## **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**

Figure S1. Architecture of VarNote index system	3
Figure S2. The construction of VarNote positioning file and index file	4
Figure S3. The VarNote index format	5
Figure S4. Process of chromosome sweeping	6
Figure S5. The genomic distribution of six query variant datasets	7
Figure S6. The genomic distribution of three annotation databases	8
Figure S7. The runtime comparisons of interval-level annotations among six methods	9
Figure S8. The runtime comparisons between VarNote and other tools	10
Figure S9. Random-sweep searching algorithm based on Tabix index	10
Figure S10. The runtime comparisons of variant annotation using multi-threading	11
Figure S11. The boxplot of rank ratio for PICS causal variant across conditions	12



**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 ( $2^{6} \sim 2^{10}$  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).

Fiel	Field		Description	Туре
Blo	Block Start Flag		<ul> <li>-1 means block start and first record distance is 1;</li> <li>-2 means block start and first record distance is a unsigned byte value;</li> <li>-3 means block start and first record distance is a unsigned short value;</li> <li>-4 means block start and first record distance is a int value.</li> </ul>	byte
	First Record in	n Block		
	Block Addres	iS	64-bit bgzip address of record in database file.	long
	Record Dista	nce	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 Offs	et	Average block offset of all records in this block.	short, int
	Other Records	s 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]
	Directio	n 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 Dista	nce	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 be Offset to Ave	etween Block rage 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
Fiel	d		Description	Туре
Blo	ck Count		Total block in current chromosome	int
	Summary of E	ach Block		
	Block Addres	s	Address of this block in the VarNote file.	long
	Start Position	1	The start position (Beg) of the first record in this block.	int
	Block Distance	се	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).

1000G_p3 : CADD_anno	1.4E5	2.2E4	2.3E4	6.2E5	6.2E5	1.5E5	1.1E4
NA12878_WGS : CADD_anno -	7.6E4	2.2E4	2.6E4	3.3E5	1.4E4	9.8E4	2E3
NA12878_WES : CADD_anno -	7.5E4	2.2E4	2.3E4	3E4	1.2E3	1E5	352
A375_chip : CADD_anno -	7.4E4	2E4	2.2E4	5.7E4	2.5E3	1E5	599
NA12878_Amp : CADD_anno -	7.5E4	2.2E4	2.3E4	5.6E3	151	2.8E4	50
A375_SM : CADD_anno -	8E4	2.1E4	2.3E4	4.3E3	271	3.8E4	76
1000G_p3 : CADD_score -	1.3E4	3.3E3	3.2E3	5.4E4	3.1E5	3.3E4	2.5E3
NA12878_WGS : CADD_score	6.2E3	3.3E3	3.5E3	2.9E4	8E3	1.5E4	1E3
NA12878_WES : CADD_score	6E3	3.2E3	2.9E3	2.4E3	695	1E4	142
A375_chip : CADD_score -	6.2E3	3E3	3.2E3	5.4E3	1.6E3	1.2E4	147
NA12878_Amp : CADD_score -	6.4E3	3.2E3	3.2E3	482	98	2.9E3	19
A375_SM : CADD_score -	6.7E3	3.2E3	3.3E3	448	172	5.3E3	39
1000G_p3 : dbNSFP -	2E3	676	751	2E5	4.3E4	2.7E3	374
NA12878_WGS : dbNSFP -	1.3E3	628	727	1.1E4	2.7E3	1.7E3	76
NA12878_WES : dbNSFP -	1.3E3	530	516	1.8E3	242	1.6E3	21
A375_chip : dbNSFP -	1.3E3	509	546	1.8E3	441	1.7E3	9
NA12878_Amp : dbNSFP -	1.3E3	540	573	714	58	1.2E3	18
A375_SM : dbNSFP -	1.6E3	526	567	430	48	1.5E3	16
1000G_p3 : Cistrome_TF -	3.1E3	1.6E3	2.3E3	2E5	2.3E5	2.4E3	8.7E3
NA12878_WGS : Cistrome_TF -	536	218	256	1.3E4	1.4E4	982	951
NA12878_WES : Cistrome_TF -	219	95	102	1.1E3	1.4E3	827	87
A375_chip : Cistrome_TF -	230	81	104	2.1E3	2.3E3	893	124
NA12878_Amp : Cistrome_TF -	223	94	103	239	218	401	17
A375_SM : Cistrome_TF -	225	81	106	185	159	499	19
Seconds fast slow	BEDTools	BEDOPS	BCFtools	VEP	GIGGLE	vcfanno	VarNote

**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.

	4-threads				8-threads						
NA12878 WGS : CADD anno-	2.1E4	1.4E5	5.1E4	1.5E4	618		2.1E4	8.1E4	3.6E4	7.8E3	304
NA12878 WES : CADD anno -	2.1E4	1.3E4	3.7E4	1.4E4	58		2.1E4	7.1E3	2.1E4	7.1E3	31
NA12878 Amp : CADD anno -	2.1E4	1.8E3	1.6E4	3.1E3	13		2.1E4	1.1E3	6.3E3	1.5E3	8
NA12878 WGS : gnomAD -	4.1E3	1.7E4	1.8E3	1.9E3	622		4.1E3	1E4	1.3E3	940	305
NA12878 WES : gnomAD -	3.8E3	1.5E3	1.3E3	1.5E3	52		3.8E3	892	770	725	29
NA12878 Amp : gnomAD -	3.9E3	238	504	358	13		3.8E3	149	425	190	7
NA12878 WGS : dbNSFP -	596	4.1E3	945	367	24		597	2.8E3	672	180	15
NA12878 WES : dbNSFP -	500	715	582	277	7		496	464	377	148	5
NA12878_Amp : dbNSFP -	500	240	532	169	6		501	165	484	93	4
Seconds		1_throad						2_throad	le.		
fast slow			1-uneau	u					2-unreau	15	
NA12878 WGS : CADD anno -	2.1E4	3.2E5	9.6E4	5.3E4	2E3		2.1E4	2.3E5	6.2E4	2.7E4	1.2E3
NA12878_WGS : CADD_anno - NA12878_WES : CADD_anno -	2.1E4 2.1E4	3.2E5 2.8E4	9.6E4 1.1E5	5.3E4 4.3E4	2E3 404		2.1E4 2.1E4	2.3E5 2E4	6.2E4 6.1E4	2.7E4 2.7E4	1.2E3 111
NA12878_WGS : CADD_anno - NA12878_WES : CADD_anno - NA12878_WES : CADD_anno - NA12878_Amp : CADD_anno -	2.1E4 2.1E4 2.1E4	3.2E5 2.8E4 5E3	9.6E4 1.1E5 3.1E4	5.3E4 4.3E4 9.8E3	2E3 404 102		2.1E4 2.1E4 2.1E4	2.3E5 2E4 2.9E3	6.2E4 6.1E4 1.7E4	2.7E4 2.7E4 5.5E3	1.2E3 111 22
hast slow NA12878_WGS : CADD_anno - NA12878_WES : CADD_anno - NA12878_Amp : CADD_anno - NA12878_WGS : gnomAD -	2.1E4 2.1E4 2.1E4 4.2E3	3.2E5 2.8E4 5E3 4.1E4	9.6E4 1.1E5 3.1E4 4.5E3	5.3E4 4.3E4 9.8E3 6E3	2E3 404 102 1.9E3		2.1E4 2.1E4 2.1E4 4.3E3	2.3E5 2E4 2.9E3 2.8E4	6.2E4 6.1E4 1.7E4 2.4E3	2.7E4 2.7E4 5.5E3 3.2E3	1.2E3 111 22 1.1E3
hast slow NA12878_WGS : CADD anno - NA12878_WES : CADD_anno - NA12878_Amp : CADD_anno - NA12878_WGS : gnomAD - NA12878_WES : gnomAD -	2.1E4 2.1E4 2.1E4 4.2E3 3.9E3	3.2E5 2.8E4 5E3 4.1E4 3.5E3	9.6E4 1.1E5 3.1E4 4.5E3 4.4E3	5.3E4 4.3E4 9.8E3 6E3 4.7E3	2E3 404 102 1.9E3 169		2.1E4 2.1E4 2.1E4 4.3E3 4.3E3	2.3E5 2E4 2.9E3 2.8E4 2.5E3	6.2E4 6.1E4 1.7E4 2.4E3 2.3E3	2.7E4 2.7E4 5.5E3 3.2E3 2.7E3	1.2E3 111 22 1.1E3 105
Iast slow NA12878_WGS: CADD_anno - NA12878_WES: CADD_anno - NA12878_Amp: CADD_anno - NA12878_WGS: gnomAD - NA12878_WGS: gnomAD - NA12878_WES: gnomAD - NA12878_Amp: gnomAD -	2.1E4 2.1E4 2.1E4 4.2E3 3.9E3 3.8E3	3.2E5 2.8E4 5E3 4.1E4 3.5E3 633	9.6E4 1.1E5 3.1E4 4.5E3 4.4E3 1.4E3	5.3E4 4.3E4 9.8E3 6E3 4.7E3 1.2E3	2E3 404 102 1.9E3 169 36		2.1E4 2.1E4 2.1E4 4.3E3 4.3E3 3.9E3	2.3E5 2E4 2.9E3 2.8E4 2.5E3 375	6.2E4 6.1E4 1.7E4 2.4E3 2.3E3 824	2.7E4 2.7E4 5.5E3 3.2E3 2.7E3 567	1.2E3 111 22 1.1E3 105 21
Institution fast slow NA12878_WGS: CADD_anno- NA12878_WES: CADD_anno- NA12878_Amp: CADD_anno- NA12878_WGS: gnomAD- NA12878_WES: gnomAD- NA12878_Amp: gnomAD- NA12878_WGS: dbNSFP-	2.1E4 2.1E4 2.1E4 4.2E3 3.9E3 3.8E3 602	3.2E5 2.8E4 5E3 4.1E4 3.5E3 633 8.8E3	9.6E4 1.1E5 3.1E4 4.5E3 4.4E3 1.4E3 1.7E3	5.3E4 4.3E4 9.8E3 6E3 4.7E3 1.2E3 976	2E3 404 102 1.9E3 169 36 95		2.1E4 2.1E4 2.1E4 4.3E3 4.3E3 3.9E3 602	2.3E5 2E4 2.9E3 2.8E4 2.5E3 375 6.3E3	6.2E4 6.1E4 1.7E4 2.4E3 2.3E3 824 1E3	2.7E4 2.7E4 5.5E3 3.2E3 2.7E3 567 568	1.2E3 111 22 1.1E3 105 21 46
Iast slow NA12878_WGS: CADD_anno- NA12878_WES: CADD_anno- NA12878_Amp: CADD_anno- NA12878_MGS: gnomAD- NA12878_WGS: gnomAD- NA12878_WGS: dbNSFP- NA12878_WGS: dbNSFP- NA12878_WES: dbNSFP-	2.1E4 2.1E4 2.1E4 4.2E3 3.9E3 3.8E3 602 497	3.2E5 2.8E4 5E3 4.1E4 3.5E3 633 8.8E3 1.5E3	9.6E4 1.1E5 3.1E4 4.5E3 4.4E3 1.4E3 1.7E3 1.7E3	5.3E4 4.3E4 9.8E3 6E3 4.7E3 1.2E3 976 856	2E3 404 102 1.9E3 169 36 95 20		2.1E4 2.1E4 2.1E4 4.3E3 4.3E3 3.9E3 602 495	2.3E5 2E4 2.9E3 2.8E4 2.5E3 375 6.3E3 1.1E3	6.2E4 6.1E4 1.7E4 2.4E3 2.3E3 824 1E3 960	2.7E4 2.7E4 5.5E3 3.2E3 2.7E3 567 568 474	1.2E3 111 22 1.1E3 105 21 46 11
Iast slow NA12878_WGS: CADD_anno- NA12878_WES: CADD_anno- NA12878_Amp: CADD_anno- NA12878_MGS: gnomAD- NA12878_WGS: gnomAD- NA12878_WGS: dbNSFP- NA12878_WGS: dbNSFP- NA12878_Amp: dbNSFP- NA12878_Amp: dbNSFP-	2.1E4 2.1E4 2.1E4 4.2E3 3.9E3 3.8E3 602 497 494	3.2E5 2.8E4 5E3 4.1E4 3.5E3 633 8.8E3 1.5E3 560	9.6E4 1.1E5 3.1E4 4.5E3 4.4E3 1.4E3 1.7E3 1.7E3 1.1E3	5.3E4 4.3E4 9.8E3 6E3 4.7E3 1.2E3 976 856 543	2E3 404 102 1.9E3 169 36 95 20 30		2.1E4 2.1E4 2.1E4 4.3E3 4.3E3 3.9E3 602 495 494	2.3E5 2E4 2.9E3 2.8E4 2.5E3 375 6.3E3 1.1E3 355	6.2E4 6.1E4 1.7E4 2.4E3 2.3E3 824 1E3 960 642	2.7E4 2.7E4 5.5E3 3.2E3 2.7E3 567 568 474 281	1.2E3 111 22 1.1E3 105 21 46 11 10

**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**

Table S1. The query datasets and annotation databases for performance evaluation	14
Table S2. The software and running parameters for interval-level overlap annotation .	15
Table S3. Summary information of 144 runtime comparisons	16
Table S4. The running parameters used in the evaluation of variant-level annotation	19
Table S5. The curation list of rare pathogenic regulatory variants	20
Table S6. The pseudocode of VarNote random-sweep searching algorithm	22

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

	Dataset	Source
	1000G_p3	Variant call result of 1000 Genomes project phase3: <u>ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_shapeit2_m_vncall_integrated_v5b.20130502.sites.vcf.gz</u>
Datasets	NA12878_WGS	Variant call result of 10X Genomics Chromium whole genome sequencing for NA12878 from GIAB: <u>ftp://ftp-</u> <u>trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongR</u> <u>anger2.1_09302016/NA12878_hg19/NA12878_hg19_phased_variants.vcf.gz</u>
	NA12878_WES	Variant call result of Nextera Rapid Capture Exome and Expanded Exome whole exome sequencing for NA12878 from GIAB: <u>ftp://ftp-</u> <u>trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome</u> / <u>project.NIST.hc.snps.indels.vcf</u>
Query	NA12878_Amp	Variant call result of Ion Ampliseq Exome capture sequencing data for NA12878 from GIAB: <u>ftp://ftp-</u> <u>trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/IonTorrent_TVC_04302015/TSV</u> <u>C_variants.vcf</u>
	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/
	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
atabases	CADD_anno	CADD v1.3 all annotations for all possible SNVs (344G): https://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs_inclAnn_ o.tsv.gz
tion Da	dbNSFP	dbNSFP v3.4a all functional predictions (16G): <u>ftp://dbnsfp@dbnsfp@dbnsfp.softgenetics.com/dbNSFPv3.4a.zip</u>
Annota	gnomAD	gnomAD r2.0.2 Genomes all known variants (86G): https://gnomad.broadinstitute.org/
7	Cistrome_TF	CistromeDB 20181120 aggregated human transcription regulator ChIP-seq peaks (12G): <u>http://cistrome.org/db</u>

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	vepdir {VEP DATABASE DIRECTORY}assembly GRCh37vcfformat vcf -i {QUERY FILE} -custom {DATABASE FILE},ANN,bed,overlap,1offlinefork 1
VarNote	v1.1	java -jar VarNote.jar Intersect -Q {QUERY FILE} -D {DATABASE FILE} -T 1

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

**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	1000 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. BCEtools	1000G n3 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	0.19
VarNote vs. GIGGLE	1000G p3 vs. Cistrome_TT	8681.00	234685.00	27.03
VarNote vs. UED	1000G_p3 vs. Cistrome_TT	8681.00	202807.00	27.03
VarNote vs. vEP	1000G_p5 vs. Cistrome_IF	8681.00	202897.00	23.37
Varinole vs. vcianno	1000G_p5 vs. Cistrome_IF	8081.00	2438.10	0.28
Varinole VS. BUFIOOIS	1000G p3 VS. dbNSFP	5/4.20	/ 30.09	2.01
varNote vs. BEDOPS	1000G_p3 vs. dbNSFP	374.20	0/5.86	1.81
varinote vs. BED fools	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 <sup>SM</sup> vs. CADD <sup>score</sup>	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. GIGGI F	A375 SM vs. dbNSFP	15.65	47.72	3 05
VarNote vs. VFP	A375 SM vs. dbNSFP	15.65	429.66	27 45
VarNote vs. vcfanno	A375 SM vs. dbNSFP	15.65	1528.87	97.60
VarNote vs. BCFtools	A375 chin ve CADD anno	500 43	21648.00	36.11
VarNote vs. BEDODS	A375 chin vs. CADD_anno	500 /2	20461.00	2/ 12
VarNote vs. BEDOr5	A375 chin vs. CADD_anno	500 /2	74452.00	10/ 20
VarNote vs. CICCLE	A375 chip vs. CADD_allilo	500 42	2500 85	124.20
VarNote vs. GIGGLE	AS / 5_ cmp vs. CADD_anno	500 42	2300.83	4.1/
Vallinute vs. VEP	AS / 3_ cnip vs. CADD_anno	599.43	3/313.00	95.61
Varinote vs. vcfanno	A3/5_cnip vs. CADD_anno	599.43	1006/2.00	167.95
varinote vs. BCFtools	A3/5_chip vs. CADD_score	146.93	3207.64	21.83
varNote vs. BEDOPS	A3/5_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
V M OFDT 1	A 275 shin va Cistuana TE	124.04	230.20	1.86

VarNote vs. GIGGLE	A375_chip vs. Cistrome_TF	124.04	2345.14	18.9
varNote vs. VEP	A3/5_chip vs. Cistrome_TF	124.04	2123.99	17.1
VarNote vs. vcfanno	A3/5_chip vs. Cistrome_TF	124.04	892.76	7.2
VarNote vs. BCFtools	A3/5_chip vs. dbNSFP	8.71	546.31	62.7
VarNote vs. BEDOPS	A375_chip vs. dbNSFP	8.71	509.24	58.4
VarNote vs. BEDTools	A375_chip vs. dbNSFP	8.71	1268.80	145.6
VarNote vs. GIGGLE	A375_chip vs. dbNSFP	8.71	441.38	50.6
VarNote vs. VEP	A375_chip vs. dbNSFP	8.71	1799.69	206.6
VarNote vs. vcfanno	A375_chip vs. dbNSFP	8.71	1715.83	197.0
VarNote vs. BCFtools	NA12878_Amp vs. CADD_anno	49.66	22689.00	456.8
VarNote vs. BEDOPS	NA12878_Amp vs. CADD_anno	49.66	21680.00	436.5
VarNote vs. BEDTools	NA12878_Amp vs. CADD_anno	49.66	75198.00	1514.2
VarNote vs. GIGGLE	NA12878_Amp vs. CADD_anno	49.66	151.46	3.0
VarNote vs. VEP	NA12878 Amp vs. CADD anno	49.66	5637.00	113.5
VarNote vs. vcfanno	NA12878 Amp vs. CADD anno	49.66	27697.00	557.7
VarNote vs. BCFtools	NA12878 Amp vs. CADD score	19.18	3232.38	168.5
VarNote vs. BEDOPS	NA12878 Amp vs. CADD score	19.18	3200.88	166.8
VarNote vs. BEDTools	NA12878 Amp vs. CADD score	19.18	6362.00	331.7
VarNote vs. GIGGLE	NA12878 Amp vs CADD score	19.18	97.80	51
VarNote vs. VEP	NA12878 Amp vs. CADD_score	19.18	481 52	25.1
VarNote vs. vcfanno	NA12878 Amp vs. CADD score	19.18	2871 12	149.6
VarNote vs. PCEtools	NA12878 Amp vs. Cistrome TE	17.10	102.88	5.0
VarNote vs. DEPt0015	NA12878 Amp vs. Cistrome_IT	17.42	04.41	5.9
VarNote vs. BEDOFS	NA12878 Amp vs. Cistrome_IT	17.42	222.01	12.9
VarNote vs. BEDTOOIS	NA12878 Amp vs. Cistrome_TF	17.42	225.01	12.0
Varinole vs. GIGGLE	NA12878_Amp vs. Cistrome_IF	17.42	210.17	12.3
VarNote vs. VEP	NA128/8_Amp vs. Cistrome_IF	17.42	238.72	13./
VarNote vs. vctanno	NA128/8_Amp vs. Cistrome_IF	17.42	400.80	23.0
VarNote vs. BCFtools	NA12878_Amp vs. dbNSFP	17.62	5/3.18	32.5
VarNote vs. BEDOPS	NA12878_Amp vs. dbNSFP	17.62	539.58	30.6
VarNote vs. BEDTools	NA12878_Amp vs. dbNSFP	17.62	1340.92	76.1
VarNote vs. GIGGLE	NA12878_Amp vs. dbNSFP	17.62	58.25	3.3
VarNote vs. VEP	NA12878_Amp vs. dbNSFP	17.62	714.00	40.5
VarNote vs. vcfanno	NA12878_Amp vs. dbNSFP	17.62	1192.21	67.6
VarNote vs. BCFtools	NA12878_WES vs. CADD_anno	352.47	22544.00	63.9
VarNote vs. BEDOPS	NA12878_WES vs. CADD_anno	352.47	21747.00	61.7
VarNote vs. BEDTools	NA12878_WES vs. CADD_anno	352.47	74965.00	212.6
VarNote vs. GIGGLE	NA12878 WES vs. CADD anno	352.47	1153.61	3.2
VarNote vs. VEP	NA12878 WES vs. CADD anno	352.47	29828.00	84.6
VarNote vs. vcfanno	NA12878 WES vs. CADD anno	352.47	102118.00	289.7
VarNote vs. BCFtools	NA12878 WES vs. CADD score	142.37	2917.42	20.4
VarNote vs. BEDOPS	NA12878 WES vs. CADD score	142.37	3212.17	22.5
VarNote vs. BEDTools	NA12878 WES vs. CADD score	142.37	6048.00	42.4
VarNote vs. GIGGLE	NA12878 WES vs. CADD score	142.37	694.71	4.8
VarNote vs. VEP	NA12878 WES vs. CADD score	142.37	2400 67	16.8
VarNote vs. vcfanno	NA12878 WES vs. CADD_score	142.37	10201 37	71.6
VarNote vs. RCEtools	NA12878 WES vs. CADD_SOLE	87.26	102 27	/1.0
VarNote ve REDODC	NA12070_WES VS. CISUOME_IF	87.20	05.00	1.1
VarNote va. DEDUPS	NA12070_WES vs. Cistrome_IF	87.20	23.07	1.0
VarNote vs. BED1001S	NA120/0_WES VS. UISTOME_IF	01.20	∠10.88 1292-11	2.3
VarNote vs. GIGGLE	NA12070_WES VS. UIStrome_IF	0/.20	1383.11	15.8
VarNote vs. VEP	NA128/8_WES VS. UIStrome_IF	87.20	1000.25	12.2
varivote vs. vctanno	NA128/8_WES vs. Cistrome_TF	87.26	827.16	9.4
VarNote vs. BCFtools	NA12878_WES vs. dbNSFP	21.38	515.86	24.1
varNote vs. BEDOPS	NA128/8_WES vs. dbNSFP	21.38	530.14	24.8
VarNote vs. BEDTools	NA12878_WES vs. dbNSFP	21.38	1255.68	58.7
VarNote vs. GIGGLE	NA12878_WES vs. dbNSFP	21.38	242.23	11.3
VarNote vs. VEP	NA12878_WES vs. dbNSFP	21.38	1807.31	84.5
VarNote vs. vcfanno	NA12878_WES vs. dbNSFP	21.38	1562.93	73.1
VarNote vs. BCFtools	NA12878_WGS vs. CADD_anno	1960.29	25960.75	13.2
VarNote vs. BEDOPS	NA12878_WGS vs. CADD_anno	1960.29	21744.00	11.0
VarNote vs. BEDTools	NA12878_WGS vs. CADD_anno	1960.29	75880.00	38.7
VarNote vs. GIGGLE	NA12878 WGS vs. CADD anno	1960.29	13855.00	7.0
VarNote vs. VEP	NA12878 WGS vs. CADD anno	1960.29	331010.00	168.8
VarNote vs. vcfanno	NA12878 WGS vs. CADD anno	1960.29	98069.00	50.0
VarNote vs. BCFtools	NA12878 WGS vs. CADD score	1004.61	3510.78	3 4
VarNote vs BEDOPS	NA12878 WGS vs CADD score	1004.61	3257.67	3.1
VarNote vs. BEDOI 5	NA12878 WGS vs. CADD_score	1004.61	6191.00	5.2
VarNote vs. GIGCI E	NA12878 WCS VS. CADD SCORE	1004.01	7050.00	7.0
VarNote vs. VED	NA12878 WGS VS. CADD_SCORE	1004.01	28002 00	7.9 ד סר
Vallaule vs. VEP	NA12070 WG5 VS. CADD_score	1004.01	26902.00	28.7
varinote vs. vcianno	NA128/8_WGS VS. CADD_score	1004.61	13301.17	15.2

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

Tool	Database dataset	Database format	Extracted Features	Command line				
DCEtaala	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}				
Der tools	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} -				
	gnomAD	VCF	AF, AN, GC, CSQ	Α (ΠΕΓΡΟ)				
	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}				
	gnomAD	VCF	AF, AN, GC, CSQ					
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	-				
	dbNSFP	BED	SIFT_score, Polyphen2_HDIV_score, MutationAssessor_score, M-CAP_score, REVEL_score	vepdir {VEP DATABASE}assembly GRCh37 vcfformat vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},bed,exact,0 offlinefork {THREADS}				
VEP	gnomAD VCF AF, AN, GC		AF, AN, GC	vepdir {VEP DATABASE}assembly GRCh37 vcfformat vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},vcf,exact,0,{DATABASE FIELDS}offline fork {THREADS}				
	CADD_anno	BED	GerpRS, fitCons, RawScore, PHRED	vepdir {VEP DATABASE}assembly GRCh37 vcfformat vcf -i {QUERY FILE} -custom {DATABASE FILE},{DATABASE ID},bed,exact,0 offlinefork {THREADS}				

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

CHR	POS	rsID	REF	ALT	Disease	Functional Assays	Molecular Effect	Source	PMID
1	11905784	rs12565	Т	G	Familial atrial fibrillation	EMSA, luciferase reporter assay (HEK293)	decreases binding affinity for REST to prompt NPPA expression	HGMD	22496669
1	101184877	rs3783605	А	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	А	Metabolic syndrome	luciferase reporter assay (CHO)	prompts HSD11B1 expression	HGMD	19934376
3	9791918	rs56387615	G	С	Gastric cancer	luciferase reporter assay (HEK293, HeLa)	decreases promoter activity to repress OGG1 expression	HGMD	21822670
3	9791948	rs1801129	А	G	Type 2 diabetes mellitus (T2DM)	luciferase reporter assay (HeLa, HEK293)	decreases promoter activity to repress OGG1 expression	HGMD	20562008
3	9791953	rs1801126	G	Т	Type II epithelial ovarian cancer (EOC)	luciferase reporter assay (HEK293)	represses OGG1 expression	HGMD	21997177
3	37035012	rs587779001	С	А	Lynch syndrome	luciferase reporter assay (HEK293)	represses MLH1 expression	ClinVar, Genomiser	21840485
5	147211355	rs191068215	С	Т	Chronic pancreatitis	luciferase reporter assay (AR42J, 266-6, HEK293T)	decreases promoter activity to repress SPINK1 expression	NCBoost, HGMD	25792561
6	1610252	rs77888940	С	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	Т	С	Pontocerebellar hypoplasia	luciferase reporter assay (HEK293T)	decreases promoter activity to repress RARS2 expression	ClinVar, Genomiser, CDTS	25809939
11	5248393	rs34883338	G	А	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	А	С	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	С	Hereditary hemorrhagic telangiectasia (HHT)	luciferase reporter assay (HUVEC), EMSA (HUVEC)	decreases binding affinity for Sp1 to repress ACVRL1 expression	NCBoost	23460919
16	67520735	rs8047574	С	Т	Low body fatness	luciferase reporter assay (NCI-h295R)	increases promoter activity to prompt AGRP expression	HGMD	15121772
16	72088461	rs5471	А	С	Ahaptoglobinaemia	luciferase reporter assay (HepG2)	decreases promoter activity to repress HP expression	HGMD	14616769

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

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	С	Hereditary hyperferritinaemia- cataract syndrome	EMSA (K562)	decreases binding affinity for iron-regulatory proteins	ClinVar, Genomiser, CDTS	10759702
20	23030292	rs16984852	С	А	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

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

<u>SearchInGroup:</u> 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 <u>https://github.com/mulinlab/VarNote</u>