Humanizing Genomics
macrogen

Exome Sequencing

Report

2024.07

WES

Research use only

Table of Contents

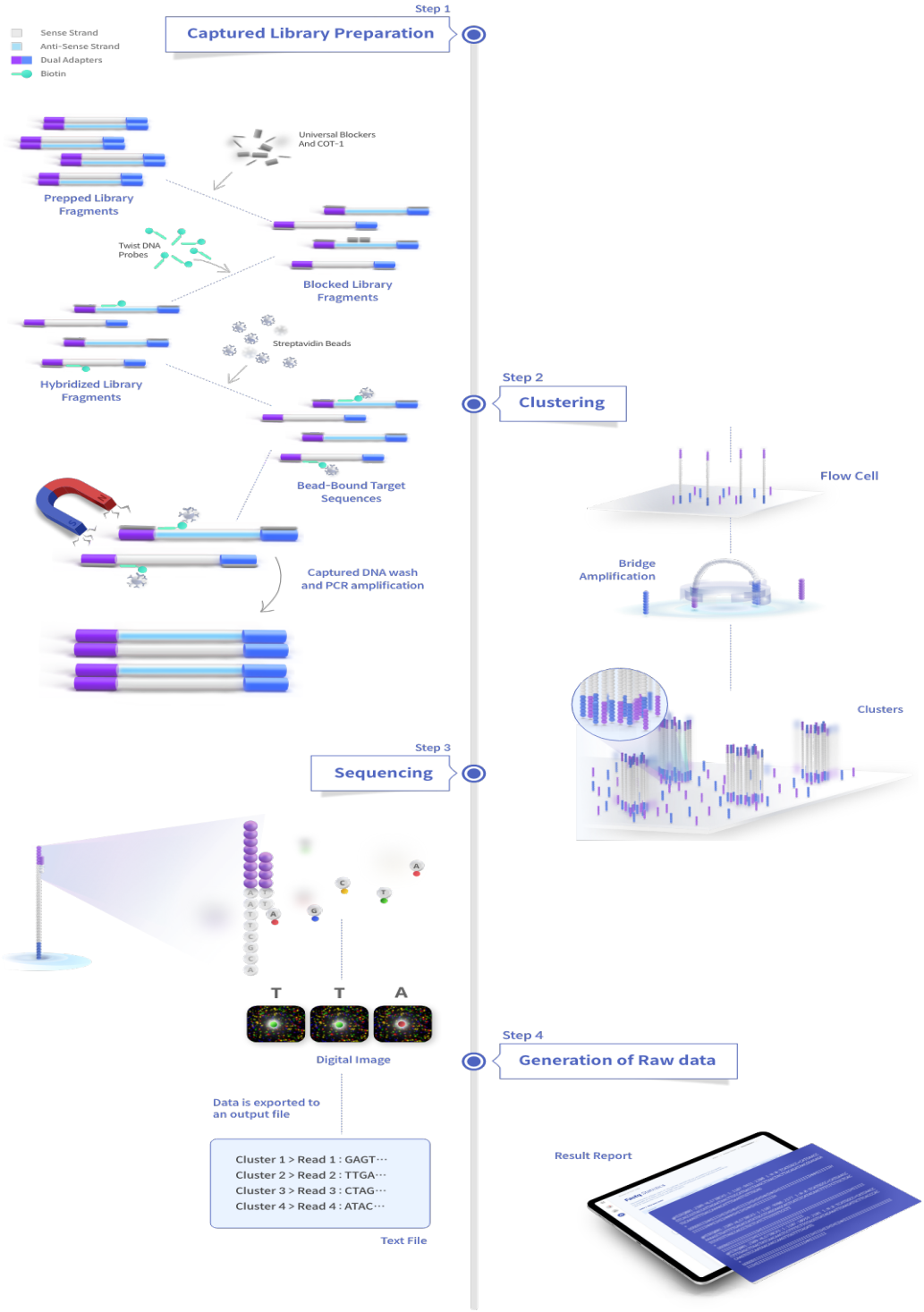
Order Information

Client Name	NGS_service
Client Organization	(주)마크로젠
Order Number	HN00000000
Application	Exome Sequencing
Type of Read	Paired-end
Read Length	101
Library kit	Twist Human Core Exome 2.0
Type of Sequencer	Illumina platform

Workflow

Experiments Workflow

The Twist NGS Target Enrichment workflow is solution-based system utilizing ultra-long 120 mer biotinylated cDNA baits – to capture regions of interest, enriching them out of NGS genomic fragment library.



Captured Library Construction

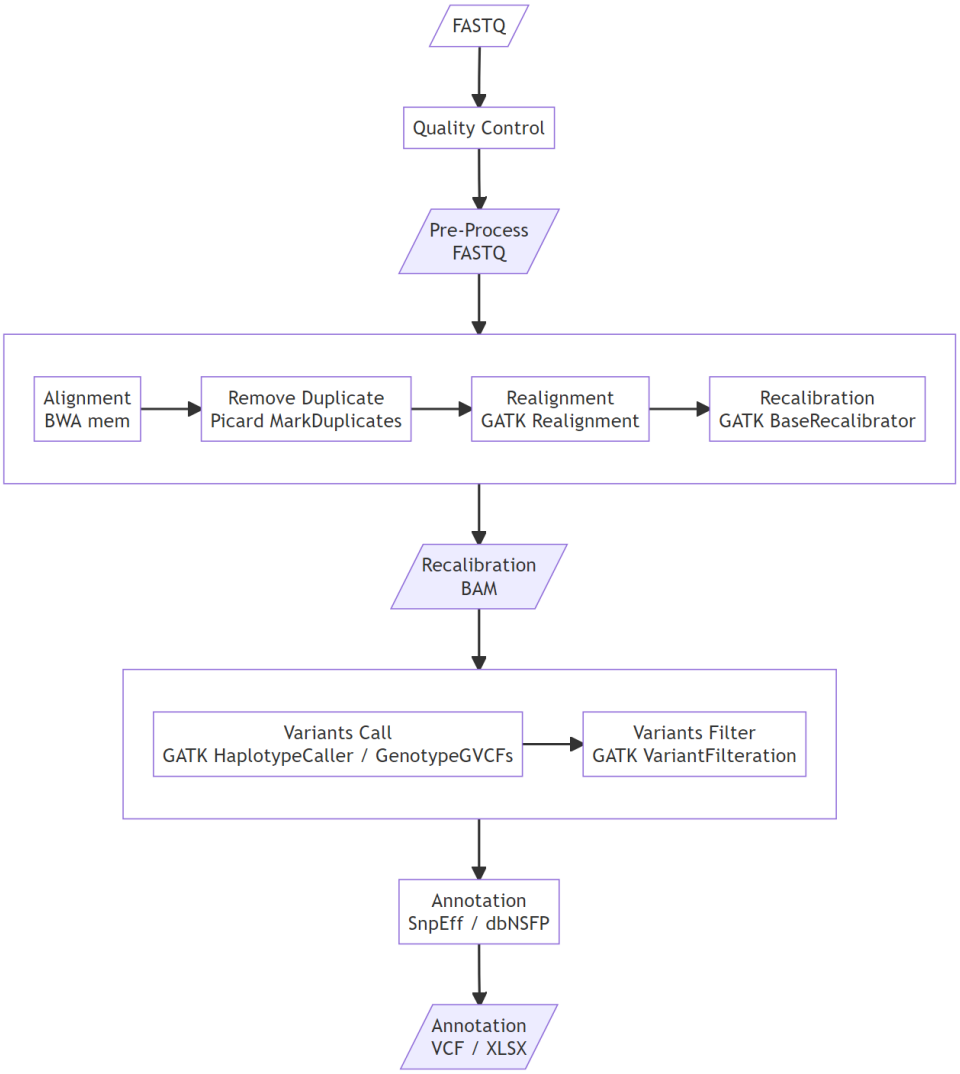
Target Enrichment and Sequencing A gDNA library was prepared from 50ng of input gDNA using the Twist Library Preparation EF Kit (96 samples, PN 101058) with full-length combinatorial dual index TruSeq-compatible Y-adapters (Illumina) according to the Twist Bioscience Library Protocol. The DNA quantity and quality is measured by PicoGreen and agarose gel electrophoresis, respectively. We use 50ng of each gDNA diluted with EB Buffer and sheared to a target peak size of 200bp with the fragmentation enzyme. Fragmentation is followed by end-repair and the addition of 'A' tail. Twist CD index adapters are then ligated to the fragments. After assessing the efficiency of ligation, the adapter ligated product is PCR amplified.
The final purified product is quantified by TapeStation DNA screentape D1000 (Agilent) and PicoGreen. For exome capture, each hybridization reaction requires a total of 1500 ng of indexed libraries, made by pooling equal amounts from 8 individual libraries. And mixed with Fast hybridization mix, Twist exome probe, Blocker solution, and Universal Blocker, according to the Twist library preparation protocol. The captured DNA is washed and amplified. The final purified product is quantified by qPCR according to the qPCR Quantification Protocol Guide (KAPA Library Quantification kits for Illumina Sequencing platforms) and qualified by the TapeStation DNA screentape D1000 (Agilent).

Clustering & Sequencing

Illumina utilizes a unique amplification reaction that occurs on the surface of the flow cell. A flow cell containing millions of unique clusters is loaded into the Illumina platform for automated cycles of extension and imaging. Sequencing-by-Synthesis utilizes four proprietary nucleotides possessing reversible fluorophore and termination properties. Each sequencing cycle occurs in the presence of all four nucleotides leading to higher accuracy than methods where only one nucleotide is present in the reaction mix at a time. This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster.

Workflow

Analysis Workflow



FASTQ

The Illumina platform generates raw images and base calling with an integrated primary analysis software called RTA. The base calling files which are expressed in binary are converted into FASTQ by Illumina package bcl2fastq v2.20.0. The demultiplexing option (—barcode-mismatches) is set as value : 0. 'FastQC' is then used to check the sequencing quality.

ALIGN

Paired-end sequences produced by NovaSeq Instrument are firstly mapped to the human reference genome using the mapping program 'BWA'. (BWA-MEM is used out of the three algorithms provided by BWA) The mapping result file is generated in BAM format, without unordered sequences and alternate haplotypes.

Mark Duplicates

PCR duplicates are removed using MarkDuplicates.jar from 'Picard-tools' package, which requires reads to be sorted. Reads with identical starting positions are considered as duplicates and reduced into a single read.

Base Quality Score Recalibration

BAM files are then recalibrated with Base Quality Score Recalibration (BQSR). BQSR is a process which uses machine learning to model the sequencing errors empirically and adjust the quality scores accordingly.

Variant Calling

Based on the BAM file previously generated, variant genotyping for each sample is performed with Haplotype Caller of GATK. In this stage SNP and short indels candidates are detected at nucleotide resolution.

Variant Filtering

We filter variants with VariantFiltration of GATK Tool. This tool is designed for hard-filtering variant calls based on certain criteria. Records are hard-filtered by changing the value in the FILTER field to something else other than PASS. Filtered records will be preserved in the output unless their removal is requested in the command line.

Annotation

Filtered variants are annotated with another program called SnpEff and filtered with dbSNP and SNPs from the 1000 genome project. The format of the final product is in vcf. Then, in-house program and SnpEff are used to annotate with additional databases, including ESP6500, ClinVar, dbNSFP, ACMG information.

Database Version

Software	Version
Mapping Reference	hg38 from UCSC (original GRCh38 from NCBI, Dec. 2013)
dbSNP	156
1000Genome	Phase3
Clinvar	2024-07-16
ESP	ESP6500SI_V2
dbNSFP	dbNSFPv4.5c

Tool Version

Software	Version
BWA	bwa-0.7.17
Picard	picard-tools-Version:3.1.1
GATK	GATKv4.5.0.0
SnpEff	SnpEff 5.2 2023-09-29

Tool Parameter

Software	Parameter	Value	Remark
BWA-MEM	-M		Mark shorter split hits as secondary(for Picard compatibility)
Picard	VALIDATION_STRINGENCY	LENIENT	Improve performance when validate of stringency
Picard	SO	coordinate	Sort order
Picard	REMOVE_DUPLICATES	true	
Picard	AS	true	Assume Sorted
Picard	CREATE_INDEX	true	Create index files
GATK		BaseRecalibrator	Generate the first pass recalibration
GATK		HaplotypeCaller	Call SNPs and indels simultaneously via local re-assembly of haplotypes in an active region
GATK		SelectVariants	Selects variants from a VCF source
GATK		VariantFiltration	Filters variant calls using a number of user-selectable, parameterizable criteria
GATK		CombineVariants / MergeVcfs	Combines VCF records from different sources
GATK	-knownSites	1000G_phase1.indels.hg38.vcf	Database of known polymorphic sites
GATK	-knownSites	dbsnp_138.hg38.vcf	Database of known polymorphic sites
GATK	-knownSites	Mills_and_1000G_gold_standard . indels.hg38.sites.vcf	Database of known polymorphic sites

Analysis Result

Result per Order

◉ Fastq Statistics

Sample ID	Total yield(bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
NA12878	4,356,415,224	43,132,824	48.95	51.05	98.51	97.45

- Sample ID : Sample name.
- Total yield (bp) : Total number of bases sequenced.
- Total reads : Total number of reads.
- GC(%) : GC Content
- AT(%) : AT Content
- Q20(%) : Ratio of bases that have phred quality score of over 20.
- Q30(%) : Ratio of bases that have phred quality score of over 30.
- Quality by Cycle : This plot shows the average quality at each cycle

◉ Alignment & Coverage & Variant Statistics

	NA12878
Total reads	43,132,824
Average read length(bp)	101.0
Total yield(bp)	4,356,415,224
Target regions(bp)	36,452,732
Average throughput depth of target regions(X)	119.5
Initial mappable reads	43,080,349
Initial mappable reads(%)	99.8
Non-redundant reads	41,504,571
Non-redundant reads(%)	96.3
On-target reads	27,775,371
On-target reads(%)	66.9
On-target yield(bp)	1,975,670,760
Mean depth of target regions(X)	54.2
Number of on-target genotypes (>=1X)	36,381,560
% Coverage of target regions (>=1X)	99.8
Number of on-target genotypes (>=10X)	36,361,626
% Coverage of target regions (>=10X)	99.7
Number of on-target genotypes (>=20X)	36,254,145
% Coverage of target regions (>=20X)	99.4
Number of on-target genotypes (>=30X)	35,168,127
% Coverage of target regions (>=30X)	96.4
Number of on-target genotypes (>=50X)	21,920,245
% Coverage of target regions (>=50X)	60.1
Number of SNP	76,299
Synonymous Variant	12,491
Missense Variant	12,996
Stop Gained	150
Stop Lost	39
Number of INDEL	13,484
Frameshift Variant	385
Inframe Insertion	223
Inframe Deletion	293
% Found in dbSNP156	99.1
Het/Hom Ratio	1.8
Ts/Tv Ratio	2.3

- Sample ID : Sample name.
- Total reads : Total Number of Reads.
- Average read length(bp) : Average length of reads after trimming.
- Total yield(bp) : Total yield(bp)
- Target regions(bp) : Target region size.
- Average throughput depth of target regions(X) : {Total yield} / {Target regions}.
- Initial mappable reads : Number of reads mapped to reference.
- Initial mappable reads(%) : 100 * {Initial mappable reads} / {Total reads}.
- Non-redundant reads : Number of de-duplicate reads from Picard tools.
- Non-redundant reads(%) : 100 * {Non-redundant reads} / {Initial mappable reads}.
- On-target reads : Number of reads mapped to target regions.
- On-target reads(%) : 100 * {On-target reads} / {Non-redundant reads}.
- On-target yield(bp) : The sum of the bases in the final alignment to the target regions.
- Mean depth of target regions(X) : {On-target yield} / {Target regions}.
- % Coverage : The percentage of bases in target regions with a depth of coverage or greater.
- Number of SNP : Total Number of SNP
- Synonymous Variant : Variant causes a codon that produces the same amino acid e.g.: Ttg/Ctg, L/L.
- Missense Variant : Variant causes a codon that produces a different amino acid e.g.: Tgg/Cgg, W/R.
- Stop Gained : Variant causes a STOP codon e.g.: Cag/Tag, Q/*.
- Stop Lost : Variant causes stop codon to be mutated into a non-stop codon e.g.: Tga/Cga, */R.
- Number of INDEL : Total Number of INDEL
- Frameshift Variant : Insertion or deletion causes a frame shift e.g.: An indel size is not multple of 3.
- Inframe Insertion : One or many codons are inserted e.g.: An insert multiple of three in a codon boundary.
- Inframe Deletion : An inframe non synonymous variant that deletes bases from the coding sequence.
- % Found in dbSNP156 : The percentage of mutations found in dbSNP.
- Het/Hom Ratio : Ratio of number of heterozygous variants to number of homozygous variants.
- Ts/Tv Ratio : Ratio of transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A,G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T).

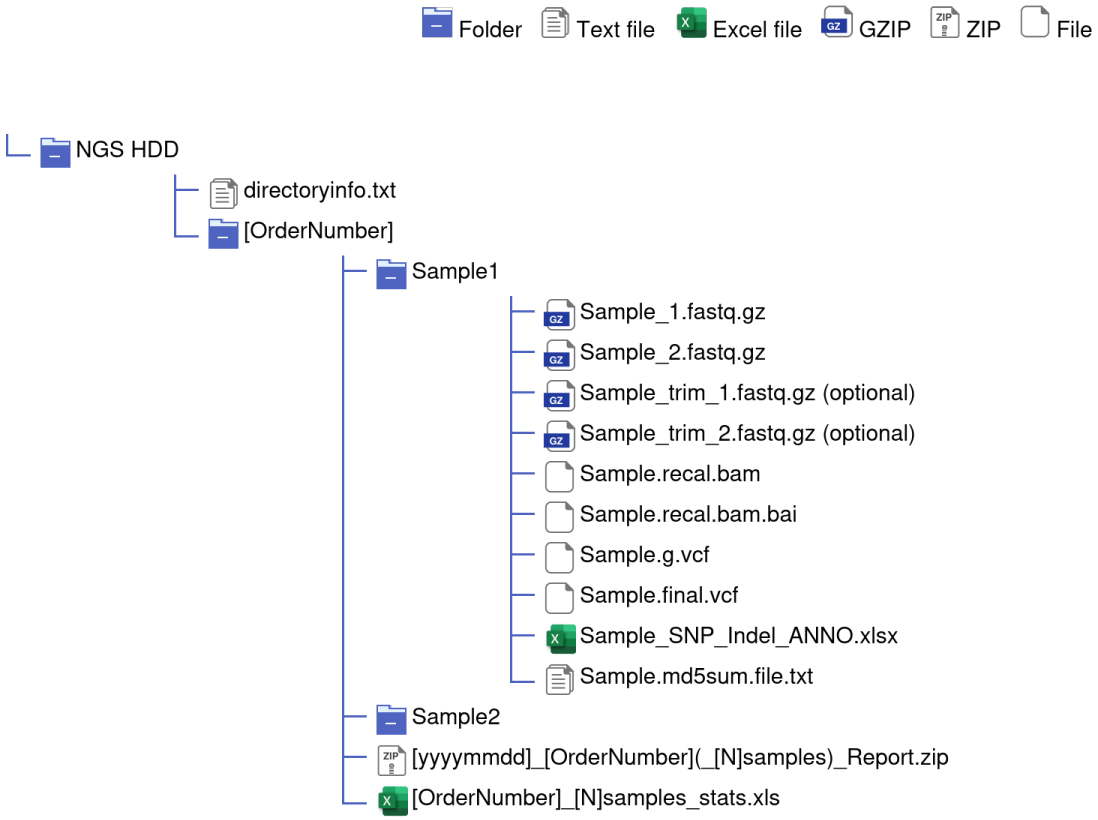
Deliverables

Analysis Result

The data can be downloaded from the links below. The download links are active for 2 weeks only, so please download your data within this period.
Once you receive/download the data, please make sure to check the integrity of the files. Please note that the sequencing files will be deleted from our server 3 months after the analysis report is released; please contact us within 3 months if you encounter a problem with the data.

File name	File Size	md5sum
HN00000000_1samples_Annotation.zip	38.7M	badce1281fcf8d6dbc6f1df08a6305e9
NA12878.tar	5.1G	51c908a15d669a12a03e426a20b91286

Analysis Result Tree



Result File Description

Deliverables List

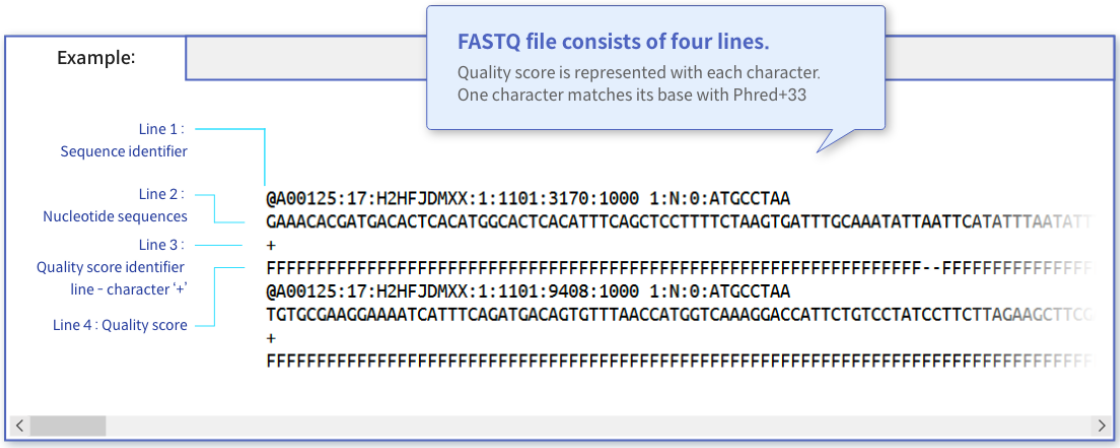
File Type	File Name	Description
FASTQ file	[Sample name 1]_[read1].fastq.gz	Raw read1 sequence data
FASTQ file	[Sample name 1]_[read2].fastq.gz	Raw read2 sequence data
BAM file	Sample.recal.bam	BWA alignment file
BAM file	Sample.recal.bam.bai	BWA alignment index file
Variant Call Result	Sample.final.vcf	SNP/INDEL file (vcf format)
Variant Call Result	Sample.g.vcf	Genomic VCF
Variant Call Result	Sample_SNP_Indel_ANNO.xlsx	Annotated variant list file (excel file)
Summary	All_samples_stats.xlsx	Analysis stats report of all samples (excel file)
md5sum	[Order#]_[#samples]_md5sum[_DownloadLink].txt	MD5 is a string of 32 hexadecimal values, which represents a 'fingerprint' of a file. By comparing the supplied MD5 value to the actual value computed by the MD5sums utility, you can make sure that the file that you downloaded off of the internet has not been tampered with or modified from the original file stored in our server.

File Format

◉ FASTQ File

FASTQ File

FASTQ Format



Phred Scores

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Q-Score Binning

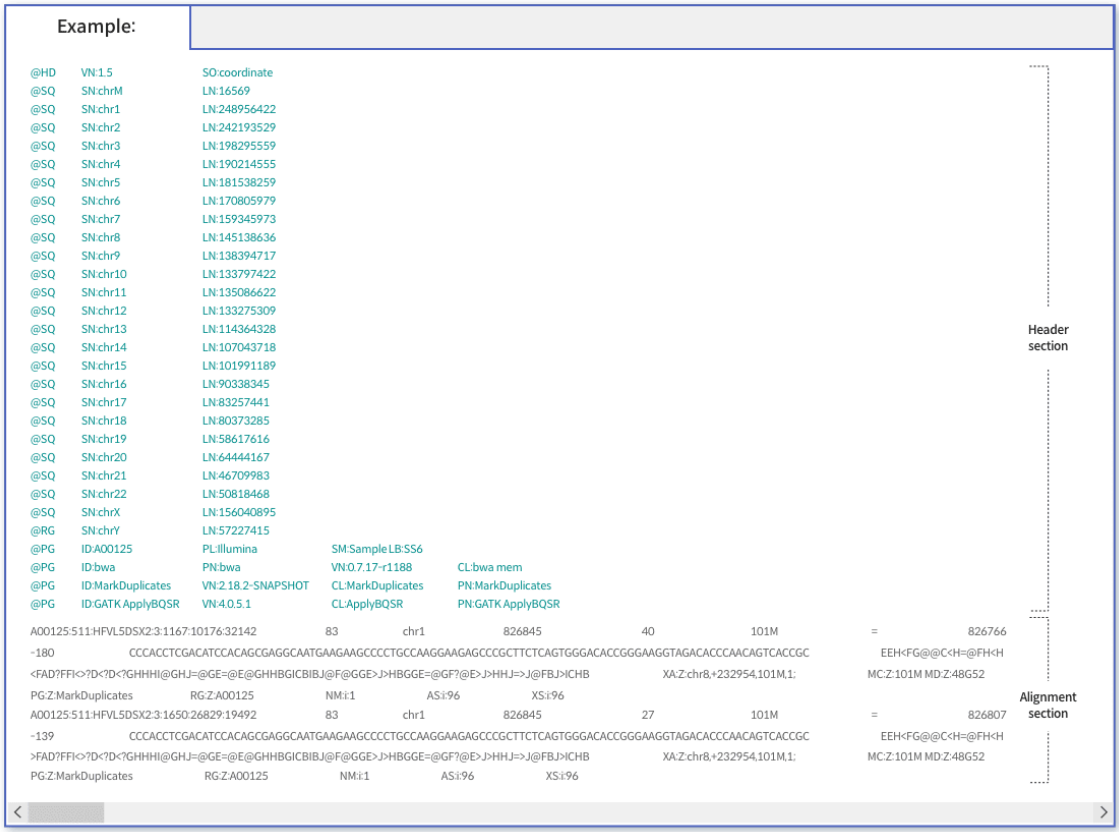
Phred Quality Score	Example of Empirically Mapped Q-Scores
N(no call)	N(no call)
2-9	7
10-19	11
20-24	22
25-29	27
30-34	32
35-39	37
40-45	42

◉ BAM File

BAM File

BAM

The BAM is a compressed binary format of a SAM(Sequence Alignment Map), The BAM file contains information about sequence alignment of Reads against a large reference sequence.



Header Line

Tag	Description
@HD	The header line
@PG	Program and command line
@RG	Read group. platform, sample name information
@SQ	Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order.
@CO	One-line text comment. Unordered multiple @CO lines are allowed.

SAM file : Alignments section mandatory fields

Field	Description
QNAME	Query template NAME
FLAG	Bitwise FLAG
RNAME	Reference sequence NAME
POS	1-based leftmost mapping POSition
MAPQ	MAPping Quality
CIGAR	CIGAR string
RNEXT	Ref. name of the mate/NEXT read
PNEXT	Position of the mate/NEXT read
TLEN	Observed Template LENgth
SEQ	segment SEQuence
QUAL	ASCII of Phred-scaled base QUALity+33

◉ VCF FIle

VCF File

VCF (SNV/INDEL)

The Variant Call Format (VCF) is a text file format that contains information about variants found at specific positions in a reference genome. The file format consists of meta-information lines, a header line, and data lines. Each data line contains information about a single variant.

Example:

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description=“Low quality">
##FILTER=<ID=MG_INDEL_Filter,Description=“QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0">
##FILTER=<ID=MG_SNP_Filter,Description=“QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description=“Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description=“Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description=“Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description=“Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description=“Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description=“Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description=“Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description=“Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description=“Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description=“Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
##INFO=<ID=AC,Number=A,Type=Integer,Description=“Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description=“Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description=“Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description=“Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=ClippingRankSum,Number=1,Type=Float,Description=“Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases">
##INFO=<ID=DP,Number=1,Type=Integer,Description=“Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description=“Were any of the samples downsampled?">
##INFO=<ID=END,Number=1,Type=Integer,Description=“Stop position of the interval">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description=“Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,Number=1,Type=Float,Description=“Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description=“Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation">
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description=“Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed">
##INFO=<ID=MLEAF,Number=A,Type=Float,Description=“Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the same order as listed">
##INFO=<ID=MQ,Number=1,Type=Float,Description=“RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description=“Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description=“Variant Confidence/Quality by Depth">
##INFO=<ID=RAW_MQ,Number=1,Type=Float,Description=“Raw data for RMS Mapping Quality">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description=“Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description=“Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##source=GenotypeGVCFs
#CHROM  POS     ID  REF  ALT  QUAL     FILTER  INFO         FORMAT                            Sample1
Chr1    817514  .   T    C    1348.77  PASS    GT:AD:DP:GQ:PL  AC=2;AF=1.00;AN=2;DP=38;ExcessHet=
3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=40.03;QD=25.36;SOR=7.328  1/1:0,38:38:99:1377,114,0
Chr1    817514  .   T    C    1348.77  PASS    GT:AD:DP:GQ:PL  AC=2;AF=1.00;AN=2;DP=38;ExcessHet=
3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=40.03;QD=25.36;SOR=7.328  1/1:0,38:38:99:1377,114,0
Chr1    817514  .   T    C    1348.77  PASS    GT:AD:DP:GQ:PL  AC=2;AF=1.00;AN=2;DP=38;ExcessHet=
3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=40.03;QD=25.36;SOR=7.328  1/1:0,38:38:99:1377,114,0
```

Meta Information lines

Header line

Data lines

Header Line

Header	Description
#CHROM	Chromosome
POS	Position (with the 1st base having position 1)
ID	The dbSNP rs identifier of the SNP
REF	Reference base(s)
ALT	Comma separated list of alternate non-reference alleles called on at least one of the samples
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant (and lower probability of errors).
FILTER	Filter status: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated below list of codes for filters that fail. See FILTER tag table for possible entries.
INFO	Additional information: INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: =. The exact format of each INFO sub-field should be specified in the meta-information. See INFO tag table for possible entries.
FORMAT	See FORMAT tag table for possible entries.

FILTER Tag

Tag	Description
LowQual	Low quality
MG_SNP_Filter	QD 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0
MG_INDEL_Filter	QD 200.0 || ReadPosRankSum < -20.0

INFO Tag

Tag	Description
AC	Allele count in genotypes, for each ALT allele, in the same order as listed
AF	Allele Frequency, for each ALT allele, in the same order as listed
AN	Total number of alleles in called genotypes
BaseQRankSum	Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
ClippingRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases
DB	dbSNP Membership
DP	Approximate read depth; some reads may have been filtered
FS	Phred-scaled p-value using Fisher's exact test to detect strand bias
HaplotypeScore	Consistency of the site with at most two segregating haplotypes
InbreedingCoeff	Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation
MLEAC	Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed
MLEAF	Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the same order as listed
MQ	RMS Mapping Quality
MQ0	Total Mapping Quality Zero Reads
MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
QD	Variant Confidence/Quality by Depth
ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
SOR	Symmetric Odds Ratio of 2x2 contingency table to detect strand bias
set	Source VCF for the merged record in CombineVariants
SNP	Variant is a SNP
MNP	Variant is an MNP
INS	Variant is an insertion
DEL	Variant is an deletion
MIXED	Variant is mixture of INS/DEL/SNP/MNP
HOM	Variant is homozygous
HET	Variant is heterozygous
VARTYPE	Comma separated list of variant types. One per allele.

FORMAT Tag

Tag	Description
GT	Genotype0/0 - the sample is homozygous reference0/1 - the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles1/1 - the sample is homozygous alternate
AD	Allelic depths for the ref and alt alleles in the order listed.
DP	Read depth at this position for this sample
GQ	Conditional genotype quality, encoded as a phred quality
PL	The normalized, Phred-scaled likelihoods for each of the 0/0, 0/1, and 1/1, without priors. The most likely genotype (given in the GT field) is scaled so that it's P = 1.0 (0 when Phred-scaled), and the other likelihoods reflect their Phred-scaled likelihoods relative to this most likely genotype.

Appendix

Annotation Column

The *_SNP_indel_ANNO.xlsx file contains information about variants found at specific positions in the reference genome. Each data line contains information about a single variant. Each column of the file has the following meaning.

Column	Description
CHROM	Chromosome
POS	Start Position (with the 1st base having position 1)
REF	Reference base(s)
ALT	Comma separated list of alternate non-reference alleles called on at least one of the samples
DP	Filtered base call depth used for site genotyping
AD	Allelic depths for the ref and alt alleles in the order listed. For indels, this value only includes reads that confidently support each allele (posterior probability 0.999 or higher that read contains indicated allele vs all other intersecting indel alleles)
QUAL	The Phred scaled probability that a REF/ALT polymorphism exists at this site given sequencing data. Because the Phred scale is -10 * log(1-p), a value of 10 indicates a 1 in 10 chance of error, while a 100 indicates a 1 in 10^10 chance.
MQ	Mapping Quality
Zygosity	Homo/Hetero
FILTER	Filter status: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated below list of codes for filters that fail.
Effect	Annotated using Sequence Ontology terms. Multiple effects can be concatenated using '&'.
Putative_Impact	A simple estimation of putative impact / deleteriousness : {HIGH, MODERATE, LOW, MODIFIER}
Gene_Name	Common gene name (HGNC). Optional: use closest gene when the variant is 'intergenic'.
Feature_Type	Which type of feature is in the next field (e.g. transcript, motif, miRNA, etc.). It is preferred to use Sequence Ontology (SO) terms, but 'custom' (user defined) are allowed.
Feature_ID	Depending on the annotation, this may be: Transcript ID (preferably using version number), Motif ID, miRNA, ChipSeq peak, Histone mark, etc. Note: Some features may not have ID (e.g. histone marks from custom Chip-Seq experiments may not have a unique ID).
Transcript_BioType	The bare minimum is at least a description on whether the transcript is {'Coding', :'Noncoding'}. Whenever possible, use ENSEMBL biotypes.
Rank/Total	Exon or Intron rank / total number of exons or introns.
HGVS.c	Variant using HGVS notation (DNA level)
HGVS.p	If variant is coding, this field describes the variant using HGVS notation (Protein level). Since transcript ID is already mentioned in 'feature ID', it may be omitted here.
REF_AA	reference amino acid
ALT_AA	alternative amino acid
cDNA_pos	Position in cDNA (one based)
cDNA_length	Transcript's cDNA length
CDS_pos	Position of coding bases (one based includes START and STOP codons).
CDS_length	Number of coding bases (one based includes START and STOP codons).
AA_pos	Position of AA (one based, including START, but not STOP).
AA_length	Number of AA (one based includes START and STOP codons).
Distance	All items in this field are options, so the field could be empty. - Up/Downstream: Distance to first / last codon - Intergenic: Distance to closest gene - Distance to closest Intron boundary in exon (+/up/downstream). If same, use positive number. - Distance to closest exon boundary in Intron (+/up/downstream) - Distance to first base in MOTIF - Distance to first base in miRNA - Distance to exonintron boundary in splice_site or splice_region - ChipSeq peak: Distance to summit (or peak center) - Histone mark / Histone state: Distance to summit (or peak center)
dbSNP138_ID	dbSNP138 rsNo.
dbSNP156_ID	dbSNP156 rsNo.
p3_1000G_AF	Non-reference allele frequency of existing variation in 1000 Genomes
p3_1000G_AFR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined African population
p3_1000G_AMR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined American population

Column	Description
p3_1000G_EAS_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined East Asian population
p3_1000G_EUR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined European population
p3_1000G_SAS_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined South Asian population
ESP6500_MAF_EA	Minor allele and frequency in the European American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set)
ESP6500_MAF_AA	Minor allele and frequency in the African American samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set)
ESP6500_MAF_ALL	Minor allele and frequency in all samples of the NHLBI GO Exome Sequencing Project (ESP6500 data set)
CLINVAR_CLNSIG	Clinical significance for this single variant. . Affects association Benign Benign/Likely_benign Conflicting_interpretations_of_pathogenicity drug_response Likely_benign not_provided other Pathogenic protective risk_factor Uncertain_significance
CLINVAR_CLNDISDB	Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN
CLINVAR_CLNDN	ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB
CLINVAR_CLNREVSTAT	ClinVar review status for the Variation ID ClinVar Review Status, mult - Classified by multiple submitters, single - Classified by single submitter, not - Not classified by submitter, exp - Reviewed by expert panel, prof - Reviewed by professional society
ACMG_SF_v3.2	
REF_AA_dbnsfp	reference amino acid in dbNSFP
ALT_AA_dbnsfp	alternative amino acid in dbNSFP
hg19_chr	chromosome as to hg19, '.' means missing
hg19_pos(1-based)	physical position on the chromosome as to hg19 (1-based coordinate). For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)
cds_strand	separated by ';'
refcodon	coding sequence (CDS) strand (+ or -)
codonpos	reference codon
codon_degeneracy	position on the codon (1, 2 or 3)
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 ',', corresponding to Ensembl_proteinid.
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 0.00963 to 0.91219.
SIFT_pred	If SIFTori is smaller than 0.05 (rankscore>0.395) the corresponding nsSNV is predicted as D(amaging); otherwise it is predicted as T(olerated). Multiple predictions 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. 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 0.00162 to 0.84324.
LRT_pred	LRT prediction, D(eleterious), N(eutral) or U(nknown), which is not solely determined by the score.
LRT_Omega	estimated nonsynonymous-to-synonymous-rate ratio (Omega, reported by LRT)
MutationTaster_score	MutationTaster p-value (MTori), ranges from 0 to 1. Multiple scores are separated by ','. Information on corresponding transcript(s) can be found by querying http://www.mutationtaster.org/ChrPos.html
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. If there are multiple scores of a SNV, only the largest MTnew was used in ranking. The rankscore is the ratio of the rank of the score over the total number of MTnew scores in dbNSFP. The scores range from 0.08979 to 0.81033.
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 MTnew and 0.31713 for the rankscore.
MutationTaster_model	MutationTaster prediction models.
MutationTaster_AAE	MutationTaster predicted amino acid change.
MutationAssessor_score	MutationAssessor functional impact combined score (MAori). The score ranges from -5.135 to 6.49 in dbNSFP.
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, 1.935 and 0.8, respectively. The rankscore cutoffs between H and M, M and L, and L and N, are 0.92922, 0.51944 and 0.19719, respectively.
FATHMM_score	FATHMM default score (weighted for human inherited-disease mutations with Disease Ontology) (FATHMMori). Scores range from -16.13 to 10.64. The smaller the score the more likely the SNP has damaging effect. Multiple scores separated by ',', corresponding to Ensembl_proteinid.

Column	Description
FATHMM_converted_rankscore	FATHMMori scores were first converted to FATHMMnew=1-(FATHMMori+16.13)/26.77, then ranked among all FATHMMnew scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMnew 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 =0.81332) the corresponding nsSNV is predicted as D(AMAGING); otherwise it is predicted as T(OLERATED). Multiple predictions separated by ',', corresponding to Ensembl_proteinid.
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 ',', corresponding to Ensembl_proteinid.
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. The scores range from 0 to 1.
PROVEAN_pred	If PROVEANori =0.543) the corresponding nsSNV is predicted as D(amaging); otherwise it is predicted as N(eutral). Multiple predictions separated by ',', corresponding to Ensembl_proteinid.
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 0.82268.
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 0.81113.
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.
M-CAP_score	M-CAP score (details in DOI: 10.1038/ng.3703). Scores range from 0 to 1. The larger the score the more likely the SNP has damaging effect.
M-CAP_rankscore	M-CAP scores were ranked among all M-CAP scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of M-CAP scores in dbNSFP.
M-CAP_pred	Prediction of M-CAP score based on the authors' recommendation, T(olerated) or D(amaging). The score cutoff between D and T is 0.025.
MutPred_score	General MutPred score. Scores range from 0 to 1. The larger the score the more likely the SNP has damaging effect.
MutPred_rankscore	MutPred scores were ranked among all MutPred scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MutPred scores in dbNSFP.
MutPred_protID	UniProt accession or Ensembl transcript ID used for MutPred_score calculation.
MutPred_AAchange	Amino acid change used for MutPred_score calculation.
MutPred_Top5features	Top 5 features (molecular mechanisms of disease) as predicted by MutPred with p values. MutPred_score > 0.5 and p 0.75 and p 0.75 and p < 0.01 are referred to as very confident hypotheses.
fathmm-MKL_coding_score	fathmm-MKL p-values. Scores range from 0 to 1. SNVs with scores >0.5 are predicted to be deleterious, and those <0.5 are predicted to be neutral or benign. Scores close to 0 or 1 are with the highest-confidence. Coding scores are trained using 10 groups of features. More details of the score can be found in doi: 10.1093/bioinformatics/btv009.
fathmm-MKL_coding_rankscore	fathmm-MKL coding scores were ranked among all fathmm-MKL coding scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of fathmm-MKL coding scores in dbNSFP.
fathmm-MKL_coding_pred	If a fathmm-MKL_coding_score is >0.5 (or rankscore >0.28317) the corresponding nsSNV is predicted as 'D(AMAGING)''; otherwise it is predicted as 'N(EUTRAL)'.
fathmm-MKL_coding_group	the groups of features (labeled A-J) used to obtained the score. More details can be found in doi: 10.1093/bioinformatics/btv009.
Eigen-raw_coding	Eigen score for coding SNVs. A functional prediction score based on conservation, allele frequencies, and deleteriousness prediction using an unsupervised learning method (doi: 10.1038/ng.3477).
Eigen-phred_coding	Eigen score in phred scale.
Eigen-PC-raw_coding	Eigen PC score for genome-wide SNVs. A functional prediction score based on conservation, allele frequencies, deleteriousness prediction (for missense SNVs) and epigenomic signals (for synonymous and non-coding SNVs) using an unsupervised learning method (doi: 10.1038/ng.3477).
Eigen-PC-phred_coding	Eigen PC score in phred scale.

Column	Description
Eigen-PC-raw_coding_rankscore	Eigen-PC-raw scores were ranked among all Eigen-PC-raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of Eigen-PC-raw scores in dbNSFP.
integrated_fitCons_score	fitCons score predicts the fraction of genomic positions belonging to a specific function class (defined by epigenomic fingerprint) that are under selective pressure. Scores range from 0 to 1, with a larger score indicating a higher proportion of nucleic sites of the functional class the genomic position belong to are under selective pressure, therefore more likely to be functional important. Integrated (i6) scores are integrated across three cell types (GM12878, H1-hESC and HUVEC). More details can be found in doi:10.1038/ng.3196.
integrated_fitCons_rankscore	integrated fitCons scores were ranked among all integrated fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of integrated fitCons scores in dbNSFP.
integrated_confidence_value	0 - highly significant scores (approx. p=.25).
GERP++_NR	GERP++ neutral rate
GERP++_RS	GERP++ RS score, the larger the score, the more conserved the site. Scores range from -12.3 to 6.17.
GERP++_RS_rankscore	GERP++ RS scores were ranked among all GERP++ RS scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GERP++ RS scores in dbNSFP.
gnomAD_exomes_AC	Alternative allele count in the whole gnomAD exome samples
gnomAD_exomes_AN	Total allele count in the whole gnomAD exome samples
gnomAD_exomes_AF	Alternative allele frequency in the whole gnomAD exome samples
gnomAD_exomes_AFR_AC	Alternative allele count in the African/African American gnomAD exome samples
gnomAD_exomes_AFR_AN	Total allele count in the African/African American gnomAD exome samples
gnomAD_exomes_AFR_AF	Alternative allele frequency in the African/African American gnomAD exome samples
gnomAD_exomes_AMR_AC	Alternative allele count in the Latino gnomAD exome samples
gnomAD_exomes_AMR_AN	Total allele count in the Latino gnomAD exome samples
gnomAD_exomes_AMR_AF	Alternative allele frequency in the Latino gnomAD exome samples
gnomAD_exomes_ASJ_AC	Alternative allele count in the Ashkenazi Jewish gnomAD exome samples
gnomAD_exomes_ASJ_AN	Total allele count in the Ashkenazi Jewish gnomAD exome samples
gnomAD_exomes_ASJ_AF	Alternative allele frequency in the Ashkenazi Jewish gnomAD exome samples
gnomAD_exomes_EAS_AC	Alternative allele count in the East Asian gnomAD exome samples
gnomAD_exomes_EAS_AN	Total allele count in the East Asian gnomAD exome samples
gnomAD_exomes_EAS_AF	Alternative allele frequency in the East Asian gnomAD exome samples
gnomAD_exomes_FIN_AC	Alternative allele count in the Finnish gnomAD exome samples
gnomAD_exomes_FIN_AN	Total allele count in the Finnish gnomAD exome samples
gnomAD_exomes_FIN_AF	Alternative allele frequency in the Finnish gnomAD exome samples
gnomAD_exomes_NFE_AC	Alternative allele count in the Non-Finnish European gnomAD exome samples
gnomAD_exomes_NFE_AN	Total allele count in the Non-Finnish European gnomAD exome samples
gnomAD_exomes_NFE_AF	Alternative allele frequency in the Non-Finnish European gnomAD exome samples
gnomAD_exomes_SAS_AC	Alternative allele count in the South Asian gnomAD exome samples
gnomAD_exomes_SAS_AN	Total allele count in the South Asian gnomAD exome samples
gnomAD_exomes_SAS_AF	Alternative allele frequency in the South Asian gnomAD exome samples
gnomAD_genomes_AC	Alternative allele count in the whole gnomAD genome samples
gnomAD_genomes_AN	Total allele count in the whole gnomAD genome samples
gnomAD_genomes_AF	Alternative allele frequency in the whole gnomAD genome samples
gnomAD_genomes_AFR_AC	Alternative allele count in the African/African American gnomAD genome samples
gnomAD_genomes_AFR_AN	Total allele count in the African/African American gnomAD genome samples
gnomAD_genomes_AFR_AF	Alternative allele frequency in the African/African American gnomAD genome samples
gnomAD_genomes_AMR_AC	Alternative allele count in the Latino gnomAD genome samples
gnomAD_genomes_AMR_AN	Total allele count in the Latino gnomAD genome samples
gnomAD_genomes_AMR_AF	Alternative allele frequency in the Latino gnomAD genome samples
gnomAD_genomes_ASJ_AC	Alternative allele count in the Ashkenazi Jewish gnomAD genome samples
gnomAD_genomes_ASJ_AN	Total allele count in the Ashkenazi Jewish gnomAD genome samples
gnomAD_genomes_ASJ_AF	Alternative allele frequency in the Ashkenazi Jewish gnomAD genome samples

Column	Description
gnomAD_genomes_EAS_AC	Alternative allele count in the East Asian gnomAD genome samples
gnomAD_genomes_EAS_AN	Total allele count in the East Asian gnomAD genome samples
gnomAD_genomes_EAS_AF	Alternative allele frequency in the East Asian gnomAD genome samples
gnomAD_genomes_FIN_AC	Alternative allele count in the Finnish gnomAD genome samples
gnomAD_genomes_FIN_AN	Total allele count in the Finnish gnomAD genome samples
gnomAD_genomes_FIN_AF	Alternative allele frequency in the Finnish gnomAD genome samples
gnomAD_genomes_NFE_AC	Alternative allele count in the Non-Finnish European gnomAD genome samples
gnomAD_genomes_NFE_AN	Total allele count in the Non-Finnish European gnomAD genome samples
gnomAD_genomes_NFE_AF	Alternative allele frequency in the Non-Finnish European gnomAD genome samples
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 ','.
GTEx_V8_gene	target gene of the (significant) eQTL SNP
GTEx_V8_tissue	tissue type of the expression data with which the eQTL/gene pair is detected
MIM_id	MIM gene id (from HGNC)
Gene_full_name	Gene full name (from HGNC)
Pathway(Uniprot)	Pathway description from Uniprot
Pathway(BioCarta)_short	Short name of the Pathway(s) the gene belongs to (from BioCarta)
Pathway(BioCarta)_full	Full name(s) of the Pathway(s) the gene belongs to (from BioCarta)
Pathway(ConsensusPathDB)	Pathway(s) the gene belongs to (from ConsensusPathDB)
Pathway(KEGG)_id	ID(s) of the Pathway(s) the gene belongs to (from KEGG)
Pathway(KEGG)_full	Full name(s) of the Pathway(s) the gene belongs to (from KEGG)
Function_description	Function description of the gene (from Uniprot)
Disease_description	Disease(s) the gene caused or associated with (from Uniprot)
MIM_phenotype_id	MIM id(s) of the phenotype the gene caused or associated with (from Uniprot)
MIM_disease	MIM disease name(s) with MIM id(s) in '[]' (from Uniprot)
Trait_association(GWAS)	Trait(s) the gene associated with (from GWAS catalog)
GO_biological_process	GO terms for biological process
GO_cellular_component	GO terms for cellular component
GO_molecular_function	GO terms for molecular function
Tissue_specificity(Uniprot)	Tissue specificity description from Uniprot
Expression(egenetics)	Tissues/organs the gene expressed in (egenetics data from BioMart)
Expression(GNF/Atlas)	Tissues/organs the gene expressed in (GNF/Atlas data from BioMart)
Interactions(IntAct)	Other genes (separated by ;) genes this gene interacting with (from IntAct). Full information (gene name followed by Pubmed id in '[]') can be found in the '.complete' table
Interactions(BioGRID)	Other genes (separated by ;) this gene interacting with (from BioGRID) Full information (gene name followed by Pubmed id in '[]') can be found in the '.complete' table
Interactions(ConsensusPathDB)	Other genes (separated by ;) this gene interacting with (from ConsensusPathDB). Full information (gene name followed by Pubmed id in '[]') can be found in the '.complete' table
P(HI)	Estimated probability of haploinsufficiency of the gene (from doi:10.1371/journal.pgen.1001154)
P(rec)	Estimated probability that gene is a recessive disease gene (from DOI:10.1126/science.1215040)
Known_rec_info	Known recessive status of the gene (from DOI:10.1126/science.1215040) lof-tolerant = seen in homozygous state in at least one 1000G individual recessive = known OMIM recessive disease (original annotations from DOI:10.1126/science.1215040)
RVIS_EVS	Residual Variation Intolerance Score, a measure of intolerance of mutational burden, the higher the score the more tolerant to mutational burden the gene is. Based on EVS (ESP6500) data. from doi:10.1371/journal.pgen.1003709
RVIS_percentile_EVS	The percentile rank of the gene based on RVIS, the higher the percentile the more tolerant to mutational burden the gene is. Based on EVS (ESP6500) data.
LoF-FDR_ExAC	'A gene's corresponding FDR p-value for preferential LoF depletion among the ExAC population. Lower FDR corresponds with genes that are increasingly depleted of LoF variants.' cited from RVIS document.
RVIS_ExAC	'ExAC-based RVIS; setting 'common' MAF filter at 0.05% in at least one of the six individual ethnic strata from ExAC.' cited from RVIS document.

Column	Description
RVIS_percentile_ExAC	'Genome-Wide percentile for the new ExAC-based RVIS; setting 'common' MAF filter at 0.05% in at least one of the six individual ethnic strata from ExAC.' cited from RVIS document.
GHIS	A score predicting the gene haploinsufficiency. The higher the score the more likely the gene is haploinsufficient. (from doi: 10.1093/nar/gkv474)
GDI	gene damage index score, 'a genome-wide, gene-level metric of the mutational damage that has accumulated in the general population' from doi: 10.1073/pnas.1518646112. The higher the score the less likely the gene is to be responsible for monogenic diseases.
GDI-Phred	Phred-scaled GDI scores
Gene_damage_prediction(all_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for all diseases
Gene_damage_prediction(all_Mendelian_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for all Mendelian diseases
Gene_damage_prediction(Mendelian_AD_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for Mendelian autosomal dominant diseases
Gene_damage_prediction(Mendelian_AR_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for Mendelian autosomal recessive diseases
Gene_damage_prediction(all_PID_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for all primary immunodeficiency diseases
Gene_damage_prediction(PID_AD_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for primary immunodeficiency autosomal dominant diseases
Gene_damage_prediction(PID_AR_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for primary immunodeficiency autosomal recessive diseases
Gene_damage_prediction(all_cancer_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for all cancer disease
Gene_damage_prediction(cancer_recessive_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for cancer recessive disease
Gene_damage_prediction(cancer_dominant_disease-causing_genes)	gene damage prediction (low/medium/high) by GDI for cancer dominant disease

Analysis Tools

BWA (Burrows-Wheeler Alignment Tool)

bwa-0.7.17

BWA is a software package for mapping low-divergent sequences to a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two are for longer sequences ranging from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment. However, BWA-MEM, the latest of all, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.
For all the algorithms, BWA first needs to construct the FM-index for the reference genome (the index command). Alignment algorithms are invoked with different sub-commands: aln/samse/sampe for BWA-backtrack, bwasw for BWA-SW and mem for the BWA-MEM algorithm.

More information can be found here:

http://bio-bwa.sourceforge.net/bwa.shtml

Picard

picard-tools-Version:3.1.1

Picard is a collection of Java-based command-line utilities that manipulate SAM files, and a Java API (SAM-JDK) for creating new programs that read and write SAM files. Both SAM text format and SAM binary (BAM) format are supported. Picard MarkDuplicates examines aligned records in the supplied SAM or BAM file to locate duplicate molecules. All records are then written to the output file with the duplicate records flagged.

More information can be found here:

http://broadinstitute.github.io/picard/

GATK (Genome Analysis Toolkit)

GATKv4.5.0.0

The Genome Analysis Toolkit or GATK is a software package developed at the Broad Institute to analyze high-throughput sequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.
HaplotypeCaller calls SNPs and indels simultaneously via local re-assembly of haplotypes in an active region.

More information can be found here:

https://www.broadinstitute.org/gatk/

SnpEff

SnpEff 5.2 2023-09-29

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes).

SnpEff can generate the following results :

- Genes and transcripts affected by the variant
- Location of the variants
- How the variant affects the protein synthesis (e.g. generating a stop codon)
- Comparison with other databases to find equal known variants

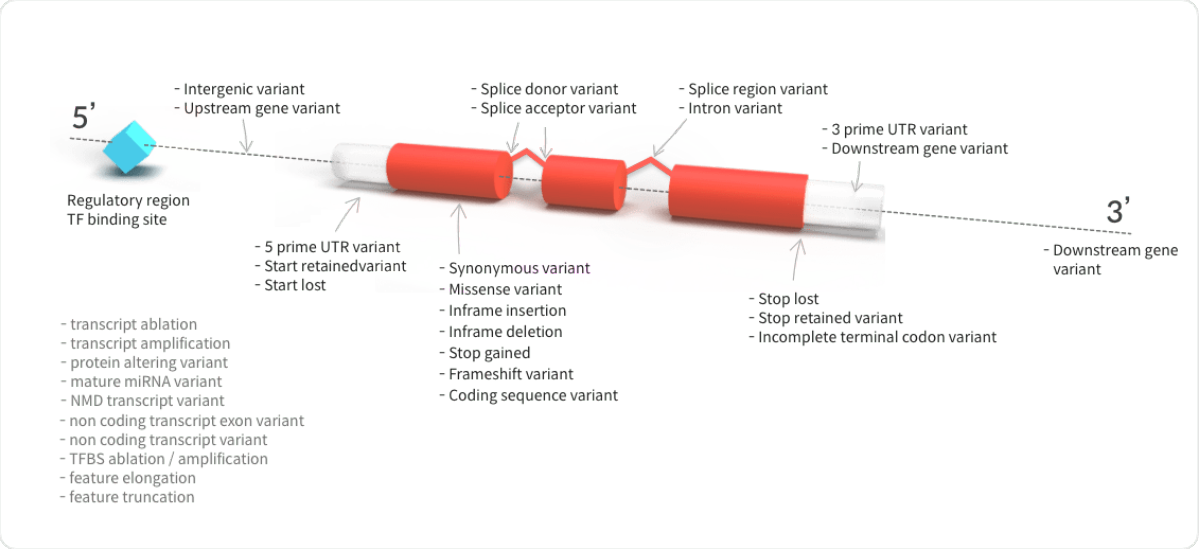
More information can be found here:

http://snpeff.sourceforge.net/SnpEff.html

Analysis Database

Effect (Sequence Ontology)

Sequence ontology (SO) allows to standardize terminology used for assessing sequence changes and impact. This allows for a common language across all variant annotation programs and makes it easier to communicate using a uniform terminology. Starting from version 4.0 VCF output uses SO terms by default. See below for the location of each display term relative to the transcript structure:



- The terms in the table below are shown in order of severity (more severe to less severe) as estimated by SnpEff.

SO Table

SO Term	SO Description	SO Accession
chromosome	Structural unit composed of a nucleic acid molecule which controls its own replication through the interaction of specific proteins at one or more origins of replication.	SO:0000340
chromosome_number_variation	A kind of chromosome variation where the chromosome complement is not an exact multiple of the haploid number.	SO:1000182
gene_fusion	A sequence variant whereby a two genes have become joined.	SO:0001565
bidirectional_gene_fusion	Fusion of two genes in opposite directions.	SO:0002086
duplication	An insertion which derives from, or is identical in sequence to, nucleotides present at a known location in the genome.	SO:1000035
feature_ablation	A sequence variant, caused by an alteration of the genomic sequence, where the deletion, is greater than the extent of the underlying genomic features.	SO:0001879
inversion	Inversion of a large chromosome segment (over 1% or 1,000,000 bases).	SO:1000036
protein_protein_contact	A binding site that, in the protein molecule, interacts selectively and non-covalently with polypeptide residues.	SO:0001093
rearranged_at_DNA_level	An attribute to describe the sequence of a feature, where the DNA is rearranged.	SO:0000904
structural_interaction_variant	A variant that impacts the internal interactions of the resulting polypeptide structure.	SO:0002093
exon_loss_variant	A sequence variant whereby an exon is lost from the transcript.	SO:0001572
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three.	SO:0001589
stop_gained	Variant causes a STOP codon. e.g.: Cag/Tag, Q/*	SO:0001587
stop_lost	Variant causes stop codon to be mutated into a non-stop codon. e.g.: Tga/Cga, */R	SO:0001578
start_lost	Variant causes start codon to be mutated into a non-start codon. e.g.: aTg/aGg, M/R	SO:0002012
splice_acceptor_variant	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	SO:0001574
splice_donor_variant	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	SO:0001575
rare_amino_acid_variant	A sequence variant whereby at least one base of a codon encoding a rare amino acid is changed, resulting in a different encoded amino acid.	SO:0002008
missense_variant	Variant causes a codon that produces a different amino acid. e.g.: Tgg/Cgg, W/R	SO:0001583
disruptive_inframe_insertion	One codon is changed and one or many codons are inserted. e.g.: An insert of size multiple of three, not at codon boundary	SO:0001824
conservative_inframe_insertion	An inframe increase in cds length that inserts one or more codons into the coding sequence between existing codons.	SO:0001823
disruptive_inframe_deletion	One codon is changed and one or more codons are deleted. e.g.: A deletion of size multiple of three, not at codon boundary	SO:0001826
conservative_inframe_deletion	An inframe decrease in cds length that deletes one or more entire codons from the coding sequence but does not change any remaining codons.	SO:0001825
5_prime_UTR_truncation	A sequence variant that causes the reduction of a the 5'UTR with regard to the reference sequence.	SO:0002013
3_prime_UTR_truncation	A sequence variant that causes the reduction of a the 3'UTR with regard to the reference sequence.	SO:0002015
splice_branch_variant	A splice branch variant is a genetic change that affects the location of the branch point, the site where the intron is cut during splicing. This type of variant can alter the splicing pattern of a gene, leading to the production of an abnormal protein or the loss of protein production.	none
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.	SO:0001630
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon (the new codon produces a different AA).	SO:0001567
initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript.	SO:0001582
non_canonical_start_codon	A start codon that is not the usual AUG sequence.	SO:0000680
synonymous_variant	Variant causes a codon that produces the same amino acid. e.g.: Ttg/Ctg, L/L	SO:0001819
coding_sequence_variant	The variant hits a CDS.	SO:0001580
5_prime_UTR_premature_start_codon_gain_variant	A 5'UTR variant where a premature start codon is gained.	SO:0001988
start_retained_variant	Variant causes start codon to be mutated into another start codon. e.g.: Ttg/Ctg, L/L (TTG and CTG can be START codons)	SO:0002019

SO Term	SO Description	SO Accession
5_prime_UTR_variant	Variant hits 5'UTR region.	SO:0001623
3_prime_UTR_variant	Variant hits 3'UTR region.	SO:0001624
upstream_gene_variant	A sequence variant located 5' of a gene.	SO:0001631
downstream_gene_variant	A sequence variant located 3' of a gene.	SO:0001632
TF_binding_site_variant	A sequence variant located within a transcription factor binding site.	SO:0001782
regulatory_region_variant	A sequence variant located within a regulatory region (non-coding).	SO:0001566
miRNA	Small, ~22-nt, RNA molecule that is the endogenous transcript of a miRNA gene or the product of other non coding RNA genes. Micro RNAs are produced from precursor molecules (SO:0001244) that can form local hairpin structures, which ordinarily are processed (usually via the Dicer pathway) such that a single miRNA molecule accumulates from one arm of a hairpin precursor molecule. Micro RNAs may trigger the cleavage of their target molecules or act as translational repressors.	SO:0000276
custom	Custom in variant priority referes to user-defined variants, which are variants found in a user-created genetic database and can be classified using custom classification using custom classification rules defined by the user.	none
sequence_feature	Any extent of continuous biological sequence.	SO:0000110
conserved_intron_variant	A transcript variant occurring within a conserved region of an intron.	SO:0002018
intron_variant	A transcript variant occurring within an intron.	SO:0001627
intragenic_variant	A variant that occurs within a gene but falls outside of all transcript features. This occurs when alternate transcripts of a gene do not share overlapping sequence.	SO:0002011
conserved_intergenic_variant	A sequence variant located in a conserved intergenic region, between genes.	SO:0002017
intergenic_region	A region containing or overlapping no genes that is bounded on either side by a gene, or bounded by a gene and the end of the chromosome.	SO:0000605
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript.	SO:0001792
exon_variant	A sequence variant that changes exon sequence.	SO:0001791
non_coding_transcript_variant	A transcript variant of a non coding RNA gene.	SO:0001619
gene_variant	A sequence variant where the structure of the gene is changed.	SO:0001564
transcript_variant	A sequence variant that changes the structure of the transcript.	SO:0001576

CLINVAR

ClinVar is a freely accessible, data archive of reports of the relationships among human variations and phenotypes hosted by the National Center for Biotechnology Information (NCBI) and funded by intramural National Institutes of Health (NIH) funding.

ESP (Exome Sequencing Project)

The ESP is a NHLBI funded exome sequencing project aiming to identify genetic variants in exonic regions from over 6000 individuals, including healthy ones as well as subjects with different diseases. The variant call data set is constantly being updated. As the size of the database is more than 1000 Genomes Project and the fold coverage is far higher, this data set will be particularly useful for users with exome sequencing data sets. As of October 2012, esp5400 and esp6500 are available, representing summary statistics from 5400 exomes and 6500 exomes, respectively. As of February 2013, the most recent version of ESP is esp6500si, so whenever possible, users should use this database for annotation. Compared to esp6500, the esp6500si contains more calls, and indel calls and chrY calls.

SIFT

SIFT(S orting I ntolerant F orm T olerant) predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids. The data provide for each amino acid substitution is a score and a qualitative prediction (either 'tolerated' or 'deleterious'). The score is the normalized probability that the amino acid change is tolerated so scores nearer to 0 are more likely to be deleterious. The qualitative prediction is derived from this score such that substitutions with a score < 0.05 are called 'deleterious' and all others are called 'tolerated'.
Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm Nature Protocols 4(8):1073-1081 (2009)

More information can be found here:

https://www.nature.com/articles/nprot.2009.86

dbNSFP

Macrogen carries out the functional annotation of variants using dbNSFP database, currently. dbNSFP is a database developed for functional annotation of non-synonymous single-nucleotide variants, which can determine whether a certain variant causes change at the amino-acid level or whether it has damaging effect. dbNSFP can provide not only the frequency Information such as 1000 Genomes Project, The Exome Aggregation Consortium (ExAC), but prediction scores are calculated through prediction algorithms like SIFT, MutationTaster, PROVEAN. And also, it shows conservation scores from related databases such as GERP++, SiPhy.

Research use only

HEADQUARTER

Macrogen Gangnam HQ

Business & Support Center

Macrogen Bldg, 238, Teheran-ro,
Gangnam-gu, Seoul, Republic of Korea
Tel: +82-2-2180-7000
Web: www.macrogen.com
LIMS: dna.macrogen.com

Macrogen Genome Center

Laboratory & IT Center

[08511] 1001, 10F, 254, Beotkkot-ro,
Geumcheon-gu, Seoul, Republic of Korea
(Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com
(Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

**Laboratory,
Business & Support Center**

Meibergdreef 57, 1105 BA, Amsterdam,
the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

**Laboratory,
Business & Support Center**

3 Biopolis Drive #05-18, Synapse,
Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Psomagen (Macrogen USA)

**Laboratory,
Business & Support Center**

1330 Piccard Drive, Suite 103, Rockville,
MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

**Laboratory,
Business & Support Center**

16F Time24 Building, 2-4-32 Aomi,
Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

**Laboratory,
Business & Support Center**

Av. Sur del Aeropuerto de Barajas,
28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Macrogen Belgium

**Laboratory,
Business & Support Center**

Oxfordlaan 70, 6229 EV Maastricht,
Netherlands
Tel: +31-20-333-7563
Email: info.be@macrogen.eu

Macrogen Italy

**Laboratory,
Business & Support Center**

Viale Ortles, 22/4, 20139 Milano,
MI, Italy
Tel: +39-02-5666-0274
Email: italy@macrogen-europe.com

Humanizing Genomics
macrogen