COMMUNICATIONS

WGSA: an annotation pipeline for human genome sequencing studies

Xiaoming Liu[1,2], Simon White[3], Bo Peng[4], Andrew D. Johnson[5,6], Jennifer A. Brody[7], Alexander H. Li[1], Zhuoyi Huang[3], Andrew Carroll[8], Peng Wei[1,9], Richard Gibbs[3], Robert J. Klein[10], Eric Boerwinkle[1,2,3]

[1]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA; [2]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA; [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA; [4]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA; [5]NHLBI Framingham Heart Study, Bethesda, MD, USA; [6] Population Sciences Branch, NHLBI Division of Intramural Research, Bethesda, MD, USA; [7]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; [8]DNAnexus, Mountain View, CA, USA; [9]Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA; [10]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Icahn Institute for Genomics and Multiscale Biology, New York, NY, USA.
e-mail: Xiaoming.Liu@uth.tmc.edu

DNA sequencing technologies continue to make progress in both increased throughput and quality, and decreased cost. As we transition from whole exome capture sequencing to whole genome sequencing (WGS), our ability to convert machine-generated variant calls, including single nucleotide variant (SNV) and insertion-deletion variants (indels), into human-interpretable knowledge has lagged far behind the ability to obtain enormous amounts of variants. To help narrow this gap, here we present WGSA (WGS Annotator), an functional annotation pipeline for human genome sequencing studies, which is runnable out-of-box on the Amazon Compute Cloud and freely downloadable at https://sites.google.com/site/jpopgen/wgsa/.

Functional annotation is a key step in a WGS analysis. In one way, annotation helps the analyst filter to a subset of elements of particular interest (e.g., cell type specific enhancers), in another way annotation helps the investigators to increase the power of identifying phenotype-associated loci (e.g., association test using functional prediction score as a weight) and interpret potentially interesting findings. Currently, there are several popular gene model based annotation tools, including ANNOVAR [1], SnpEff [2] and the Ensembl Variant Effect Predictor (VEP) [3]. These can annotate a variety of gene models both protein coding and non-coding from a range of

species. It is well known among practitioners that different databases (e.g., RefSeq [4] and Ensembl [5]) use different models for the same gene. Even when the same gene structure is implemented, predicted consequences of a given variant from different annotation tools may not be in agreement [6]. Therefore, it has been suggested to obtain annotation from tools across multiple databases for a more complete interpretation of the variants discovered in WGS [6]. Annotations of both coding and non-coding variants include scores pertaining to functionality, conservation, population allele frequencies and disease-related annotations, i.e., known disease-causing variants and disease-associated variants identified in genome-wide association analyses (GWAS). Recent large-scale epigenomics projects provide rich datasets of cell-specific regulatory elements. Unfortunately, there are currently few tools available to integrate all those functional annotation resources and provide a convenient and efficient pipeline for annotating millions of variants discovered in a WGS study.

To facilitate the functional annotation step of WGS, we developed WGSA. Currently WGSA supports the annotation of SNVs and indels locally without remote database requests, allowing it to scale up for large WGS studies. The overview of the WGSA pipeline is presented in **figure 1**. The complete list of the resources (and their references) contained in WGSA can be found in **online supplementary table S1**. For gene-model based annotation, WGSA integrates the outputs from three annotation tools (ANNOVAR, SnpEff and VEP) versus two databases (RefSeq and Ensembl), and provides a summary of variant consequences from the six annotation results. To further speed up the process for large-scale WGS studies, we have pre-computed annotations for all potential human SNVs (a total of 8,584,031,106) based on human reference hg19 non-N bases and use it as a local database. For SNV-centric resources, WGSA integrates five functional prediction scores, eight conservation scores, allele frequencies from four large-scale sequencing studies, variants in four disease-related databases, among others (**figure 1** and **online supplementary table S1**). For regulatory region-centric resources, WGSA includes cell type specific transcription factor binding sites, DNAse I hypersensitivity regions and chromosome activity predictions from three epigenomics projects (**figure 1** and **online supplementary table S1**). WGSA also contains rich functional annotations for non-synonymous SNVs and genes from our dbNSFP database [7, 8].

Annotating the consequences of indels raises special challenges. In addition to the allele frequencies and consequences predicted by the three annotation tools (ANNOVAR, SnpEff and VEP), we take an approach to first "translate" an indel to local SNVs by incorporating their effect (insertion, deletion, replacement) on nearby flanking sequences, annotating those SNVs with other available resources, and then summarizing these results for the indel (**figure 1**, **online supplementary figure S1** and details in **online supplementary notes**). Although this approach is a simplification of potentially complicate impact of an indel, it will provide additional information on an indel regarding the focal region it resides, such as whether it is a STR (short tandem repeat) mutation, whether it breaks a TFBS (transcription factor binding site), local functional prediction scores, local conservation scores, etc.

To provide convenient access for a broad community, we have built an Amazon Machine Image (AMI) for running WGSA on the cloud via Amazon Web Services (AWS). To run WGSA, the user only needs to upload a variant list file (or a .vcf file) and a configuration file, containing a list of the resources to be included in the annotation. The annotation pipeline can be run with just two command line calls. We also provide WGSA as a downloadable version at https://sites.google.com/site/jpopgen/wgsa for bioinformaticians who prefer to build WGSA locally. It can be used as a foundation resource for customized annotation pipelines, as is the case for Baylor College of Medicine's Human Genome Sequencing Center annotation software, Cassandra (https://www.hgsc.bcm.edu/software/cassandra). The runtime of two experimental runs with 46.6 million variants is shown in **online supplementary table S2**. WGSA was written in Java so that most of its annotation modules (the whole SNV annotation and most of the indel annotation except VEP) can be easily run across different platforms. Detailed protocols for using WGSA in the cloud and building it locally can be found in **online supplementary notes** and at https://sites.google.com/site/jpopgen/wgsa/. As changes inevitably occur in the source databases as well as new annotation resources emerge, WGSA will be updated in response. Users will be able to receive update notice and detailed instruction on updating steps (for local version).

Acknowledgments We gratefully acknowledge contributions from the researchers of the CHARGE sequencing project, especially Drs. L Adrienne Cupples and Fuli Yu and other members of the CHARGE Analysis & Bioinformatics Committee. We thank Mike Dahdouli for providing technical support. We thank Jin Yu for advising on AWS usage.

Funding This study was supported by the US National Institutes of Health (5RC2HL102419 and U54HG003273).

Contributors X.L., A.D.J., J.A.B., A.H.L., A.C., P.W., Z.H., R.J.K. and E.B. designed the study. X.L. collected the annotation resources and developed the tool. S.W. tested the pipeline. B.P. provided tools for retrieving the RegulomeDB data set. E.B. and R.G. supervised the study. X.L., S.W. and E.B. wrote the draft manuscript and all authors provided critical edits.

Competing interests The authors declare that they have no competing interests.

Reference:

1 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

2 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;**6**:80–92.

3 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010;**26**:2069–70.

4 Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;**42**:D756–63.

5 Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SMJ, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. Ensembl 2015. *Nucleic Acids Res* 2015;**43**:D662–9.

6 McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, The WGS500 Consortium, Cazier J-B, Donnelly P. Choice of transcripts and software has a large effect on variant annotation. *Genome Med* 2014;**6**:26.

7 Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* 2011;**32**:894–9.

8 Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* 2013;**34**:E2393–402.

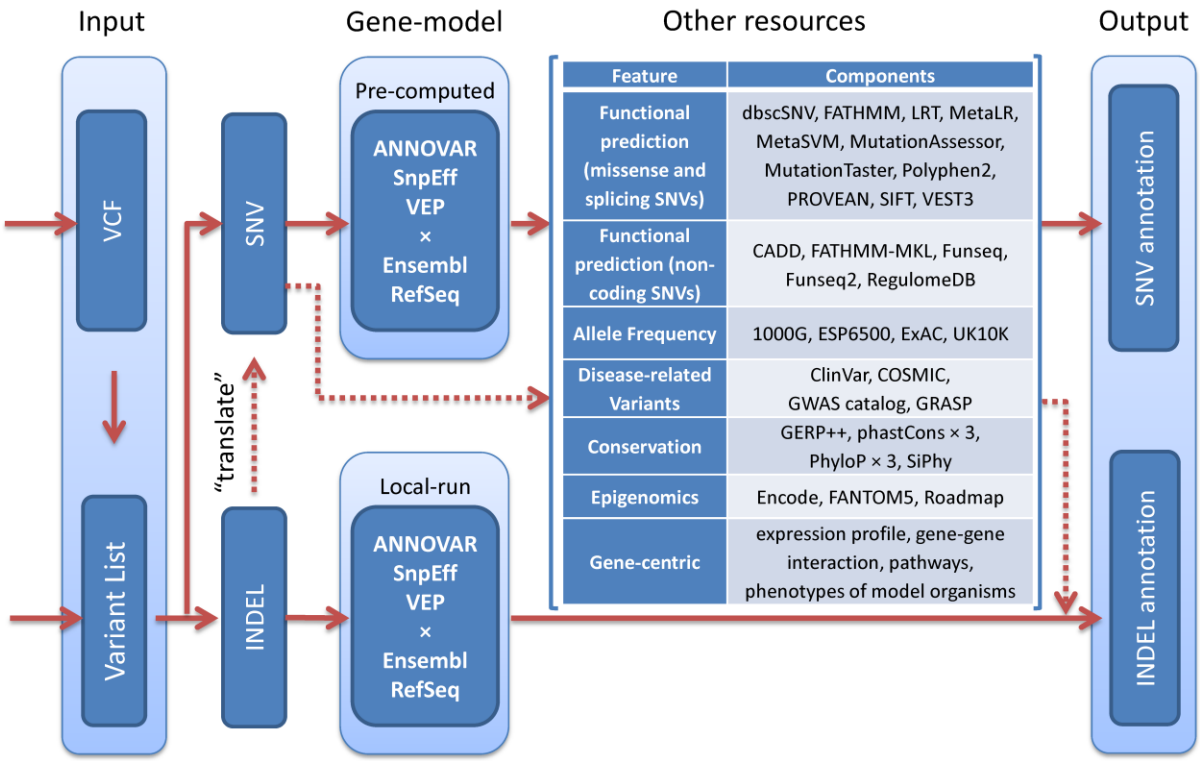


Figure 1: Flowchart of the WGSA annotation pipeline. Dotted lines show the "detour" of the indel annotation via the SNV annotation pipes.

Supplementary Table S1: Annotation resources integrated in WGSA (v0.5).

Resource	Brief Description	Ref.	URL
Functional annotation for missense and splicing SNVs & gene-centric annotation			
dbNSFP (v2.9)	An integrated functional annotation database for missense SNVs and splicing SNVs	[1,2]	https://sites.google.com/site/jpopgen/dbNSFP
dbscSNV (v1.0)	A deleteriousness prediction score for SNVs within splicing consensus regions (scSNVs)	[3]	https://sites.google.com/site/jpopgen/dbNSFP
Functional prediction scores for non-coding SNVs			
CADD (v1.2)	A genome-wide deleteriousness prediction score for DNA variants based on 63 sequence features (only SNV annotations are in WGSA)	[4]	http://cadd.gs.washington.edu/
FATHMM-MKL	A genome-wide deleteriousness prediction score for SNVs based on 10 feature groups	[5]	http://fathmm.biocompute.org.uk/fathmmMKL.htm
Funseq	A genome-wide categorical deleteriousness prediction score for DNA variants (only non-coding SNV annotations are in WGSA)	[6]	http://funseq.gersteinlab.org/
Funseq2	A genome-wide deleteriousness prediction score designed for non-coding somatic SNVs	[7]	http://funseq2.gersteinlab.org/
RegulomeDB (v1.0)	A genome-wide categorical functional prediction score for SNVs based on ENCODE annotation	[8]	http://regulomedb.org/

Allele frequencies (SNVs and indels)

1000G	Whole genome allele frequencies from the 1000 Genomes Project phase 3 data	9	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
ESP6500	Exome allele frequencies from the Exome Variant Server ESP6500SI-V2 release	10	http://evs.gs.washington.edu/EVS/
ExAC (r0.3)	Exome allele frequencies from the Exome Aggregation Consortium		http://exac.broadinstitute.org/
UK10K	Whole genome allele frequencies from TWINSUK cohort		http://www.uk10k.org/studies/cohorts.html

Disease-related variants (SNVs and indels)

ClinVar (2014/09/02)	DNA variants related to human diseases/phenotypes	11	http://www.ncbi.nlm.nih.gov/clinvar/
COSMIC (v71)	Somatic variants discovered in cancer	12	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
GWAS catalog (2015/03/05)	DNA variants associated with human diseases/phenotypes discovered in GWAS studies (downloaded 03/05/2015)	13	http://www.genome.gov/gwastudies/
GRASP 2.0	DNA variants associated with human diseases/phenotypes discovered in GWAS studies, including eQTLs and other quantitative trait scans	14	http://apps.nhlbi.nih.gov/Grasp/

Conservation scores

GERP++	A conservation score measured by "Rejected Substitutions"	[15]	http://mendel.stanford.edu/SidowLab/downloads/gerp/
phastCons46way primate	A conservation score based on 46way alignment primate set	[16]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates/
phastCons46way placental	A conservation score based on 46way alignment placental set	[16]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/
phastCons100way vertebrate	A conservation score based on 100way alignment vertebrate set	[16]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons/
phyloP46way primate	A conservation score based on 46way alignment primate set	[17]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates/
phyloP46way placental	A conservation score based on 46way alignment placental set	[17]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/
phyloP100way vertebrate	A conservation score based on 100way alignment vertebrate set	[17]	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way/
SiPhy	A conservation score based on 29 mammals genomes	[18]	http://www.broadinstitute.org/mammals/2x/siphy_hg19/

Epigenomics

ENCODE	DNase clusters (across 125 cell types), uniform TFBS	[19]	http://genome.ucsc.edu/ENCODE/downloads.html
FANTOM5	Predicted enhancers, CAGE peaks including TSS (promoter)	[20]	http://fantom.gsc.riken.jp/data/
Roadmap +ENCODE	Regulatory segmentations , TFBS binding probability	[21]	http://ngs.sanger.ac.uk/production/ensembl/regulation/hg19/

Ancestral information			
Ancestral allele	Ancestral allele inferred via 6 primates EPO + RSRS allele (for mitochondrial variants)	[22,23]	ftp://ftp.ebi.ac.uk/pub/databases/ensembl/ancestral_alleles/
AltaiNeandertal	Genotype of a deep sequenced Altai Neandertal	[24]	http://cdna.eva.mpg.de/neandertal/altai/AltaiNeandertal/VCF/
Denisova	Genotype of a deep sequenced Denisova	[25]	http://www.eva.mpg.de/denisova
Read mappability / genome accessibility			
MAP20	Average Duke mappability score based on 20bp read	[19]	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeMapability
MAP35	Average Duke mappability score based on 35bp read	[19]	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeMapability
1000G strict mask	Regions which are considered callable by the 1000 Genomes Project when analyzed with a stricter stringency (20120824_strict_mask)	[9]	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/
RepeatMasker	Regions masked by RepeatMasker		http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=rmsk
Other annotations			
dbSNP	rs number from dbSNP 142	[26]	http://www.ncbi.nlm.nih.gov/SNP/
snoRNA/miRNA	snoRNA and miRNA in human genome	[27,28]	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgRna

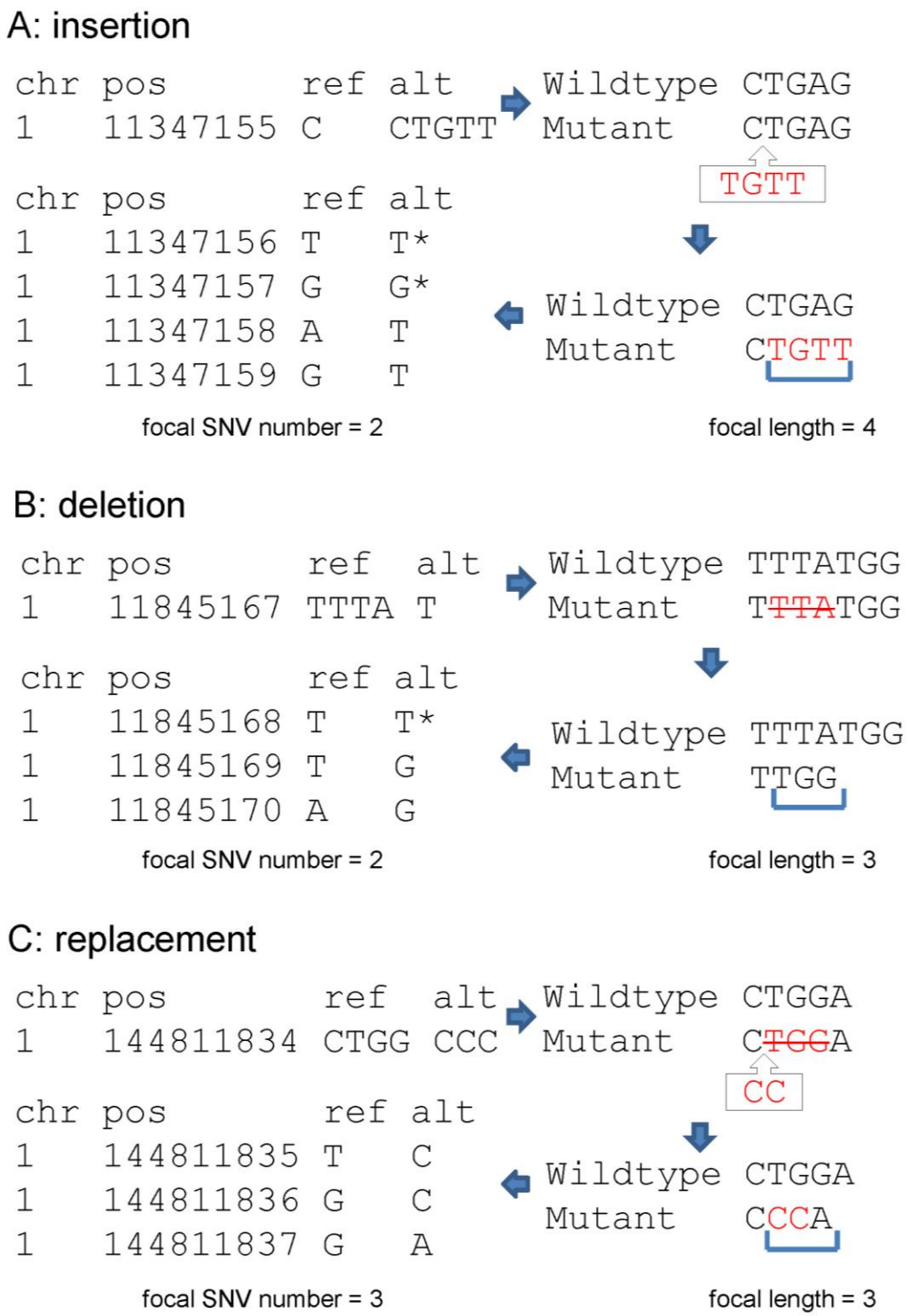
miRNA target	3'UTR miRNA target in human genome	29	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=targetScanS
ORegAnno	Known regulatory elements in human genome	30	http://www.oreganno.org/oregano/

Supplementary Table S2: Local and AWS runtime for WGSA (v0.5) on the UK10K cohort dataset. Variant file includes 42.4 million SNVs, 4.2 million indels (13.2 million focal SNVs and pseudo-SNVs).

	Local machine	**AWS EC2 instance**
Computing environment	AMD Opteron 6272 @ 2.1GHz, 8 threads, 60 GiB memory, HDD (7200 RPM 64MB Cache SATA 6.0Gb/s), SUSE Linux Release 11	r3.xlarge (Intel Xeon E5-2670 v2 @ 2.5GHz), 4 threads, 29 GiB memory, SSD (General Purpose EBS), Ubuntu 14.04.1 LTS
SNV annotations		
Integrated SNV annotation	5 h 48 min	5 h 15 min
dbSNP	12 min	18 min
snoRNA/miRNA	9 min	5 min
miRNA target	8 min	6 min
dbscSNV	15 min	12 min
GWAS catalog	9 min	5 min
GRASP 2.0	4 min	6 min
ClinVar	10 min	5 min
COSMIC	10 min	8 min
Duke mapability	59 min	34 min
1000G strict mask	22 min	14 min
RepeatMasker	19 min	11 min
EPO ancestral allele	10 min	12 min
Altai Neandertal geno.	54 min	39 min
Denisova geno.	52 min	41 min
PhyloP_primate	48 min	31 min
PhyloP_placental	45 min	32 min
PhyloP_vertebrate	50 min	32 min
PhastCons_primate	48 min	30 min
PhastCons_placental	46 min	28 min
PhastCons_vertebrate	46 min	28 min
GERP++	1 h	43 min
SiPhy	1 h 21 min	1 h 11 min
1000G allele freq.	1 h 14 min	42 min
UK10K cohort allele freq.	35 min	25 min
ESP6500 allele freq.	19 min	18 min
ExAC allele freq.	22 min	23 min
RegulomeDB	2 h 3 min	1 h 25 min
Funseq-like noncoding	2 h 55 min	1 h 42 min
Funseq2 noncoding	2 h 16 min	1 h 26 min
CADD	3 h 17 min	2 h 11 min
fathmm-MKL	4 h 2 min	2 h 20 min
OregAnno	13 min	23 min
ENCODE TFBS	25 min	32 min

ENCODE Dnase	17 min	28 min
Ensembl Regulatory Build Overviews	18 min	29 min
FAMTOM5 enhancer	14 min	26 min
FAMTOM5 CAGE	13 min	27 min
Ensembl Regulatory Build TFBS	16 min	28 min
dbNSFP	1 h 4 min	28 min
Ensembl Regulatory Build cell type activity	1 h 4 min	36 min
Ensembl Regulatory Build cell type specific segmentations	4 h 41 min	2 h 38 min
ENCODE Gm12878 segmentation	1 h 3 min	42 min
ENCODE H1hesc segmentation	1 h 4 min	42 min
ENCODE Helas3 segmentation	1 h 3 min	42 min
ENCODE Hepg2 segmentation	1 h 4 min	43 min
ENCODE Huvec segmentation	1 h 5 min	43 min
ENCODE K562 segmentation	1 h 5 min	42 min
Indel annotations		
ANNOVAR+SnpEff+VEP	7 h 2 min	6 h 12 min
dbSNP	1 min	1 min
GWAS catalog	< 1 min	< 1 min
GRASP	< 1 min	< 1 min
ClinVar	< 1 min	< 1 min
COSMIC	< 1 min	< 1 min
1000G allele freq.	4 min	2 min
UK10K cohort allele freq.	3 min	2 min
ESP6500 allele freq.	2 min	1 min
ExAC allele freq.	1 min	1 min
Indel annotations via focal SNV/pseudo-SNV annotations		
dbscSNV	5 min	3 min
snoRNA/miRNA	2 min	1 min
miRNA target	1 min	1 min
Duke mapability	42 min	21 min
1000G strict mask	11 min	6 min
RepeatMasker	8 min	3 min
PhyloP_primate	27 min	15 min
PhyloP_placental	28 min	14 min
PhyloP_vertebrate	30 min	14 min
PhastCons_primate	27 min	14 min
PhastCons_placental	25 min	13 min
PhastCons_vertebrate	26 min	13 min
GERP++	48 min	25 min
SiPhy	1 h 9 min	37 min
RegulomeDB	1 h 52 min	56 min
Funseq-like noncoding	2 h 34 min	1 h 14 min
Funseq2 noncoding	1 h 51 min	55 min
CADD	2 h 47 min	1 h 27 min

fathmm-MKL	3 h 22 min	1 h 42 min
OregAnno	2 min	1 min
ENCODE TFBS	8 min	4 min
ENCODE Dnase	4 min	1 min
Ensembl Regulatory Build Overviews	5 min	2 min
FAMTOM5 enhancer	2 min	1 min
FAMTOM5 CAGE	3 min	1 min
Ensembl Regulatory Build TFBS	3 min	2 min
Ensembl Regulatory Build cell type activity	13 min	6 min
Ensembl Regulatory Build cell type specific segmentations	2 h 40 min	1 h 18 min
ENCODE Gm12878 segmentation	12 min	5 min
ENCODE H1hesc segmentation	12 min	6 min
ENCODE Helas3 segmentation	11 min	4 min
ENCODE Hepg2 segmentation	12 min	5 min
ENCODE Huvec segmentation	12 min	5 min
ENCODE K562 segmentation	12 min	6 min
Summarizing focal SNV/pseudo-SNV annotations	35 min	21 min



Supplementary Figure S1: Examples illustrating the process of "translating" an indel to SNVs. A: insertion. B: deletion. C: replacement. *: a pseudo-SNV when the alternative allele is identical to the reference allele.

Supplementary Notes

1. Indel annotation via summarized SNV annotation

Step 1: Convert an indel to local SNVs. Similar to predicting the effect of an indel in the coding region of a protein (an example is given in Figure 1 of Zia and Moses (2011)[31]), we assume an indel may change the local DNA motif (if one or more exists) by introducing SNVs. Assuming there are no indels closely nearby (otherwise they shall be called as a single longer indel in the variant calling step) on the same haplotype, the largest impact of the indel shall be on the local region no larger than the length of the inserted or deleted allele, which we call the focal length of an indel.

Supplementary Figure S1A illustrates an example of an insertion. The insertion is defined by chromosome, position, reference allele and alternative allele, as defined in the vcf format[32]. Excluding the anchor nucleotide "C", the inserted allele is "TGTT", therefore the focal length of this insertion is 4. This inserted allele will correspond to "TGAG" on the reference sequence when anchoring left. Then within the focal length 4 we can define 4 SNVs: T>T, G>G, A>T and G>T. The first two SNVs have identical reference allele and alternative allele, which we call pseudo-SNVs. With a pseudo-SNV one cannot get an annotation from a SNV-centric annotation resource (e.g., CADD) but can get annotation from a position-centric annotation resource (e.g., GERP++). The remaining two SNVs are usual SNVs introduced by the insertion; therefore, the "focal SNV number" for this insertion is 2. Similarly, focal SNVs and pseudo-SNVs can be defined for deletions and replacements (**Supplementary Figure S1B,C**).

Please note we defined the corresponding positions of an indel allele to the reference sequence by anchoring the allele to the left. The reason we choose anchoring left is that many of the insertions (and deletions) are produced by inserting (or deleting) units of short tandem repeats. Typically such an insertion (or deletion) is presented as inserting (or deleting) the first (i.e. leftmost) such unit(s) in the reference sequence in vcf files. Therefore, by anchoring left we can easily identify those events as the focal SNV number will be 0.

Step 2: Annotating focal SNVs and pseudo-SNVs. The second step is to annotate the focal SNVs and pseudo-SNVs through the SNV annotation pipeline, except for the resources that can annotate indels directly (e.g., allele frequencies in human populations). As mentioned above, pseudo-SNVs will get "missing data" from SNV-centric annotations but can get annotations from position-centric annotations, such as conservation scores and regulatory segmentations.

Step 3: Summarizing annotations of focal SNVs and pseudo-SNVs. The last step is to summarize the annotations of focal SNVs and pseudo-SNVs for each indel. In the current implementation, we count the number of each unique annotation we got from the SNVs. For example, for the insertion shown in **Supplementary Figure S1A**, we obtain two CADD phred scores (2.074 and 5.290) for the two focal SNVs and "missing data" for the remaining two pseudo-SNV, and we summarize as ".{2}2.074{1}5.290{1}", where "." represents missing data.

2. Setting up WGSA on a local Linux machine

EQUIPMENT

▲CRITICAL As new features and optimizations will be developed and released for this pipeline in the future, we recommend using the most up-to-date version of this protocol, which will be hosted at https://sites.google.com/site/jpopgen/wgsa.

Computer hardware

- The WGSA pipeline is not computationally intensive but memory sensitive. The memory requirement depends on the number of variants to be annotated. For large whole genome sequencing studies with tens of millions of variants to be annotated, 32 GB or even larger memory may be required. Currently most of the annotating steps only need one thread to run but we highly recommend users to provide 4 or more threads for the VEP indel annotation step to speed up the process. By default, the pipeline will maximize the CPU and RAM resources available for the process. If the machine to be installed is not entirely devoted to this pipeline, users have an option to set the maximum memory and threads available to the pipeline (see **PROCEDURE** in **3. Protocol for using WGSA via Amazon Web Service (AWS)**).

Operating system

- We assume users have a Unix-like operating system with a bash shell on the machine for installing this pipeline. We have successfully installed it on machines running Ubuntu Linux and SUSE Linux Enterprise. It may be possible to install the pipeline following this protocol on MacOS X or Microsoft Windows with a Unix-like environment such as Cygwin (https://cygwin.com/) but additional steps may be required. Here we use Ubuntu Linux with bash shell as a model system. All commands below can be run in a terminal window.

Software

- If only SNV annotations are needed, Java 1.6 or higher is the only required software.
- To run ANNOVAR[33], SnpEff[34] and VEP[22] for indel annotations, Perl and Java 1.7 or higher is needed, as well as the main packages and gene models for ANNOVAR, SnpEff and VEP. This protocol shows the installation of ANNOVAR (2014Nov12), SnpEff v4.1 and VEP v78.

EQUIPMENT SETUP

Folders for the pipeline

- (Optional) We recommend to put all annotation resources within a folder dedicated for the pipeline, such as /WGSA
  - `mkdir /WGSA`
- (Optional) A tmp folder with writing permission is required, such as /WGSA/tmp
  - `mkdir /WGSA/tmp`

- `chmod 777 /WGSA/tmp`

Install ANNOVAR

- (Optional) The following steps are required if users want to annotate indels.
- Download the ANNOVAR main package from http://www.openbioinformatics.org/annovar/annovar_download.html. Please note a license is needed for commercial use of ANNOVAR.
- The package comes as annovar.latest.tar.gz, save it to /WGSA/annovar. Unzip it to /WGSA/annovar:
  - `mkdir /WGSA/annovar`
  - `cd /WGSA/annovar`
  - `tar -zxvf annovar.latest.tar.gz`
- Download RefSeq[35] and Ensembl[36] gene models for ANNOVAR:
  - `cd /WGSA/annovar/annovar`
  - `perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar refGene humandb/`
  - `perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar ensGene humandb/`

Install SnpEff

- (Optional) The following steps are required if users want to annotate indels.
- Download SnpEff v4.1 main package from http://snpeff.sourceforge.net/download.html and save the zip file to /WGSA/snpeff:
  - `mkdir /WGSA/snpeff`
  - `cd /WGSA/snpeff`
  - `wget` http://sourceforge.net/projects/snpeff/files/snpEff_latest_core.zip
  - `unzip snpEff_latest_core.zip`
- Download RefSeq[35] and Ensembl[36] gene models for SnpEff:
  - `cd /WGSA/snpeff/snpEff`
  - `java -jar snpEff.jar download -v hg19`
  - `java -jar snpEff.jar download -v GRCh37.75`

Install VEP

- (Optional) The following steps are required if users want to annotate indels.
- Download VEP 78 main package from http://useast.ensembl.org/info/docs/tools/vep/index.html and save it to /WGSA/vep:
  - `mkdir /WGSA/vep`
  - `cd /WGSA/vep`
  - `wget` https://github.com/Ensembl/ensembl-tools/archive/release/78.zip
  - `unzip 78.zip`

- Install some additional Perl modules that may be required for successfully installing VEP API, including Archive:Extract, Archive:Tar, Archive:Zip, CGI and DBI. Here are some example commands for Ubuntu Linux:
    - `sudo apt-get install libarchive-extract-perl`
    - `sudo apt-get install libarchive-zip-perl`
    - `sudo apt-get install libarchive-tar-perl`
    - `sudo apt-get install libcgi-pm-perl`
    - `sudo apt-get install libdbi-perl`
- Install VEP API to /WGSA/vep and download RefSeq and Ensembl gene models to /WGSA/.vep
    - `cd /WGSA/vep/ensembl-tools-release-78/scripts/variant_effect_predictor/`
    - `mkdir /WGSA/.vep`
    - `sudo perl INSTALL.pl -c /WGSA/.vep --ASSEMBLY GRCh37`
    - Go through the steps of the installing process and following the guidance at http://useast.ensembl.org/info/docs/tools/vep/script/vep_tutorial.html. When being asked for the cache files, choose "33 : homo_sapiens_merged_vep_78_GRCh37.tar.gz". When being asked for fasta files, choose "27 : homo_sapiens". The fasta file downloading is optional for the current version of WGSA.

Download the pipeline programs and other resources

- The main pipeline program is WGSA##.class where ## is the version number. We recommend putting the main program under /WGSA. The download links for the pipeline program and other resources are provided at https://sites.google.com/site/jpopgen/wgsa.
- Which resources need to be downloaded depend on which resources the users want to use for their annotation. The current available resources can be found in **Supplementary Table S1** and the up-to-date version is provided at https://sites.google.com/site/jpopgen/wgsa/list-of-resources. Two resource folders, *javaclass* and *hg19*, are necessary for running the pipeline and shall be downloaded. For integrated SNV annotations from ANNOVAR, SnpEff and VEP, users shall download *IntegratedSNV*.
- We recommend putting all downloaded resources under the folder /WGAS/resources.
- The resource files are either plain text file or compressed using lz4 algorithm implemented in Java (https://github.com/jpountz/lz4-java). We also provide tools for decompressing at https://sites.google.com/site/jpopgen/wgsa/lz4-utilities.

PROCEDURE

Please refer to the **PROCEDURE** subsection of the next section for the usage of the WGSA pipeline.

3. Using WGSA via Amazon Web Service (AWS)

EQUIPMENT

Computing environment

- A computer with internet connection.
- A secure shell (SSH) client installed in the operating system (e.g. PuTTY).
- A SCP or SFTP client installed in the operating system (e.g. FileZilla).

EQUIPMENT SETUP

Create an AWS account

- If you already have an AWS account, skip this step.
- Following the steps at http://aws.amazon.com/ to create an account.
- A graphical guidance of this step can be found at https://sites.google.com/site/jpopgen/wgsa/create-an-aws-account.

Launch an instance from an AMI of WGSA

- Sign in to your AWS account. Navigate to the EC2 Dashboard.
- Search for a WGSA public AMI and choose to launch an instance from it. A list of available WGSA AMI can be found at https://sites.google.com/site/jpopgen/wgsa.
- Configure the type and details of the instance and launch the instance.
- A graphical guidance of this step can be found at https://sites.google.com/site/jpopgen/wgsa/launch-an-instance.

Terminate an instance (after the PROCEDURE is finished)

- Navigate to the EC2 Dashboard.
- Select the instance and choose “Terminate” from “Instance State”.
- Find the EBS volume that is leftover from the instance and choose “Delete Volume”.
- A graphic guidance of this step can be found at https://sites.google.com/site/jpopgen/wgsa/terminate-an-instance.

PROCEDURE

1| Prepare input files ▲CRITICAL

- Two input files are needed. One is a variant file and the other is a configuration/setting file.

- The standard variant file is a plain text format file with TAB-delimited columns. The first row must be a title row and then followed by variant rows, with one variant per row. The first four columns must be chromosome, position, reference allele and alternative allele, with their formats defined by the vcf file format[32]. Multiple alternative alleles for the same reference allele shall be separated into multiple rows. Additional columns can be included. An example is shown in **Figure: *Standard variant file example***.
- Alternatively a plain text vcf format file (must with an extension .vcf) can be used as a variant file. The pipeline will automatically recognize its format and converted to a standard variant file.
- A configuration file is a plain text format file, in which the users provide information for the name of input file, name of the output file, directory to various resources and options for annotation. The first 30 lines of a configuration file are given in **Figure: *Configuration file example***. A complete template file is also given at https://sites.google.com/site/jpopgen/wgsa/using-wgsa-via-aws. The file is self-explanatory and users shall change the contents of the second column to specify their options. The available options are given in the third column (beginning with #). (**! CAUTION** To run the pipeline on a local machine, the directories settings (line 3 to 8) shall be modified to reflect the absolute paths to the corresponding directories on the local machine.)
- Upload the variant file and the configuration file to the folder where WGSA##.class resides, e.g. /WGSA.

#chr	pos	ref	alt	clinvar_rs	CLNSIG
1	955597	G	T	rs115173026	3
2	179510636	G	GCATT	rs397517560	3
3	37081782	T	TAAGT	rs267607699	5
5	172660144	CCCG	CAT	rs397516908	4
22	41320486	G	T	rs267607179	5
X	13764534	GAGAAAT	G	rs312262839	5
Y	545379	G	T	rs137852558	5
MT	3700	G	A	rs397515508	5

Figure: *Standard variant file example.*

2| Upload input files

- SSH to the machine with WGSA installed (e.g. a WGSA AMI instance).
- Upload the input files with SCP or SFTP.
- A graphical guidance for accessing a WGSA instance using PuTTY and FileZilla as examples can be found at https://sites.google.com/site/jpopgen/wgsa/ssh-and-sftp-to-an-instance. Guidance for Linux environment or across platforms using MindTerm can be

found at
http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstances.html.

3| Create the pipeline shell script

- Within the terminal, change directory to the WGSA folder where the WGSA##.class resides (i.e. /WGSA for a WGSA AMI instance)
    - `cd /WGSA`
- Run the WGSA main program (e.g. WGSA05.class) following the configuration file name, for example, my_configuration_file, which will create a shell script with the name my_configuration_file.sh and two text files with descriptions for the columns of the final SNV and indel annotation files.
    - `java WGSA05 my_configuration_file`
- (Optional) If you want to limit the maximum memory and threads available to WGSA, you can add additional parameters specifying the maximum memory (in GiB) and number of threads. For example, here is an example setting a maximum of 30 GiB memory and 4 threads to the pipeline:
    - `java WGSA05 my_configuration_file 30 4`

```
input file name:              test.vcf                                        #name of the input file
output file name:             test.annotated                                  #name of the output file
resources dir:                /WGSA/resources/
annovar dir:                  /WGSA/annovar/annovar/
snpeff dir:                   /WGSA/snpeff/snpEff/
vep dir:                      /WGSA/vep/ensembl-tools-release-78/scripts/variant_effect_predictor/
.vep dir:                     /WGSA/.vep/
tmp dir:                      /WGSA/tmp/
retain intermediate file:          no     #supported option: snp or s, indel or i, both or b, no or n
dbSNP:                             both   #supported option: snp or s, indel or i, both or b, no or n
snoRNA miRNA:                      no     #supported option: snp or s, indel or i, both or b, no or n
UTR3 miRNA target:                 no     #supported option: snp or s, indel or i, both or b, no or n
scSNV deleteriousness prediction: snp     #supported option: snp or s, indel or i, both or b, no or n
GWAS catalog:                      both   #supported option: snp or s, indel or i, both or b, no or n
GRASP:                             no     #supported option: snp or s, indel or i, both or b, no or n
Clinvar:                           both   #supported option: snp or s, indel or i, both or b, no or n
COSMIC                             no     #supported option: snp or s, indel or i, both or b, no or n
Duke mapability:                   both   #supported option: snp or s, indel or i, both or b, no or n
1000G mask:                        both   #supported option: snp or s, indel or i, both or b, no or n
RepeatMasker:                      both   #supported option: snp or s, indel or i, both or b, no or n
EPO ancestral:                     snp    #supported option: snp or s, no or n
AltaiNeandertal genotypes:         no     #supported option: snp or s, no or n
Denisova genotypes:                no     #supported option: snp or s, no or n
PhyloP_primate:                    no     #supported option: snp or s, indel or i, both or b, no or n
PhyloP_placental:                  no     #supported option: snp or s, indel or i, both or b, no or n
PhyloP_vertebrate:                 no     #supported option: snp or s, indel or i, both or b, no or n
PhastCons_primates:                no     #supported option: snp or s, indel or i, both or b, no or n
PhastCons_placental:               no     #supported option: snp or s, indel or i, both or b, no or n
PhastCons_vertebrate:              no     #supported option: snp or s, indel or i, both or b, no or n
GERP++:                            both   #supported option: snp or s, indel or i, both or b, no or n
```

Figure: ***Configuration file example.*** Only the first 30 lines are shown.

4| Run the pipeline shell script

- Run the shell script created in **3|**, for example, my_configuration_file.sh
  - `bash my_configuration_file.sh`
- (Optional) We recommend saving the standard output and error messages to files for record purpose.
  - `bash my_configuration_file.sh >output.txt 2>error.txt`
- The time needed for finishing all annotation steps may take hours even days, mostly depending on the total number of variants to be annotated and the annotation steps users specified in the configuration file. The time spent for annotating the UK10K cohort dataset (http://www.uk10k.org/data.html; 3,781 individuals, whole genome sequencing with 6x coverage, 4.2 million indels, 42.4 million SNVs) with two experimental runs are shown in **Supplementary Table S2**.

5| Download output files

- SNV and indel annotation files will be outputted separately, as well as the descriptions of their columns (see also **5. Column description for SNV annotation files** and **6. Column description for indel annotation files**).

- If indel annotations were chosen in the configuration file, the transcript-specific indel annotation file for RefSeq and Ensembl (.IntegratedRefseq file and .IntegratedEnsembl file) will be outputted.
- If intermediate files were chosen to be retained, intermediate files at each step will be retained. They can be useful if the annotation pipeline is interrupted (e.g. when using spot priced instances).
- For record purposes the actual SNV and indel variant files as well as the list of focal SNVs and pseudo-SNVs used for annotation will be retained. The SNV file can be used for searching transcript-specific SNV annotations (see **4. Integrated transcript-specific SNV annotation**).
- Download the output files using SFTP or SCP (see step **2**|).

4. Integrated transcript-specific SNV annotation

WGSA also provides integrated transcript-specific SNV annotation from ANNOVAR, SnpEff and VEP × Ensembl and RefSeq. The resources are provided as both an AWS AMI and a downloadable version.

Use downloadable version

- The link to the downloadable version can be found at https://sites.google.com/site/jpopgen/wgsa.
- The setup is similar as described in **2. Setting up WGSA on a local Linux machine**.
- Download the directory "resources2" and its contents (keeping its structure) to a folder, such as "/WGSA". Download the Java class "search_integrated_output5.class" to the same folder.
- Upload your standard variant input file, such as the ".snp" file outputted by WGSA during the **PROCEDURE** in **3. Using WGSA via Amazon Web Service (AWS)**. If it is compressed, decompress it.
- Run the Java class as "java search_integrated_output5 [input_file] [searchEnsembl(true or false)] [searchRefseq(true or false)] <sourcedir>". The first argument is the name of SNV standard variant file. The second argument is either true or false for searching the integrated annotation with Ensembl. The third argument is either true or false for searching the integrated annotation with Refseq. The forth argument is optional: if the path to the resources2 directory is "/WGSA/resources2", it can be omitted; otherwise, provide the path to the resources2 directory. For large variant files, specify a large enough memory for the usage of Java by using "-Xmx". The following is an example:
    - `java -Xmx29g input.snp true true ~/WGSA/resources2`
- If "searchEnsembl" is set "true", an ".IntegratedEnsembl" file will be outputted. The description of the columns of this file can be found at **7. Column description of Ensembl transcript-specific SNV annotation**.

- If “searchRefseq” is set “true”, an “.IntegratedRefseq” file will be outputted. The description of the columns of this file can be found at **8. Column description of RefSeq transcript-specific SNV annotation**.
- (Optional) Compress the output file (with gzip).
- Download the output files.

Use AWS AMI

- The setup is similar as described in **3. Using WGSA via Amazon Web Service (AWS)**
- Launch an instance from an AMI of WGSA resources2. A list of available AMI can be found at https://sites.google.com/site/jpopgen/wgsa.
- Upload your standard variant input file, such as the “.snp” file outputted by WGSA during the PROCEDURE in **3. Using WGSA via Amazon Web Service (AWS)**. We recommend compressing the input file before uploading and decompressing it after uploading (with gzip).
- Run the Java class as “java search_integrated_output5 [input_file] [searchEnsembl(true or false)] [searchRefseq(true or false)]”. The first argument is the name of SNV standard variant file. The second argument is either true or false for searching the integrated annotation with Ensembl. The third argument is either true or false for searching the integrated annotation with Refseq. For large variant files, specify a large enough memory for the usage of Java by using “-Xmx”. The following is an example:
    - `java -Xmx29g input.snp true true`
- If “searchEnsembl” is set “true”, an “.IntegratedEnsembl” file will be outputted. The description of the columns of this file can be found at **7. Column description of Ensembl transcript-specific SNV annotation**.
- If “searchRefseq” is set “true”, an “.IntegratedRefseq” file will be outputted. The description of the columns of this file can be found at **8. Column description of RefSeq transcript-specific SNV annotation**.
- Compress the output file (with gzip).
- Download the output files.

5. Column description for SNV annotation files (WGSA v0.5)

#chr: chromosome number

pos: position (hg19)

ref: reference allele

alt: alternative allele

ANNOVAR_ensembl_summary: ANNOVAR consequence summary with Ensembl as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_ensembl_summary: SnpEff consequence summary with Ensembl as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

VEP_ensembl_summary: VEP consequence summary with Ensembl as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

ANNOVAR_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

VEP_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_ensembl_LOF: SnpEff Loss-Of-Function summary with Ensembl as gene model.

Format: GeneID(total number of transcripts):

consequence#1(percentage of transcripts affected*total number of coding transcripts)

consequence#2(percentage of transcripts affected*total number of coding transcripts)...

SnpEff_refseq_LOF: SnpEff Loss-Of-Function summary with Refseq as gene model.

Format: GeneID(total number of transcripts):

consequence#1(percentage of transcripts affected*total number of coding transcripts)

consequence#2(percentage of transcripts affected*total number of coding transcripts)...

rs_dbSNP141: rs number from dbSNP141

sno_miRNA_name: the name of snoRNA or miRNA if the site is located within (from UCSC)

sno_miRNA_type: the type of snoRNA or miRNA (from UCSC)

UTR3_miRNA_target: the gene-miRNA pair, if the site is located within a 3'UTR miRNA target (from UCSC)

splicing_consensus_adaboost_score: splicing-change prediction for splicing consensus SNPs

based on adaboost. If the score >0.6, it predicts that the splicing will be changed,

otherwise it predicts the splicing will not be changed.

splicing_consensus_rf_score: splicing-change prediction for splicing consensus SNPs

based on random forest. If the score >0.6, it predicts that the splicing will be changed,

otherwise it predicts the splicing will not be changed

GWAS_catalog_rs: rs number according to GWAS catalog

GWAS_catalog_trait: associated trait according to GWAS catalog

GWAS_catalog_pubmedid: pubmedid of the paper describing the association

GRASP_rs: rs number by GRASP

GRASP_PMID: PMID number by GRASP

GRASP_p-value: p-value of the association test based on the SNP

GRASP_phenotype: phenotype the SNP associated with

GRASP_phenotype: phenotype the SNP associated with

GRASP_ancestry: population ancestry of the samples on which the association test was based

GRASP_platform: SNP platform on which the association test was based

clinvar_rs: rs number by clinvar

clinvar_clnsig: clinical significance by clinvar

2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response,

7 - histocompatibility. A negative score means the the score is for the ref allele

clinvar_trait: the trait/disease the clinvar_clnsig referring to

COSMIC_ID: ID of the SNV at the COSMIC (Catalogue Of Somatic Mutations In Cancer) database

COSMIC_CNT: number of samples having this SNV in the COSMIC database

MAP20: average Duke mappability score based on 20bp read,

0-1, higher score means higher mappability

MAP35: average Duke mappability score based on 35bp read

0-1, higher score means higher mappability

1000G_strict_masked: whether the site is within the 1000G strict masked region

Y (Yes) or N (No), Y means generally good mapping quality

RepeatMasker_masked: whether the site is masked by RepeatMasker

Y (Yes) or N (No), Y means generally lower mapping quality

Ancestral_allele: Ancestral allele (based on the EPO pipeline). The following

comes from its original README file:

ACTG - high-confidence call, ancestral state supported by the other two sequences

actg - low-confidence call, ancestral state supported by one sequence only

N - failure, the ancestral state is not supported by any other sequence

- - the extant species contains an insertion at this position

. - no coverage in the alignment

AltaiNeandertal: genotype of a deep sequenced Altai Neandertal

Denisova: genotype of a deep sequenced Denisova

phyloP46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative

phyloP46way_placental: a conservation score based on 46way alignment placental set, the higher the more conservative

phyloP100way_vertebrate: a conservation score based on 100way alignment vertebrate set, the higher the more conservative

phastCons46way_primate: a conservation score based on 46way alignment primate set, the higher the more conservative

phastCons46way_placental: a conservation score based on 46way alignment placental set, the higher the more conservative

phastCons100way_vertebrate: a conservation score based on 100way alignment vertebrate set, the higher the more conservative

GERP++_NR: GERP++ neutral rate

GERP++_RS: GERP++ RS score, the larger the score, the more conserved the site

SiPhy_29way_logOdds: SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site

1000Gp3_AC: Alternative allele counts in the whole 1000 genomes phase 3 (1000Gp3) data.

1000Gp3_AF: Alternative allele frequency in the whole 1000Gp3 data.

1000Gp3_AFR_AC: Alternative allele counts in the 1000Gp3 African descendent samples.

1000Gp3_AFR_AF: Alternative allele frequency in the 1000Gp3 African descendent samples.

1000Gp3_EUR_AC: Alternative allele counts in the 1000Gp3 European descendent samples.

1000Gp3_EUR_AF: Alternative allele frequency in the 1000Gp3 European descendent samples.

1000Gp3_AMR_AC: Alternative allele counts in the 1000Gp3 American descendent samples.

1000Gp3_AMR_AF: Alternative allele frequency in the 1000Gp3 American descendent samples.

1000Gp3_EAS_AC: Alternative allele counts in the 1000Gp3 East Asian descendent samples.

1000Gp3_EAS_AF: Alternative allele frequency in the 1000Gp3 East Asian descendent samples.

1000Gp3_SAS_AC: Alternative allele counts in the 1000Gp3 South Asian descendent samples.

1000Gp3_SAS_AF: Alternative allele frequency in the 1000Gp3 South Asian descendent samples.

TWINSUK_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

TWINSUK_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

ALSPAC_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

ALSPAC_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

ESP6500_AA_AC: Alternative allele counts in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_AA_AF: Alternative allele frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_EA_AC: Alternative allele counts in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_EA_AF: Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ExAC_AC: Allele count in total ExAC samples (~60,706 unrelated individuals)

ExAC_AF: Allele frequency in total ExAC samples

ExAC_Adj_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC samples

ExAC_Adj_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC samples

ExAC_AFR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC samples

ExAC_AFR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC samples

ExAC_AMR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC samples

ExAC_AMR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC samples

ExAC_EAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC samples

ExAC_EAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC samples

ExAC_FIN_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC samples

ExAC_FIN_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC samples

ExAC_NFE_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

ExAC_NFE_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

ExAC_SAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC samples

ExAC_SAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC samples

RegulomeDB_motif: motif the SNP resides (from RegulomeDB)

RegulomeDB_score: categorical score from RegulomeDB. The smaller, the more likely the SNP affects binding

Motif_breaking: whether break a known motif (in-house script)

network_hub: whether the target gene is a network hub based on funseq-0.1

ENCODE_annotated: whether annotated by ENCODE based on funseq-0.1

sensitive: whether defined as sensitive region based on funseq-0.1

ultra_sensitive: whether defined as ultra-sensitive region based funseq-0.1

target_gene: target gene (for promoter, enhancer, etc.) based on funseq-0.1

funseq_noncoding_score: funseq-like noncoding score range 0-6, each of the previous 5 columns contribute 1 if "YES", or 0 if "NO"; the column Motif_breaking contribute 1 if it is not a "."

funseq2_noncoding_score: funseq2 noncoding score range 0-5.4 a weighted score designed for damaging prediction of cancer somatic SNPs

CADD_raw: CADD raw score, the larger the number the more likely damaging

CADD_phred: CADD phred-like score, ranges 1-99, the larger the number the more likely damaging; score >10 means the variant in the top 10% (0.1) among the total 8.6 billion possible SNVs, >20 means in the top 1%, >30 means in the top 0.1%, etc. CADD suggests a cutoff between 10 and 20 (e.g. 15)

fathmm-MKL_non-coding_score: fathmm-MKL non-coding prediction probability, the larger the number the more likely damaging; the threshold separating deleterious prediction and neutral prediction is 0.5.

fathmm-MKL_non-coding_group: fathmm-MKL non-coding group, the feature group used for the non-coding prediction

fathmm-MKL_coding_score; fathmm-MKL coding prediction probability, the larger the number the more likely damaging the threshold separating deleterious prediction and neutral prediction is 0.5.

fathmm-MKL_coding_group: fathmm-MKL coding group, the feature group used for the coding prediction.

ORegAnno_type: the type of regulatory region by ORegAnno

ORegAnno_PMID: the PMID of the paper describing the regulation

ENCODE_TFBS: name of the transcription factors (separated by ;) if the site is within a TFBS

ENCODE_TFBS_score: the higher the score the stronger the evidence of the TFBS

ENCODE_TFBS_cells: the cell lines (separated by ;) the TFBS was detected

ENCODE_Dnase_score: the higher the score the stronger the evidence of a DNase I hypersensitive site

ENCODE_Dnase_cells: number of cell lines supporting a DNase I hypersensitive site

Ensembl_Regulatory_Build_Overviews: genome segment prediction based on 17 cell types from ENCODE and Roadmap. Predicted states: ctcf - CTCF binding sites, distal - Predicted enhancers open - Unannotated open chromatin regions, proximal - Predicted promoter flanking regions, tfbs - Unannotated transcription factor binding sites, tss - Predicted promoters

FAMTOM5_enhancer: whether the site is within a FAMTOM5 predicted enhancer region Y (Yes) or N (No)

FAMTOM5_CAGE_peak: whether the site is within a FAMTOM5 Cap Analysis of Gene Expression (CAGE) peak. Y (Yes) or N (No). A CAGE peak generally suggests a promoter region

Ensembl_Regulatory_Build_TFBS: TFBS from Ensembl Regulatory Build. Multiple TFBS separated by ";"

Ensembl_Regulatory_Build_TFBS_prob: the probability of observing TFBS binding Multiple probabilities (corresponding to Ensembl_Regulatory_Build_TFBS) separated by ";"

The following columns are unique to SNPs with an entry in dbNSFP v2.9 (nonsynonymous or splicing):

aaref: reference amino acid "-" if the variant is a splicing site SNP (2bp on each end of an intron)

aaalt: alternative amino acid "-" if the variant is a splicing site SNP (2bp on each end of an intron)

hg18_pos(1-based): physical position on the chromosome as to hg18 (1-based coordinate)

genename: gene name; if the NScan be assigned to multiple genes, gene names are separated by ";"

Uniprot_acc: Uniprot accession number. Multiple entries separated by ";".

Uniprot_id: Uniprot ID number. Multiple entries separated by ";".

Uniprot_aapos: amino acid position as to Uniprot. Multiple entries separated by ";".

Interpro_domain: domain or conserved site on which the variant locates. Domain annotations come from Interpro database. The number in the brackets following a specific domain is the count of times Interpro assigns the variant position to that domain, typically coming from different predicting databases. Multiple entries separated by ";".

cds_strand: coding sequence (CDS) strand (+ or -)

refcodon: reference codon

SLR_test_statistic: SLR test statistic for testing natural selection on codons. A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence.

codonpos: position on the codon (1, 2 or 3)

fold-degenerate: degenerate type (0, 2 or 3)

Ensembl_geneid: Ensembl gene id

Ensembl_transcriptid: Ensembl transcript ids (separated by ";")

aapos: amino acid position as to the protein "-1" if the variant is a splicing site SNP (2bp on each end of an intron)

aapos_SIFT: ENSP id and amino acid positions corresponding to SIFT scores. Multiple entries separated by ";"

aapos_FATHMM: ENSP id and amino acid positions corresponding to FATHMM scores.
Multiple entries separated by ";"

SIFT_score: SIFT score (SIFTori). Scores range from 0 to 1. The smaller the score the more likely the SNP has damaging effect. Multiple scores separated by ";".

SIFT_converted_rankscore: SIFTori scores were first converted to SIFTnew=1-SIFTori, then ranked among all SIFTnew scores in dbNSFP. The rankscore is the ratio of the rank the SIFTnew score over the total number of SIFTnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The rankscores range from 0.02654 to 0.87932.

SIFT_pred: If SIFTori is smaller than 0.05 (rankscore>0.55) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". Multiple predictions separated by ";"

Polyphen2_HDIV_score: Polyphen2 score based on HumDiv, i.e. hdiv_prob. The score ranges from 0 to 1. Multiple entries separated by ";".

Polyphen2_HDIV_rankscore: Polyphen2 HDIV scores were first ranked among all HDIV scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.02656 to 0.89917.

Polyphen2_HDIV_pred: Polyphen2 prediction based on HumDiv, "D" ("porobably damaging", HDIV score in [0.957,1] or rankscore in [0.52996,0.89917]), "P" ("possibly damaging", HDIV score in [0.453,0.956] or rankscore in [0.34412,0.52842]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.02656,0.34399]). Score cutoff for binary classification is 0.5 for HDIV score or 0.35411 for rankscore, i.e. the prediction is

"neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.35411), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.35411). Multiple entries are separated by ";".

Polyphen2_HVAR_score: Polyphen2 score based on HumVar, i.e. hvar_prob. The score ranges from 0 to 1. Multiple entries separated by ";".

Polyphen2_HVAR_rankscore: Polyphen2 HVAR scores were first ranked among all HVAR scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0.01281 to 0.9711.

Polyphen2_HVAR_pred: Polyphen2 prediction based on HumVar, "D" ("porobably damaging", HVAR score in [0.909,1] or rankscore in [0.62955,0.9711]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.44359,0.62885]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01281,0.44315]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45998 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.45998), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.45998). Multiple entries are separated by ";".

LRT_score: The original LRT two-sided p-value (LRTori), ranges from 0 to 1.

LRT_converted_rankscore: LRTori scores were first converted as LRTnew=1-LRTori*0.5 if Omega<1, or LRTnew=LRTori*0.5 if Omega>=1. Then LRTnew scores were ranked among all LRTnew scores in dbNSFP. The rankscore is the ratio of the rank over the total number of the scores in dbNSFP. The scores range from 0.00166 to 0.85682.

LRT_pred: LRT prediction, D(eleterious), N(eutral) or U(nknown), which is not solely

determined by the score.

MutationTaster_score: MutationTaster p-value (MTori), ranges from 0 to 1.

MutationTaster_converted_rankscore: The MTori scores were first converted: if the prediction is "A" or "D" MTnew=MTori; if the prediction is "N" or "P", MTnew=1-MTori. Then MTnew scores were ranked among all MTnew scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.0931 to 0.80722.

MutationTaster_pred: MutationTaster prediction, "A" ("disease_causing_automatic"), "D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic"). The score cutoff between "D" and "N" is 0.5 for MTori and 0.328 for the rankscore.

MutationAssessor_score: MutationAssessor functional impact combined score (MAori). The score ranges from -5.545 to 5.975 in dbNSFP. Please refer to Reva et al. (2011) Nucl. Acids Res. 39(17):e118 for details.

MutationAssessor_rankscore: MAori scores were ranked among all MAori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MAori scores in dbNSFP. The scores range from 0 to 1.

MutationAssessor_pred: MutationAssessor's functional impact of a variant : predicted functional, i.e. high ("H") or medium ("M"), or predicted non-functional, i.e. low ("L") or neutral ("N"). The MAori score cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 3.5, 1.9 and 0.8, respectively. The rankscore cutoffs between "H" and "M", "M" and "L", and "L" and "N", are 0.9416, 0.61387 and 0.26162, respectively.

FATHMM_score: FATHMM default score (weighted for human inherited-disease mutations with

Disease Ontology) (FATHMMori). Scores range from -18.09 to 11.0. Multiple scores separated by ";" Please refer to Shihab et al. (2013) Human Mutation 34(1):57-65 for details.

FATHMM_rankscore: FATHMMori scores were ranked among all FATHMMori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMori scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented. The scores range from 0 to 1.

FATHMM_pred: If a FATHMMori score is <=-1.5 (or rankscore <=0.81415) the corresponding NS is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)". Multiple predictions separated by ";"

MetaSVM_score: Our support vector machine (SVM) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP.

MetaSVM_rankscore: MetaSVM scores were ranked among all MetaSVM scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaSVM scores in dbNSFP. The scores range from 0 to 1.

MetaSVM_pred: Prediction of our SVM based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0. The rankscore cutoff between "D" and "T" is 0.83357.

MetaLR_score: Our logistic regression (LR) based ensemble prediction score, which incorporated 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster,

Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from 0 to 1.

MetaLR_rankscore: MetaLR scores were ranked among all MetaLR scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MetaLR scores in dbNSFP. The scores range from 0 to 1.

MetaLR_pred: Prediction of our MetaLR based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.82268.

Reliability_index: Number of observed component scores (except the maximum frequency in the 1000 genomes populations) for MetaSVM and MetaLR. Ranges from 1 to 10. As MetaSVM and MetaLR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions.

VEST3_score: VEST 3.0 score. Score ranges from 0 to 1. The larger the score the more likely the mutation may cause functional change. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. Please refer to Carter et al., (2013) BMC Genomics. 14(3) 1-16 for details.
Please note this score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

VEST3_rankscore: VEST3 scores were ranked among all VEST3 scores in dbNSFP.
The rankscore is the ratio of the rank of the score over the total number of VEST3 scores in dbNSFP. The scores range from 0 to 1.

Please note VEST score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.

PROVEAN_score: PROVEAN score (PROVEANori). Scores range from -14 to 14. The smaller the score the more likely the SNP has damaging effect. Multiple scores separated by ";". Details can be found in DOI: 10.1371/journal.pone.0046688

PROVEAN_converted_rankscore: PROVEANori were first converted to PROVEANnew=1-(PROVEANori+14)/28, then ranked among all PROVEANnew scores in dbNSFP. The rankscore is the ratio of the rank the PROVEANnew score over the total number of PROVEANnew scores in dbNSFP. If there are multiple scores, only the most damaging (largest) rankscore is presented.

PROVEAN_pred: If PROVEANori <= -2.5 (rankscore>=0.59) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "N(eutral)". Multiple predictions separated by ";"

The following columns are cell type specific:

Ensembl_A549_activity: A549 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_DND41_activity: DND41 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_GM12878_activity: GM12878 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_H1HESC_activity: H1HESC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HELAS3_activity: HELAS3 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HEPG2_activity: HEPG2 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HMEC_activity: HMEC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMM_activity: HSMM specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMMT_activity: HSMMT specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HUVEC_activity: HUVEC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_K562_activity: K562 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_MONO_activity: MONO specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHA_activity: NHA specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHDFAD_activity: NHDFAD specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHEK_activity: NHEK specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHLF_activity: NHLF specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_OSTEO_activity: OSTEO specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_A549_segmentation: A549 specific genome segment prediction from Ensembl Regulatory

Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_DND41_segmentation: DND41 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_GM12878_segmentation: GM12878 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_H1HESC_segmentation: H1HESC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HELAS3_segmentation: HELAS3 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HEPG2_segmentation: HEPG2 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HMEC_segmentation: HMEC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HSMM_segmentation: HSMM specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HSMMT_segmentation: HSMMT specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HUVEC_segmentation: HUVEC specific genome segment prediction from Ensembl Regulatory

Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_K562_segmentation: K562 specific genome segment prediction from Ensembl Regulatory

Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_MONO_segmentation: MONO specific genome segment prediction from Ensembl Regulatory

Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHA_segmentation: NHA specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHDFAD_segmentation: NHDFAD specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHEK_segmentation: NHEK specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHLF_segmentation: NHLF specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_OSTEO_segmentation: OSTEO specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

ENCODE_Gm12878_segmentation: the genome segmentation of the cell line Gm12878 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_H1hesc_segmentation: the genome segmentation of the cell line H1hesc using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Helas3_segmentation: the genome segmentation of the cell line Helas3 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Hepg2_segmentation: the genome segmentation of the cell line Hepg2 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Huvec_segmentation: the genome segmentation of the cell line Huvec using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted

Repressed or Low Activity region

ENCODE_K562_segmentation: the genome segmentation of the cell line K562 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

6. Column description for indel annotation files (WGSA v0.5)

#chr: chromosome number

pos: position (hg19)

ref: reference allele

alt: alternative allele

ANNOVAR_ensembl_summary: ANNOVAR consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_ensembl_summary: SnpEff consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

VEP_ensembl_summary: VEP consequence summary with Ensembl as gene model. Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

ANNOVAR_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

VEP_refseq_summary: SnpEff consequence summary with Refseq as gene model.

Format: GeneID(total number of transcripts):consequence#1(number of transcripts affected) consequence#2(number of transcripts affected)... Multiple genes are separated by "|"

SnpEff_ensembl_LOF: SnpEff Loss-Of-Function summary with Ensembl as gene model.

Format: GeneID(total number of transcripts): consequence#1(percentage of transcripts affected*total number of coding transcripts) consequence#2(percentage of transcripts affected*total number of coding transcripts)...

SnpEff_refseq_LOF: SnpEff Loss-Of-Function summary with Refseq as gene model.

Format: GeneID(total number of transcripts): consequence#1(percentage of transcripts affected*total number of coding transcripts) consequence#2(percentage of transcripts affected*total number of coding transcripts)...

rs_dbSNP141: rs number from dbSNP141

GWAS_catalog_rs: rs number according to GWAS catalog

GWAS_catalog_trait: associated trait according to GWAS catalog

GWAS_catalog_pubmedid: pubmedid of the paper describing the association

GRASP_rs: rs number by GRASP

GRASP_PMID: PMID number by GRASP

GRASP_p-value: p-value of the association test based on the SNP

GRASP_phenotype: phenotype the SNP associated with

GRASP_phenotype: phenotype the SNP associated with

GRASP_ancestry: population ancestry of the samples on which the association test was based

GRASP_platform: SNP platform on which the association test was based

clinvar_rs: rs number by clinvar

clinvar_clnsig: clinical significance by clinvar

2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response, 7 - histocompatibility. A negative score means the score is for the ref allele

clinvar_trait: the trait/disease the clinvar_clnsig referring to

COSMIC_ID: ID of the SNV at the COSMIC (Catalogue Of Somatic Mutations In Cancer) database

COSMIC_CNT: number of samples having this SNV in the COSMIC database

1000Gp3_AC: Alternative allele counts in the whole 1000 genomes phase 3 (1000Gp3) data.

1000Gp3_AF: Alternative allele frequency in the whole 1000Gp3 data.

1000Gp3_AFR_AC: Alternative allele counts in the 1000Gp3 African descendent samples.

1000Gp3_AFR_AF: Alternative allele frequency in the 1000Gp3 African descendent samples.

1000Gp3_EUR_AC: Alternative allele counts in the 1000Gp3 European descendent samples.

1000Gp3_EUR_AF: Alternative allele frequency in the 1000Gp3 European descendent samples.

1000Gp3_AMR_AC: Alternative allele counts in the 1000Gp3 American descendent samples.

1000Gp3_AMR_AF: Alternative allele frequency in the 1000Gp3 American descendent samples.

1000Gp3_EAS_AC: Alternative allele counts in the 1000Gp3 East Asian descendent samples.

1000Gp3_EAS_AF: Alternative allele frequency in the 1000Gp3 East Asian descendent samples.

1000Gp3_SAS_AC: Alternative allele counts in the 1000Gp3 South Asian descendent samples.

1000Gp3_SAS_AF: Alternative allele frequency in the 1000Gp3 South Asian descendent samples.

TWINSUK_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

TWINSUK_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

ALSPAC_AC: Alternative allele count in called genotypes in UK10K TWINSUK cohort.

ALSPAC_AF: Alternative allele frequency in called genotypes in UK10K TWINSUK cohort.

ESP6500_AA_AC: Alternative allele counts in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_AA_AF: Alternative allele frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_EA_AC: Alternative allele counts in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ESP6500_EA_AF: Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set).

ExAC_AC: Allele count in total ExAC samples (~60,706 unrelated individuals)

ExAC_AF: Allele frequency in total ExAC samples

ExAC_Adj_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC samples

ExAC_Adj_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC samples

ExAC_AFR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC samples

ExAC_AFR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC samples

ExAC_AMR_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC samples

ExAC_AMR_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC samples

ExAC_EAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC samples

ExAC_EAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC samples

ExAC_FIN_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC samples

ExAC_FIN_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC samples

ExAC_NFE_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

ExAC_NFE_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC samples

ExAC_SAS_AC: Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC samples

ExAC_SAS_AF: Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC samples

The following columns are based on per site/SNV annotations with a number within {} presenting the count of that annotation:

splicing_consensus_adaboost_score: splicing-change prediction for splicing consensus SNPs

based on adaboost. If the score >0.6, it predicts that the splicing will be changed,

otherwise it predicts the splicing will not be changed.

splicing_consensus_rf_score: splicing-change prediction for splicing consensus SNPs

based on random forest. If the score >0.6, it predicts that the splicing will be changed,

otherwise it predicts the splicing will not be changed

sno_miRNA_name: the name of snoRNA or miRNA if the site is located within (from UCSC)

sno_miRNA_type: the type of snoRNA or miRNA (from UCSC)

UTR3_miRNA_target: the gene-miRNA pair, if the site is located within a 3'UTR miRNA target (from UCSC)

MAP20: average Duke mappability score based on 20bp read,

0-1, higher score means higher mappability

MAP35: average Duke mappability score based on 35bp read

0-1, higher score means higher mappability

1000G_strict_masked: whether the site is within the 1000G strict masked region

Y (Yes) or N (No), Y means generally good mapping quality

RepeatMasker_masked: whether the site is masked by RepeatMasker

Y (Yes) or N (No), Y means generally lower mapping quality

phyloP46way_primate: a conservation score based on 46way alignment primate set,

the higher the more conservative

phyloP46way_placental: a conservation score based on 46way alignment placental set,

the higher the more conservative

phyloP100way_vertebrate: a conservation score based on 100way alignment vertebrate set,

the higher the more conservative

phastCons46way_primate: a conservation score based on 46way alignment primate set,

the higher the more conservative

phastCons46way_placental: a conservation score based on 46way alignment placental set,

the higher the more conservative

phastCons100way_vertebrate: a conservation score based on 100way alignment vertebrate set,

the higher the more conservative

GERP++_NR: GERP++ neutral rate

GERP++_RS: GERP++ RS score, the larger the score, the more conserved the site

SiPhy_29way_logOdds: SiPhy score based on 29 mammals genomes. The larger the score,

the more conserved the site

RegulomeDB_motif: motif the SNP resides (from RegulomeDB)

RegulomeDB_score: categorical score from RegulomeDB. The smaller, the more likely the SNP

affects binding

Motif_breaking: whether break a known motif (in-house script)

network_hub: whether the target gene is a network hub based on funseq-0.1

ENCODE_annotated: whether annotated by ENCODE based on funseq-0.1

sensitive: whether defined as sensitive region based on funseq-0.1

ultra_sensitive: whether defined as ultra-sensitive region based funseq-0.1

target_gene: target gene (for promoter, enhancer, etc.) based on funseq-0.1

funseq_noncoding_score: funseq-like noncoding score range 0-6, each of the previous 5 columns contribute 1 if "YES", or 0 if "NO"; the column Motif_breaking contribute 1 if it is not a "."

funseq2_noncoding_score: funseq2 noncoding score range 0-5.4 a weighted score designed for damaging prediction of cancer somatic SNPs

CADD_raw: CADD raw score, the larger the number the more likely damaging

CADD_phred: CADD phred-like score, ranges 1-99, the larger the number the more likely damaging; score >10 means the variant in the top 10% (0.1) among the total 8.6 billion possible SNVs, >20 means in the top 1%, >30 means in the top 0.1%, etc. CADD suggests a cutoff between 10 and 20 (e.g. 15)

fathmm-MKL_non-coding_score: fathmm-MKL non-coding prediction probability, the larger the number the more likely damaging; the threshold separating deleterious prediction and neutral prediction is 0.5.

fathmm-MKL_non-coding_group: fathmm-MKL non-coding group, the feature group used for the non-coding prediction

fathmm-MKL_coding_score; fathmm-MKL coding prediction probability, the larger the number the more likely damaging the threshold separating deleterious prediction and neutral prediction is 0.5.

fathmm-MKL_coding_group: fathmm-MKL coding group, the feature group used for the coding prediction.

ORegAnno_type: the type of regulatory region by ORegAnno

ORegAnno_PMID: the PMID of the paper describing the regulation

ENCODE_TFBS: name of the transcription factors (separated by ;) if the site is within a TFBS

ENCODE_TFBS_score: the higher the score the stronger the evidence of the TFBS

ENCODE_TFBS_cells: the cell lines (separated by ;) the TFBS was detected

ENCODE_Dnase_score: the higher the score the stronger the evidence of a DNase I hypersensitive site

ENCODE_Dnase_cells: number of cell lines supporting a DNase I hypersensitive site

Ensembl_Regulatory_Build_Overviews: genome segment prediction based on 17 cell types from ENCODE and Roadmap. Predicted states: ctcf - CTCF binding sites, distal - Predicted enhancers open - Unannotated open chromatin regions, proximal - Predicted promoter flanking regions, tfbs - Unannotated transcription factor binding sites, tss - Predicted promoters

FAMTOM5_enhancer: whether the site is within a FAMTOM5 predicted enhancer region Y (Yes) or N (No)

FAMTOM5_CAGE_peak: whether the site is within a FAMTOM5 Cap Analysis of Gene Expression (CAGE) peak. Y (Yes) or N (No). A CAGE peak generally suggests a promoter region

Ensembl_Regulatory_Build_TFBS: TFBS from Ensembl Regulatory Build. Multiple TFBS separated by ";"

Ensembl_Regulatory_Build_TFBS_prob: the probability of observing TFBS binding Multiple probabilities (corresponding to Ensembl_Regulatory_Build_TFBS) separated by ";"

The following columns are cell type specific:

Ensembl_A549_activity: A549 specific activity prediction from Ensembl Regulatory Build. Predicted states: open - Unannotated active open chromatin regions proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_DND41_activity: DND41 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_GM12878_activity: GM12878 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_H1HESC_activity: H1HESC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HELAS3_activity: HELAS3 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HEPG2_activity: HEPG2 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HMEC_activity: HMEC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMM_activity: HSMM specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HSMMT_activity: HSMMT specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_HUVEC_activity: HUVEC specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_K562_activity: K562 specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_MONO_activity: MONO specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHA_activity: NHA specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHDFAD_activity: NHDFAD specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHEK_activity: NHEK specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_NHLF_activity: NHLF specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_OSTEO_activity: OSTEO specific activity prediction from Ensembl Regulatory Build.

Predicted states: open - Unannotated active open chromatin regions

proximal - Predicted active promoter flanking regions/proximal enhancer

tss - Predicted active promoters

ctcf - Active CTCF binding sites

distal - Predicted active enhancers

tfbs - Unannotated active transcription factor binding sites

Ensembl_A549_segmentation: A549 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_DND41_segmentation: DND41 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_GM12878_segmentation: GM12878 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_H1HESC_segmentation: H1HESC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HELAS3_segmentation: HELAS3 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HEPG2_segmentation: HEPG2 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HMEC_segmentation: HMEC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HSMM_segmentation: HSMM specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HSMMT_segmentation: HSMMT specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_HUVEC_segmentation: HUVEC specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_K562_segmentation: K562 specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_MONO_segmentation: MONO specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHA_segmentation: NHA specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHDFAD_segmentation: NHDFAD specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHEK_segmentation: NHEK specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_NHLF_segmentation: NHLF specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

Ensembl_OSTEO_segmentation: OSTEO specific genome segment prediction from Ensembl Regulatory Build. Predicted states: proximal - Proximal enhancer

gene - Transcription associated

tss - Active promoter

ctcf - Distal CTCF

weak - Weak signal

distal - Distal enhancer

dead - Polycomb repressed

ENCODE_Gm12878_segmentation: the genome segmentation of the cell line Gm12878 using two

different unsupervised machine learning techniques (ChromHMM and Segway).
TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_H1hesc_segmentation: the genome segmentation of the cell line H1hesc using two different unsupervised machine learning techniques (ChromHMM and Segway).
TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Helas3_segmentation: the genome segmentation of the cell line Helas3 using two different unsupervised machine learning techniques (ChromHMM and Segway).
TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Hepg2_segmentation: the genome segmentation of the cell line Hepg2 using two different unsupervised machine learning techniques (ChromHMM and Segway).
TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_Huvec_segmentation: the genome segmentation of the cell line Huvec using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

ENCODE_K562_segmentation: the genome segmentation of the cell line K562 using two different unsupervised machine learning techniques (ChromHMM and Segway). TSS - Predicted promoter region including TSS, PF - Predicted promoter flanking region, E - Predicted enhancer, WE - Predicted weak enhancer or open chromatin cis regulatory element, CTCF - CTCF enriched element, T - Predicted transcribed region, R - Predicted Repressed or Low Activity region

7. Column description of Ensembl transcript-specific SNV annotation (WGSA v0.5)

#chr: chromosome number

pos: position (hg19)

ref: reference allele

alt: alternative allele

transcript_id: Ensembl transcript id

consequence_ANNOVAR: consequence by ANNOVAR annotation

gene_id_or_closest_genes_ANNOVAR: Ensembl gene id or the closest genes in the form of gene id:transcript id(dist=distance to the transcript)

HGVSc_ANNOVAR: SNV presented in the form of HGVSc by ANNOVAR

HGVSp_ANNOVAR: SNV presented in the form of HGVSp by ANNOVAR

exon_rank_ANNOVAR: which exon the SNV resides (if applicable) by ANNOVAR

consequence_snpEff: consequence by SnpEff

gene_name_snpEff: gene name (HGNC id) by SnpEff

codon_change_or_distance_snpEff: codon change in form of old codon/new codon, or distance to the transcript (for upstream or downstream) by SnpEff

amino_acid_change_snpEff: a HGVSc or HGVSp style presentation of the SNV by SnpEff

amino_acid_length_snpEff: transcription length divided by 3 by SnpEff

exon_or_intron_rank_snpEff: which intron/exon the SNV resides by SnpEff

impact_snpEff: SnpEff impact prediction (High, Moderate, Low, Modifier)

consequence_VEP: consequence by VEP

gene_name_VEP: gene name (HGNC id) by VEP

gene_id_VEP: Ensembl gene id by VEP

protein_id_VEP: Ensembl protein id by VEP

ccds_id_VEP: CCDS id by VEP

swissprot_id_VEP: SWISSPROT id by VEP

codon_change_or_distance_VEP: codon change in form of old codon/new codon, or distance to the transcript (for upstream or downstream) by VEP

amino_acid_change_VEP: amino acid change predicted by VEP

HGVSc_VEP: a HGVSc style presentation of the SNV by VEP

HGVSp_VEP: a HGVSp style presentation of the SNV by VEP

cDNA_position_VEP: cDNA position of the SNV by VEP

CDS_position_VEP: CDS position of the SNV by VEP

protein_position_VEP: protein position of the SNV by VEP

exon_or_intron_rank_VEP: which intron/exon the SNV resides by VEP

strand_VEP: which strand the gene is on by VEP

canonical_VEP: whether the transcript is canonical by VEP

Note: for all columns, missing data is presented as "."

Note: for all 3 annotation tools, upstream/downstream is defined as within 5 kb to the transcript

8. Column description of RefSeq transcript-specific SNV annotation (WGSA v0.5)

#chr: chromosome number

pos: position (hg19)

ref: reference allele

alt: alternative allele

transcript_id: RefSeq transcript id

consequence_ANNOVAR: consequence by ANNOVAR annotation

gene_name_or_closest_genes_ANNOVAR: gene name (HGNC id) or the closest genes in the form o

gene name:transcript id(dist=distance to the transcript)

HGVSc_ANNOVAR: SNV presented in the form of HGVSc by ANNOVAR

HGVSp_ANNOVAR: SNV presented in the form of HGVSp by ANNOVAR

exon_rank_ANNOVAR: which exon the SNV resides (if applicable) by ANNOVAR

consequence_snpEff: consequence by SnpEff

gene_name_snpEff: gene name (HGNC id) by SnpEff

codon_change_or_distance_snpEff: codon change in form of old codon/new codon, or distance

to the transcript (for upstream or downstream) by SnpEff

amino_acid_change_snpEff: a HGVSc or HGVSp style presentation of the SNV by SnpEff

amino_acid_length_snpEff: transcription length divided by 3 by SnpEff

exon_or_intron_rank_snpEff: which intron/exon the SNV resides by SnpEff

impact_snpEff: SnpEff impact prediction (High, Moderate, Low, Modifier)

consequence_VEP: consequence by VEP

gene_name_VEP: gene name (HGNC id) by VEP

gene_id_VEP: RefSeq gene id by VEP

protein_id_VEP: RefSeq protein id by VEP

codon_change_or_distance_VEP: codon change in form of old codon/new codon, or distance to the transcript (for upstream or downstream) by VEP

amino_acid_change_VEP: amino acid change predicted by VEP

HGVSc_VEP: a HGVSc style presentation of the SNV by VEP

HGVSp_VEP: a HGVSp style presentation of the SNV by VEP

cDNA_position_VEP: cDNA position of the SNV by VEP

CDS_position_VEP: CDS position of the SNV by VEP

protein_position_VEP: protein position of the SNV by VEP

exon_or_intron_rank_VEP: which intron/exon the SNV resides by VEP

strand_VEP: which strand the gene is on by VEP

canonical_VEP: whether the transcript is canonical by VEP

Note: for all columns, missing data is presented as "."

Note: for all 3 annotation tools, upstream/downstream is defined as within 5 kb to the transcript

References:

1. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32,** 894–899 (2011).
2. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34,** E2393–2402 (2013).
3. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42,** 13534–13544 (2014).
4. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
5. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinforma. Oxf. Engl.* (2015). doi:10.1093/bioinformatics/btv009
6. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
7. Fu, Y. *et al.* FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15,** 480 (2014).
8. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).
9. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
10. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
11. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
12. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).

13. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).

14. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30,** i185–i194 (2014).

15. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* **6,** e1001025 (2010).

16. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034 –1050 (2005).

17. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).

18. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

20. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).

21. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

22. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

23. Behar, D. M. *et al.* A 'Copernican' Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).

24. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).

25. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).

26. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

27. Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* **34,** D158–162 (2006).

28. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–D73 (2014).

29. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).

30. Griffith, O. L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36,** D107–D113 (2008).

31. Zia, A. & Moses, A. M. Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. *BMC Bioinformatics* **12,** 299 (2011).

32. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* **27,** 2156–2158 (2011).

33. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

34. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

35. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–763 (2014).

36. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2015).