Jurkat)	increases binding affinity for ETS2	HGMD	17431880
1	209878286	rs13306421	G	A	Metabolic syndrome	luciferase reporter assay (CHO)	prompts HSD11B1 expression	HGMD	19934376
3	9791918	rs56387615	G	C	Gastric cancer	luciferase reporter assay (HEK293, HeLa)	decreases promoter activity to repress OGG1 expression	HGMD	21822670
3	9791948	rs1801129	A	G	Type 2 diabetes mellitus (T2DM)	luciferase reporter assay (HeLa, HEK293)	decreases promoter activity to repress OGG1 expression	HGMD	20562008
3	9791953	rs1801126	G	T	Type II epithelial ovarian cancer (EOC)	luciferase reporter assay (HEK293)	represses OGG1 expression	HGMD	21997177
3	37035012	rs587779001	C	A	Lynch syndrome	luciferase reporter assay (HEK293)	represses MLH1 expression	ClinVar, Genomiser	21840485
5	147211355	rs191068215	C	T	Chronic pancreatitis	luciferase reporter assay (AR42J, 266-6, HEK293T)	decreases promoter activity to repress SPINK1 expression	NCBoost, HGMD	25792561
6	1610252	rs77888940	C	G	Primary congenital glaucoma	luciferase reporter assay (HEK293T)	increases protein levels and transactivation activity of FOXC1 by disrupt a consensus sequence for a terminal oligopyrimidine tract	ClinVar, Genomiser	26220699
6	88299677	rs201150141	T	C	Pontocerebellar hypoplasia	luciferase reporter assay (HEK293T)	decreases promoter activity to repress RARS2 expression	ClinVar, Genomiser, CDTS	25809939
11	5248393	rs34883338	G	A	Familial hypercholesterolemia	luciferase reporter assay (HepG2), EMSA (SL-2)	decreases binding affinity for Sp1 and decreases promoter activity to repress LDLR	NCBoost, HGMD, CDTS	11792717
11	102596480	rs11225395	A	C	Preterm premature rupture of the membranes	luciferase reporter assay (BeWo, HTR-8/Svneo, JEG-3), EMSA (BeWo)	decreases binding affinity for unknown TF(s) and increases promoter activity to prompts MMP8 expression	HGMD	15367487
12	52308396	rs201378973	G	C	Hereditary hemorrhagic telangiectasia (HHT)	luciferase reporter assay (HUVEC), EMSA (HUVEC)	decreases binding affinity for Sp1 to repress ACVRL1 expression	NCBoost	23460919
16	67520735	rs8047574	C	T	Low body fatness	luciferase reporter assay (NCI-h295R)	increases promoter activity to prompt AGRP expression	HGMD	15121772
16	72088461	rs5471	A	C	Ahaptoglobinaemia	luciferase reporter assay (HepG2)	decreases promoter activity to repress HP expression	HGMD	14616769

17	69108655	rs1859961	A	G	Prostate cancer	ChIP (LNCaP), luciferase reporter assayq (LNCaP)	decreases binding affinity for FOXA1, increases binding affinity for AP-1 and increases enhancer activity to prompt SOX9 expression	HGMD	22665440
19	49468616	rs398124638	G	C	Hereditary hyperferritinaemia-cataract syndrome	EMSA (K562)	decreases binding affinity for iron-regulatory proteins	ClinVar, Genomiser, CDTS	10759702
20	23030292	rs16984852	C	A	Venous thrombosis	luciferase reporter assay (HUVECs, HEK293, COS-7)	repress THBD expression	ClinVar, Genomiser	23332921

Supplemental Table S6. The pseudocode of VarNote random-sweep searching algorithm

***Main*:**
read input file
map input queries into groups according to given CPU threads
foreach group do
    SearchInGroup
end
reduce group results to final output

***SearchInGroup*:**
foreach query in group do
   foreach database do
      SearchHitsInDatabase
    end
    merge hits for each query
end

***SearchHitsInDatabase*:**
if query is first read in chromosome then
   foreach database do
      loadIndexForChr
   end

foreach SROB in index do
    if query.end < SROB.min then nextBlock
   elseif query.beg <= SROB.max then do
      SearchHitsInBlock
   break loop
end

***SearchHitsInBlock*:**
foreach linklist.record in linklist do
    if linklist.record.end < query.beg then remove linklist.record from linklist
    elseif linklist.record.beg < query.end then **RandomAccessAnno** from linklist.record
end

while (next ROB in Block from pointer is not null)
   do
      if ROB.record.end <= query.beg then continue
      elseif ROB.record.beg < query.end then **RandomAccessAnno** and put ROB.record into linklist
      else set a pointer to current ROB
   end

* ***loadIndexForChr*:** will load all SROBs for current chromosome from index file
* ***RandomAccessAnno*:** will randomly access annotations from OB of database file

Supplemental Codes

Supplemental_Code.zip: The file contains all scripts and source codes for methods evaluation, manuscript results and software functions.

All latest source code and scripts for methods evaluation and manuscript results are also available at https://github.com/mulinlab/VarNote