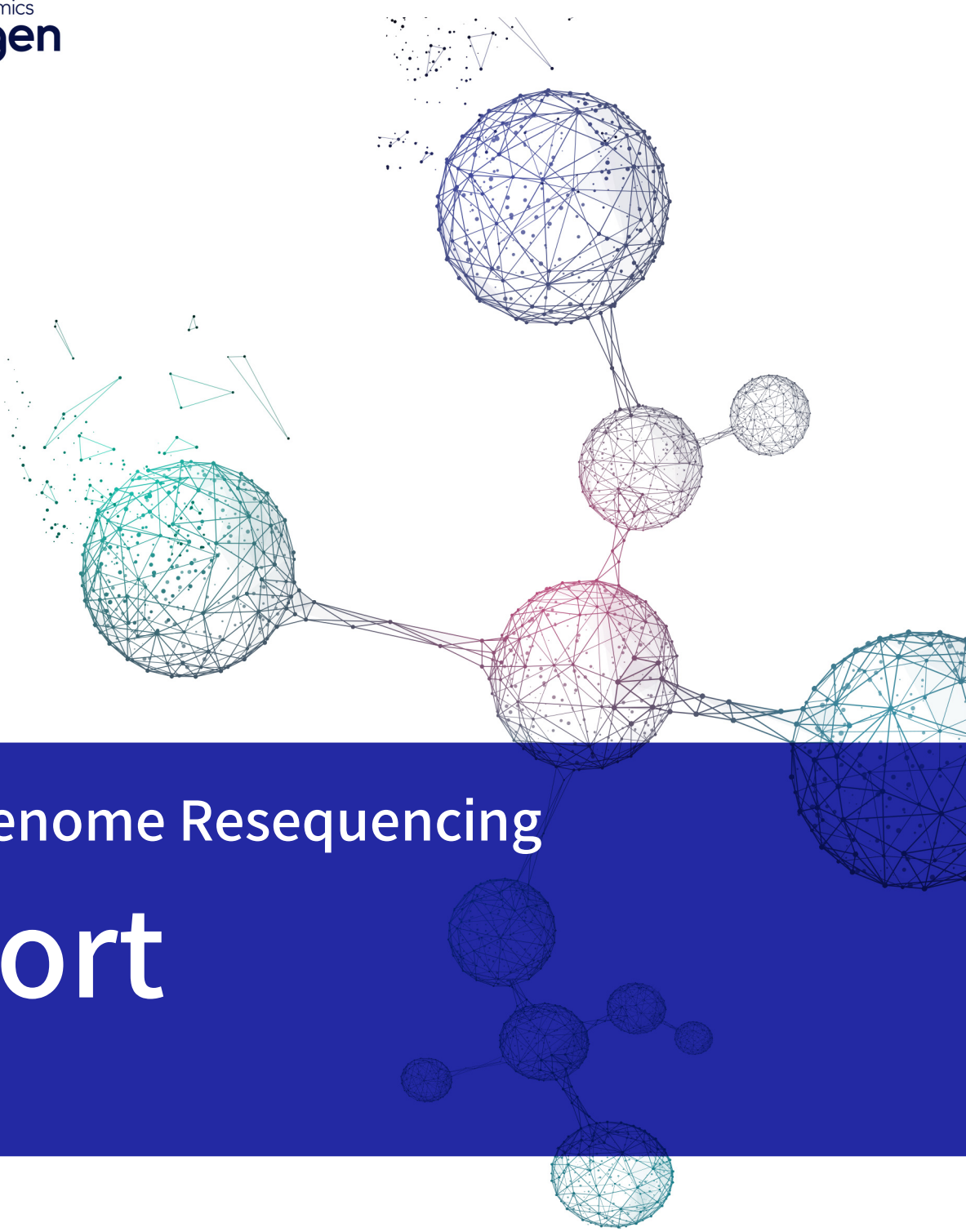
Humanizing Genomics
macrogen



Whole Genome Resequencing

Report

2022.12

WGS

Research use only

Table of Contents

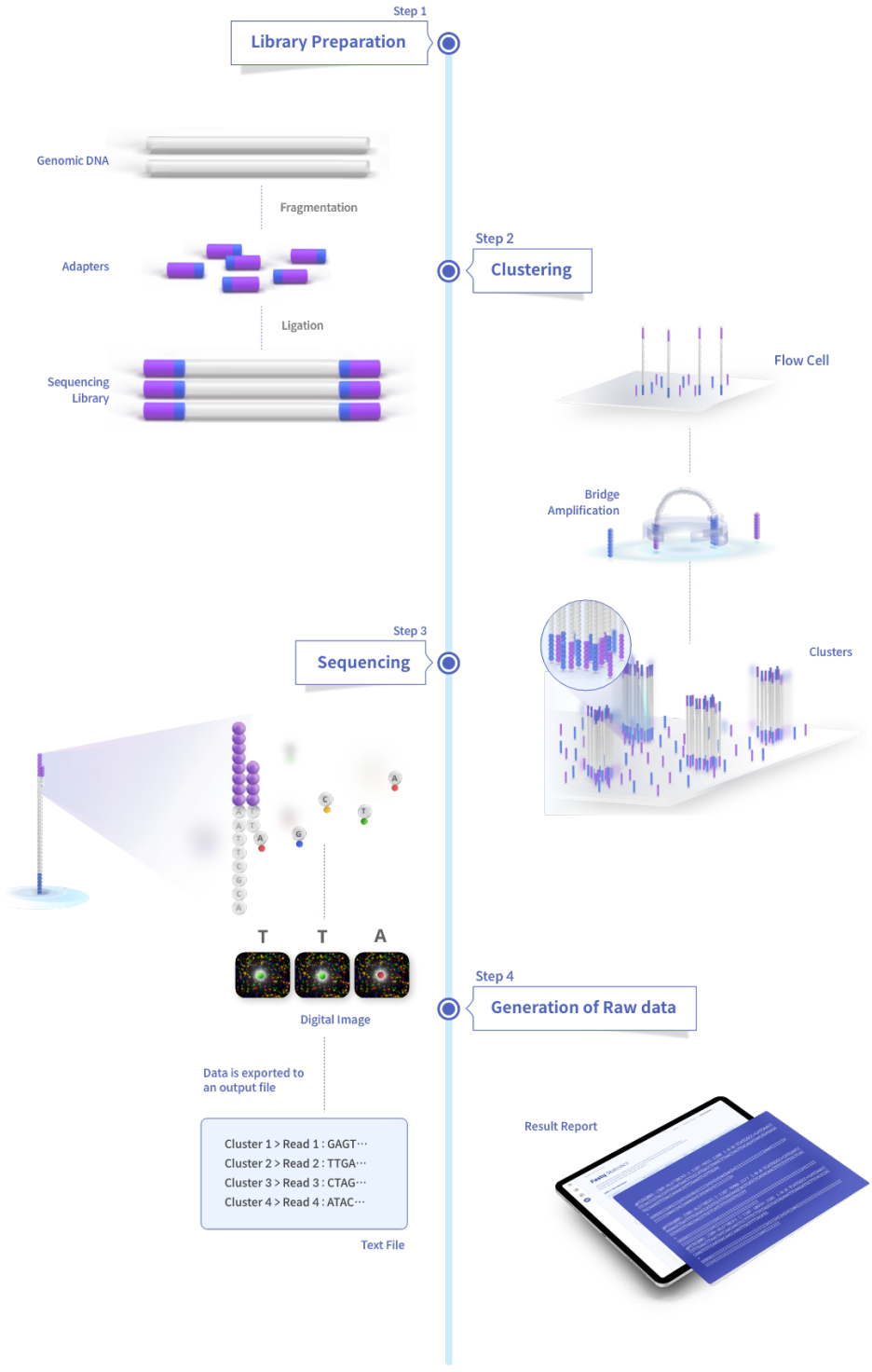
Order Information

Client Name	Macrogen
Client Organization	Macrogen HQ
Order Number	HN00998989
Application	Whole Genome Resequencing
Type Of Read	Paired-end
Read Length	151

Workflow

Experiments Workflow

The samples are prepared according to NGS library preparation workflow, and sequenced using Illumina platform. The workflow illustrated below shows the common ligation based method of library preparation. The process may differ based on the library preparation protocol followed.



Library Construction

• DNA fragmentation : Each sequenced sample is prepared according to the Illumina TruSeq DNA sample preparation guide to obtain a final library of 300-400 bp average insert size. 1 μg (TruSeq DNA PCR-free library) or 100 ng (TruSeq Nano DNA library) of genomic DNA is fragmented by covaris systems, which generates dsDNA fragments with 3' or 5' overhangs.
• End repair and size selection : The dsDNA fragments with 3' or 5' overhangs are converted into blunt ends using an end repair mix. The 3' to 5' exonuclease removes the 3' overhangs, and the polymerase fills in the 5' overhangs. Following the end repair, the appropriate library size is selected using different ratios of the sample purification beads.
• Adenylation of 3' end : A single 'A' nucleotide is added to the 3' ends of the blunted fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment.
• Adapters ligation : Multiple indexing adapters are ligated to the ends of the DNA fragments to prepare them for hybridization onto a flow cell.
• DNA fragments enrichment (TruSeq Nano DNA library only) : PCR is used to amplify the enriched DNA library for sequencing. The PCR is performed with a PCR primer solution that anneals to the ends of each adapters.
• Library validation : Macrogen performs quality control analysis on the sample library and quantification of the DNA library templates.

Clustering & Sequencing

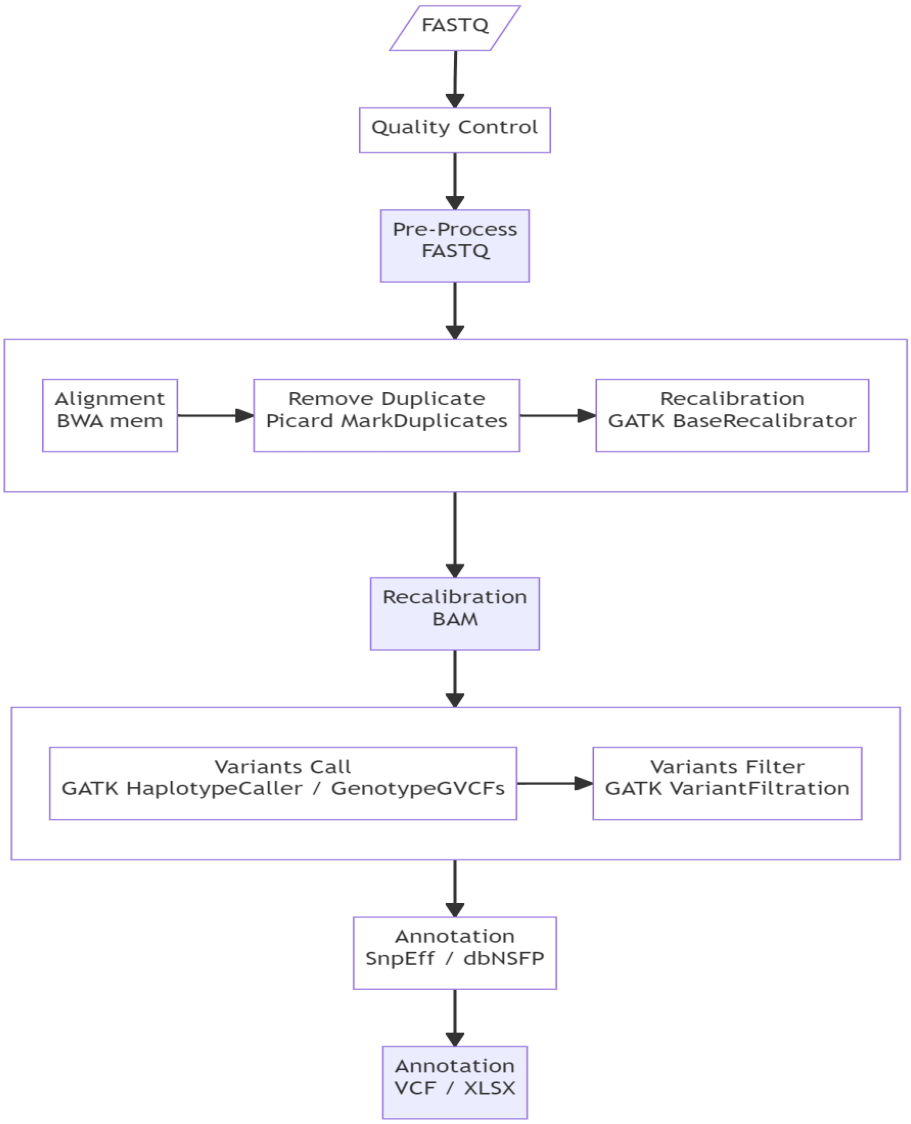
For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing. Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4 reversible, terminator-bound dNTPs are persent during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive sequence regions and homopolymers.

Generation of Raw data

The Illumina Platform generates raw images and base calling through an integrated primary analysis software called RTA (real time analysis). The BCL/cBCL (base calls) binary is converted into FASTQ using illumina package bcl2fastq2-v2.20.0. The demultiplexing option (--barcode-mismatches) was set to perfect match (value : 0).

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 HiSeq 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	138, 154
1000Genome	1000 Genome Phase3
Clinvar	07/2021
ESP	ESP6500SI_V2
dbNSFP	dbNSFPv4.2c

Tool Version

Software	Version
BWA	bwa-0.7.17
Picard	picard-tools-2.18.2-SNAPSHOT
GATK	GATKv4.0.5.1
SnpEff	SnpEff 5.0e

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	false	
Picard	AS	true	Assume Sorted
Picard	CREATIVE_INDEX	true	Creative 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	Combines VCF records from different sources
GATK	-knownSites	1000G_phase1.indels.hg38.vcf	Database of known polymorphic sites
GATK	-knownSites	dbSNP154.vcf.gz	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%
SampleA	115,003,722,456	761,614,056	40.95	59.05	96.97	92.18
SampleB	114,087,334,562	755,545,262	40.91	59.09	96.92	92.05

- 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

	SampleA	SampleB
Reference size(Mbp)	2,934	2,934
Throughput mean depth (X)	39.18	38.87
De-duplicated reads	653,247,843	644,728,535
De-duplicated reads %(out of total reads)	85.77	85.33
Mappable reads(reads mapped to reference)	652,050,177	643,538,152
Mappable reads %(out of de-duplicated reads)	99.81	99.81
Mappable mean depth (X)	33.17	32.74
% >= 1x coverage	99.8	99.8
% >= 5x coverage	99.4	99.4
% >= 10x coverage	98.8	98.7
% >= 15x coverage	96.6	96.4
% >= 20x coverage	92.7	92.3
% >= 30x coverage	65.1	62.7
SNPs	4,083,387	4,080,227
Smallinsertions	203	206
Smalldeletions	232	223
Synonymouscodingvariants	12,883	12,901
Non-synonymouscodingvariants	12,917	12,952
Splicing variants	4,322	4,324
Stop gained	153	152
Stop lost	29	27
Frameshift	343	350
% found indbSNP138	88.1	88.1
% found indbSNP154	88.3	88.4
het/homratio	1.4	1.4
Ts/Tvratio	1.9	1.9
Copy number gains(>2)	703	633
Copy number losses(<2)	209	193
Duplications	712	720
Insertions	3,094	3,037
Deletions	5,068	5,114
Inversions	349	331
Translocations	1,282	1,222

- Sample ID : Sample name.
- Reference size(Mbp) : Reference size.
- Throughput mean depth(X) : {Total yield} / {Reference size}.
- De-duplicated reads : Number of de-duplicated reads.
- De-duplicated reads %(out of total reads) : 100 * {Number of de-duplicated reads} / {Total reads}.
- Mappable reads(reads mapped to reference) : Number of mappable reads.
- Mappable reads %(out of de-duplicated reads) : 100 * {Number of mappable reads} / {Number of de-duplicated reads}.
- Mappable mean depth(X) : {Mappable yield} / {Reference size}.
- % Coverage : The percentage of bases in target regions with a depth of coverage or greater.
- SNPs : The number of Single-nucleotide polymorphism
- Synonymous variants : Variant causes a codon that produces the same amino acid e.g.: Ttg/Ctg, L/L.
- Non-synonymous variants : Variant causes a codon that produces a different amino acid e.g.: Tgg/Cgg, W/R.
- Splicing variants : The variant hits a splice acceptor and donor site.
- 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.
- Frameshift : Insertion or deletion causes a frame shift e.g.: An indel size is not multple of 3.
- % found in dbSNP138 : rs number.
- % found in dbSNP154 : rs number.
- Het/hom ratio : Ratio of number of heterozygous variants to number of homozygous variants.
- Ts/Tv ration : Ratio of 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).
- Copy number gains (>2) : CN > 2
- Copy number losses (<2) : CN < 2
- Duplications : a section of DNA is duplicated and both copies end up in the same chromosome
- Insertions : extra base pairs are inserted into DNA sequence
- Deletions : a section of DNA is lost, or deleted
- Inversions : a section of DNA is put in backwards
- Translocations : two non-homologous chromosomes exchange sections of DNA

Deliverables

Analysis Result

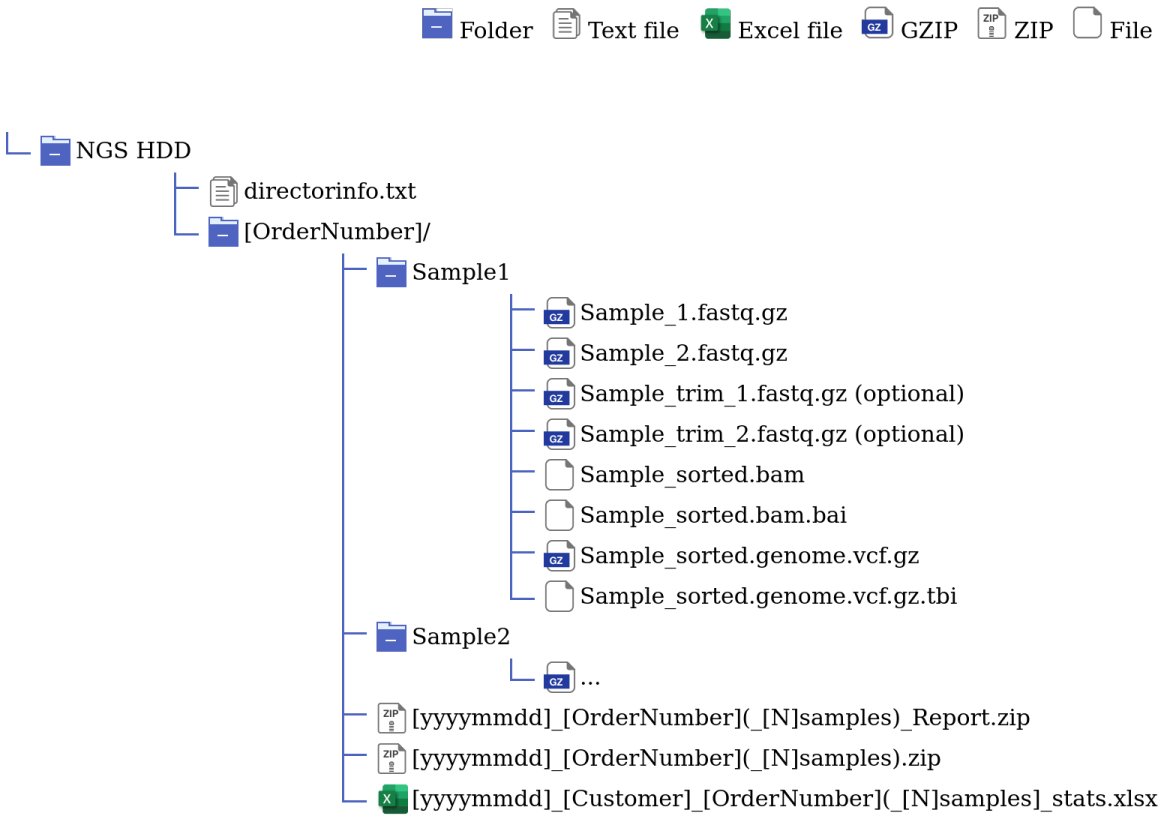
Your data will be retained in our server for 3 months.
Should you wish to extend the retention period, please contact us.

File name	File Size	md5sum
SampleA_R1.fastq.gz	25G	aa25d0bcfa5c1218d67a7d4ad6ee8de7
SampleA_R2.fastq.gz	26G	2c97babf50b53307dadacd6be842f0a3
SampleA_sorted.bam	107G	dd4321a8a560b38e73641958e48919f7
SampleA_sorted.bam.bai	8.6M	7c0cbce35542227b2fedcdc3f0f2c1de
SampleA_sorted.genome.vcf.gz	6.2G	0642bb2e90c99cfca0e242a773ab813a
SampleA_sorted.genome.vcf.gz.tbi	1.3M	7af8aa62ece136ae5f7b8fa02c759604
SampleB_R1.fastq.gz	25G	3c0e5bc5dc71e4d54e9d519561752d87
SampleB_R2.fastq.gz	26G	3ad702c658fd0b00d66f63825fe1684f
SampleB_sorted.bam	107G	680f84bf31f8f21dbd95ab3d3a469f6b
SampleB_sorted.bam.bai	8.6M	7515b37702a145782e40d18e92c1012b
SampleB_sorted.genome.vcf.gz	6.2G	ca3871ec91e8fe5dc38a9d7a3595a3bb
SampleB_sorted.genome.vcf.gz.tbi	1.3M	b66e179ca3aebf9aeda3f79392d7f140
HN00998989_2samples_md5sum.xlsx	1.3M	b66e179ca3aebf9aeda3f79392d7f140
221222_Macrogen_HN00998989_sample.zip	10M	b66e179ca3aebf9aeda3f79392d7f140

Download Data

- md5sum : 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.

Analysis Result Tree



Result File Description

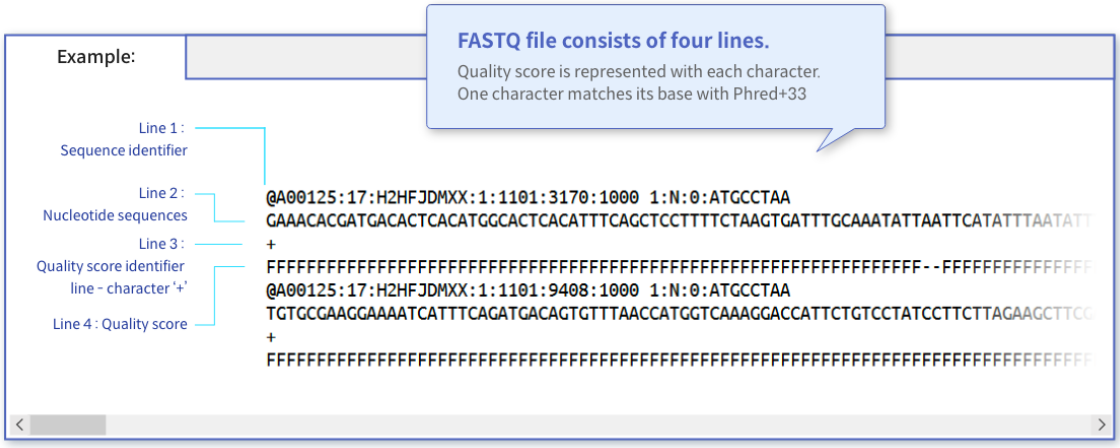
Deliverables List

File Type	File Name	Description
Raw Data	Sample_R1.fastq.gz	Raw read1 sequence data
Raw Data	Sample_R2.fastq.gz	Raw read2 sequence data
Alignment file	Sample_sorted.bam	BWA alignment file
Alignment file	Sample_sorted.bam.bai	BWA alignment index file
SNP/INDEL Result	Sample_SNP_INDEL.vcf	SNP/INDEL file (vcf format)
SNP/INDEL Result	Sample_[chr*].xlsx	Convert SNP_INDEL result (excel file)
SNP/INDEL Result	Sample_sorted.genome.vcf.gz	Genomic VCF
SNP/INDEL Result	Sample_sorted.genome.vcf.gz.tbi	Genomic VCF index file
CNV Result	Sample_CNVs.xlsx	CNV result file (excel file)
SV Result	Sample_SV.vcf	SV result file (vcf format)
SV Result	Sample_SV.xlsx	Convert SV result (excel file)
Summary	[yyyymmdd]_[Customer]_[OrderNumber#]_[#samples]_stats.xlsx	Analysis stats report of all samples (excel file)
md5sum	[Ordernumber#]_[#samples]_md5sum.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 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 = -10 log10(error rate) Encoding : ASCII Character Code=Phred Quality Value + 33

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

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

- The quality score table above is typically updated when significant characteristics of the sequencing platform change, such as new hardware, software, or chemistry versions.

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

◉ 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.

Example:

```
@HD	VN:1.5	SO:coordinate
@SQ	SN:chrM	LN:16569
@SQ	SN:chr1	LN:248956422
@SQ	SN:chr2	LN:242193529
@SQ	SN:chr3	LN:198295559
@SQ	SN:chr4	LN:190214555
@SQ	SN:chr5	LN:181538259
@SQ	SN:chr6	LN:170805979
@SQ	SN:chr7	LN:159345973
@SQ	SN:chr8	LN:145138636
@SQ	SN:chr9	LN:138394717
@SQ	SN:chr10	LN:133797422
@SQ	SN:chr11	LN:135086622
@SQ	SN:chr12	LN:133275309
@SQ	SN:chr13	LN:114364328
@SQ	SN:chr14	LN:107043718
@SQ	SN:chr15	LN:101991189
@SQ	SN:chr16	LN:90338345
@SQ	SN:chr17	LN:83257441
@SQ	SN:chr18	LN:80373285
@SQ	SN:chr19	LN:58617616
@SQ	SN:chr20	LN:64444167
@SQ	SN:chr21	LN:46709983
@SQ	SN:chr22	LN:50818468
@SQ	SN:chrX	LN:156040895
@RG	SN:chrY	LN:57227415
@PG	ID:A00125	PL:Illumina	SM:Sample LB:SS6
@PG	ID:bwa	PN:bwa	VN:0.7.17-r1188	CL:bwa mem
@PG	ID:MarkDuplicates	VN:2.18.2-SNAPSHOT	CL:MarkDuplicates	PN:MarkDuplicates
@PG	ID:GATK ApplyBQSR	VN:4.0.5.1	CL:ApplyBQSR	PN:GATK ApplyBQSR
A00125:511:HFVL5DSX2:3:1167:10176:32142	83	chr1	826845	40	101M	=	826766
-180	CCCACCTCGACATCCACAGCGAGGCAATGAAGAAGCCCCTGCCAAGGAAGAGCCCGCTTCTCAGTGGGACACCGGGAAGGTAGACACCCAACAGTCACCGC	EEH<FG@@C<H=@FH<H
<FAD?FFI<>?D<?D<?GHHHI@GHJ=@GE=@E@GHHBGICBIBJ@F@GGE>J>HBGGE=@GF?@E>J>HHJ=>J@FBJ>ICHB	XA:Z:chr8,+232954,101M,1;	MC:Z:101M MD:Z:48G52
PG:Z:MarkDuplicates	RG:Z:A00125	NM:i:1	AS:i:96	XS:i:96
A00125:511:HFVL5DSX2:3:1650:26829:19492	83	chr1	826845	27	101M	=	826807
-139	CCCACCTCGACATCCACAGCGAGGCAATGAAGAAGCCCCTGCCAAGGAAGAGCCCGCTTCTCAGTGGGACACCGGGAAGGTAGACACCCAACAGTCACCGC	EEH<FG@@C<H=@FH<H
>FAD?FFI<>?D<?D<?GHHHI@GHJ=@GE=@E@GHHBGICBIBJ@F@GGE>J>HBGGE=@GF?@E>J>HHJ=>J@FBJ>ICHB	XA:Z:chr8,+232954,101M,1;	MC:Z:101M MD:Z:48G52
PG:Z:MarkDuplicates	RG:Z:A00125	NM:i:1	AS:i:96	XS:i:96
```

Header section

Alignment section

More information about BAM format can be found here:

https://samtools.github.io/hts-specs/SAMv1.pdf

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.

SAM file : Alignments section mandatory fields

Field	Regexp / Range	Description
QNAME	[!-?A-~]{1,255}	Query template NAME
FLAG	[0,2^16-1]	Bitwise FLAG
RNAME	*|[!-()+---~][!-~]*	Reference sequence NAME
POS	[0,2^31-1]	1-based leftmost mapping POSition
MAPQ	[0,2^8-1]	MAPping Quality
CIGAR	*|([0-9]+[MIDNSHPX=])+	CIGAR string
RNEXT	*|=|[!-()+---~][!-~]*	Ref. name of the mate/NEXT read
PNEXT	[0,2^31-1]	Position of the mate/NEXT read
TLEN	[-2^31+1,2^31-1]	Observed Template LENgth
SEQ	*|[A-Za-z=.]+	segment SEQuence
QUAL	[!-~]+	ASCII of Phred-scaled base QUALity+33

◉ 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.

Appendix

Annotation Column

The [Sample]_[chr*].xlsx file contains information about variants found at specific positions in the reference genome. Each data line contains information about a single variant. The [Sample]_SV.xlsx file includes structural variants (SVs) infromation such as the variant location on the reference genome, the type of variant and the predicted effect of the variant event. Each data line describes each SV event.

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
REF_DP	Allelic depths for the ref alleles
ALT_DP	Allelic depths for the alt alleles. For indels this value only includes reads which confidently support each allele.(posterior prob 0.51 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

Column	Description
ALT_AA	alternative amino acid
cDNA_pos	Position in cDNA (one based)
cDNA_length	Trancript'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.
dbSNP154_ID	dbSNP154 rsNo.
KGphase3_AF	Non-reference allele frequency of existing variation in 1000 Genomes
KGphase3_AFR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined African population
KGphase3_AMR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined American population
KGphase3_EAS_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined East Asian population
KGphase3_EUR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined European population
KGphase3_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

Column	Description
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 v2.0 Disease	Disease when ACMG gene matches MIM ID
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)
hg18_chr	chromosome as to hg18, '.' means missing
hg18_pos(1-based)	physical position on the chromosome as to hg18 (1-based coordinate) For mitochondrial SNV, this position refers to a YRI sequence (GenBank: AF347015)
genename	gene name; if the nsSNV can be assigned to multiple genes, gene names are
cds_strand	separated by ';'
refcodon	coding sequence (CDS) strand (+ or -)
codonpos	reference codon
codon_degeneracy	position on the codon (1, 2 or 3)
Ancestral_allele	degenerate type (0, 2 or 3) ancestral allele based on 8 primates EPO. 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 Neanderthal
Denisova	genotype of a deep sequenced Denisova
Ensembl_geneid	Ensembl gene id
Ensembl_transcriptid	Ensembl transcript ids (Multiple entries separated by ',')
Ensembl_proteinid	Ensembl protein ids Multiple entries separated by ',', corresponding to Ensembl_transcriptids

Column	Description
aapos	amino acid position as to the protein. -1 if the variant is a splicing site SNP (2bp on each end of an intron). Multiple entries separated by ',', corresponding to Ensembl_proteinid
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.

Column	Description
MutationAssessor_score_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.
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.

Column	Description
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.

Column	Description
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_coding_or_noncoding	Whether Eigen-raw and Eigen-phred scores are based on coding model or noncoding model.
Eigen-raw	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	Eigen score in phred scale.
Eigen-PC-raw	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	Eigen PC score in phred scale.
Eigen-PC-raw_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_score_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).
GM12878_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. GM12878 fitCons scores are based on cell type GM12878. More details can be found in doi:10.1038/ng.3196.
GM12878_fitCons_score_rankscore	GM12878 fitCons scores were ranked among all GM12878 fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GM12878 fitCons scores in dbNSFP.

Column	Description
GM12878_confidence_value	0 - highly significant scores (approx. p=.25).
H1-hESC_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. GM12878 fitCons scores are based on cell type H1-hESC. More details can be found in doi:10.1038/ng.3196.
H1-hESC_fitCons_score_rankscore	H1-hESC fitCons scores were ranked among all H1-hESC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of H1-hESC fitCons scores in dbNSFP.
H1-hESC_confidence_value	0 - highly significant scores (approx. p=.25).
HUVEC_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. GM12878 fitCons scores are based on cell type HUVEC. More details can be found in doi:10.1038/ng.3196.
HUVEC_fitCons_score_rankscore	HUVEC fitCons scores were ranked among all HUVEC fitCons scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of HUVEC fitCons scores in dbNSFP.
HUVEC_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.
phyloP100way_vertebrate	phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 100 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from -20.0 to 10.003 in dbNSFP.
phyloP100way_vertebrate_rankscore	phyloP100way_vertebrate scores were ranked among all phyloP100way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP100way_vertebrate scores in dbNSFP.
phyloP20way_mammalian	phyloP (phylogenetic p-values) conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from -13.282 to 1.199 in dbNSFP.

Column	Description
phyloP20way_mammalian_rankscore	phyloP20way_mammalian scores were ranked among all phyloP20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP20way_mammalian scores in dbNSFP.
phastCons100way_vertebrate	phastCons conservation score based on the multiple alignments of 100 vertebrate genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.
phastCons100way_vertebrate_rankscore	phastCons100way_vertebrate scores were ranked among all phastCons100way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons100way_vertebrate scores in dbNSFP.
phastCons20way_mammalian	phastCons conservation score based on the multiple alignments of 20 mammalian genomes (including human). The larger the score, the more conserved the site. Scores range from 0 to 1.
phastCons20way_mammalian_rankscore	phastCons20way_mammalian scores were ranked among all phastCons20way_mammalian scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons20way_mammalian scores in dbNSFP.
SiPhy_29way_pi	The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.
SiPhy_29way_logOdds	SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site. Scores range from 0 to 37.9718 in dbNSFP.
SiPhy_29way_logOdds_rankscore	SiPhy_29way_logOdds scores were ranked among all SiPhy_29way_logOdds scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of SiPhy_29way_logOdds scores in dbNSFP.
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 ALSPAC cohort.
ALSPAC_AF	Alternative allele frequency in called genotypes in UK10K ALSPAC cohort.
ExAC_AC	Allele count in total ExAC samples (60,706 samples)
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

Column	Description
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
ExAC_nonTCGA_AC	Allele count in total ExAC_nonTCGA samples (53,105 samples)
ExAC_nonTCGA_AF	Allele frequency in total ExAC_nonTCGA samples
ExAC_nonTCGA_Adj_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC_nonTCGA samples
ExAC_nonTCGA_Adj_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC_nonTCGA samples
ExAC_nonTCGA_AFR_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC_nonTCGA samples
ExAC_nonTCGA_AFR_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC_nonTCGA samples
ExAC_nonTCGA_AMR_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC_nonTCGA samples
ExAC_nonTCGA_AMR_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC_nonTCGA samples
ExAC_nonTCGA_EAS_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC_nonTCGA samples
ExAC_nonTCGA_EAS_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC_nonTCGA samples
ExAC_nonTCGA_FIN_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC_nonTCGA samples
ExAC_nonTCGA_FIN_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC_nonTCGA samples

Column	Description
ExAC_nonTCGA_NFE_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC_nonTCGA samples
ExAC_nonTCGA_NFE_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC_nonTCGA samples
ExAC_nonTCGA_SAS_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC_nonTCGA samples
ExAC_nonTCGA_SAS_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC_nonTCGA samples
ExAC_nonpsych_AC	Allele count in total ExAC_nonpsych samples (45,376 samples)
ExAC_nonpsych_AF	Allele frequency in total ExAC_nonpsych samples
ExAC_nonpsych_Adj_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in total ExAC_nonpsych samples
ExAC_nonpsych_Adj_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in total ExAC_nonpsych samples
ExAC_nonpsych_AFR_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in African & African American ExAC_nonpsych samples
ExAC_nonpsych_AFR_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in African & African American ExAC_nonpsych samples
ExAC_nonpsych_AMR_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in American ExAC_nonpsych samples
ExAC_nonpsych_AMR_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in American ExAC_nonpsych samples
ExAC_nonpsych_EAS_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in East Asian ExAC_nonpsych samples
ExAC_nonpsych_EAS_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in East Asian ExAC_nonpsych samples
ExAC_nonpsych_FIN_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Finnish ExAC_nonpsych samples
ExAC_nonpsych_FIN_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Finnish ExAC_nonpsych samples
ExAC_nonpsych_NFE_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC_nonpsych samples
ExAC_nonpsych_NFE_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in Non-Finnish European ExAC_nonpsych samples
ExAC_nonpsych_SAS_AC	Adjusted Alt allele counts (DP >= 10 & GQ >= 20) in South Asian ExAC_nonpsych samples

Column	Description
ExAC_nonpsych_SAS_AF	Adjusted Alt allele frequency (DP >= 10 & GQ >= 20) in South Asian ExAC_nonpsych samples
gnomAD_exomes_AC	Alternative allele count in the whole gnomAD exome samples (123,136 samples)
gnomAD_exomes_AN	Total allele count in the whole gnomAD exome samples (123,136 samples)
gnomAD_exomes_AF	Alternative allele frequency in the whole gnomAD exome samples (123,136 samples)
gnomAD_exomes_AFR_AC	Alternative allele count in the African/African American gnomAD exome samples (7,652 samples)
gnomAD_exomes_AFR_AN	Total allele count in the African/African American gnomAD exome samples (7,652 samples)
gnomAD_exomes_AFR_AF	Alternative allele frequency in the African/African American gnomAD exome samples (7,652 samples)
gnomAD_exomes_AMR_AC	Alternative allele count in the Latino gnomAD exome samples (16,791 samples)
gnomAD_exomes_AMR_AN	Total allele count in the Latino gnomAD exome samples (16,791 samples)
gnomAD_exomes_AMR_AF	Alternative allele frequency in the Latino gnomAD exome samples (16,791 samples)
gnomAD_exomes_ASJ_AC	Alternative allele count in the Ashkenazi Jewish gnomAD exome samples (4,925 samples)
gnomAD_exomes_ASJ_AN	Total allele count in the Ashkenazi Jewish gnomAD exome samples (4,925 samples)
gnomAD_exomes_ASJ_AF	Alternative allele frequency in the Ashkenazi Jewish gnomAD exome samples (4,925 samples)
gnomAD_exomes_EAS_AC	Alternative allele count in the East Asian gnomAD exome samples (8,624 samples)
gnomAD_exomes_EAS_AN	Total allele count in the East Asian gnomAD exome samples (8,624 samples)
gnomAD_exomes_EAS_AF	Alternative allele frequency in the East Asian gnomAD exome samples (8,624 samples)
gnomAD_exomes_FIN_AC	Alternative allele count in the Finnish gnomAD exome samples (11,150 samples)
gnomAD_exomes_FIN_AN	Total allele count in the Finnish gnomAD exome samples (11,150 samples)
gnomAD_exomes_FIN_AF	Alternative allele frequency in the Finnish gnomAD exome samples (11,150 samples)

Column	Description
gnomAD_exomes_NFE_AC	Alternative allele count in the Non-Finnish European gnomAD exome samples (55,860 samples)
gnomAD_exomes_NFE_AN	Total allele count in the Non-Finnish European gnomAD exome samples (55,860 samples)
gnomAD_exomes_NFE_AF	Alternative allele frequency in the Non-Finnish European gnomAD exome samples (55,860 samples)
gnomAD_exomes_SAS_AC	Alternative allele count in the South Asian gnomAD exome samples (15,391 samples)
gnomAD_exomes_SAS_AN	Total allele count in the South Asian gnomAD exome samples (15,391 samples)
gnomAD_exomes_SAS_AF	Alternative allele frequency in the South Asian gnomAD exome samples (15,391 samples)
gnomAD_genomes_AC	Alternative allele count in the whole gnomAD genome samples (15,496 samples)
gnomAD_genomes_AN	Total allele count in the whole gnomAD genome samples (15,496 samples)
gnomAD_genomes_AF	Alternative allele frequency in the whole gnomAD genome samples (15,496 samples)
gnomAD_genomes_AFR_AC	Alternative allele count in the African/African American gnomAD genome samples (4,368 samples)
gnomAD_genomes_AFR_AN	Total allele count in the African/African American gnomAD genome samples (4,368 samples)
gnomAD_genomes_AFR_AF	Alternative allele frequency in the African/African American gnomAD genome samples (4,368 samples)
gnomAD_genomes_AMR_AC	Alternative allele count in the Latino gnomAD genome samples (419 samples)
gnomAD_genomes_AMR_AN	Total allele count in the Latino gnomAD genome samples (419 samples)
gnomAD_genomes_AMR_AF	Alternative allele frequency in the Latino gnomAD genome samples (419 samples)
gnomAD_genomes_ASJ_AC	Alternative allele count in the Ashkenazi Jewish gnomAD genome samples (151 samples)
gnomAD_genomes_ASJ_AN	Total allele count in the Ashkenazi Jewish gnomAD genome samples (151 samples)
gnomAD_genomes_ASJ_AF	Alternative allele frequency in the Ashkenazi Jewish gnomAD genome samples (151 samples)
gnomAD_genomes_EAS_AC	Alternative allele count in the East Asian gnomAD genome samples (811 samples)

Column	Description
gnomAD_genomes_EAS_AN	Total allele count in the East Asian gnomAD genome samples (811 samples)
gnomAD_genomes_EAS_AF	Alternative allele frequency in the East Asian gnomAD genome samples (811 samples)
gnomAD_genomes_FIN_AC	Alternative allele count in the Finnish gnomAD genome samples (1,747 samples)
gnomAD_genomes_FIN_AN	Total allele count in the Finnish gnomAD genome samples (1,747 samples)
gnomAD_genomes_FIN_AF	Alternative allele frequency in the Finnish gnomAD genome samples (1,747 samples)
gnomAD_genomes_NFE_AC	Alternative allele count in the Non-Finnish European gnomAD genome samples (7,509 samples)
gnomAD_genomes_NFE_AN	Total allele count in the Non-Finnish European gnomAD genome samples (7,509 samples)
gnomAD_genomes_NFE_AF	Alternative allele frequency in the Non-Finnish European gnomAD genome samples (7,509 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_V6p_gene	target gene of the (significant) eQTL SNP
GTEx_V6p_tissue	tissue type of the expression data with which the eQTL/gene pair is detected
Gene_old_names	Old gene symbol (from HGNC)
Gene_other_names	Other gene names (from HGNC)
Uniprot_acc	Uniprot acc number (from HGNC and Uniprot)
Uniprot_id	Uniprot id (from HGNC and Uniprot)
Entrez_gene_id	Entrez gene id (from HGNC)
CCDS_id	CCDS id (from HGNC)
Refseq_id	Refseq gene id (from HGNC)
ucsc_id	UCSC gene id (from HGNC)
MIM_id	MIM gene id (from HGNC)
Gene_full_name	Gene full name (from HGNC)
Pathway(Uniprot)	Pathway description from Uniprot

Column	Description
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

Column	Description
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.
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)
ExAC_pLI	'the probability of being loss-of-function intolerant (intolerant of both heterozygous and homozygous lof variants)' based on ExAC r0.3 data
ExAC_pRec	'the probability of being intolerant of homozygous, but not heterozygous lof variants' based on ExAC r0.3 data
ExAC_pNull	'the probability of being tolerant of both heterozygous and homozygous lof variants' based on ExAC r0.3 data
ExAC_nonTCGA_pLI	'the probability of being loss-of-function intolerant (intolerant of both heterozygous and homozygous lof variants)' based on ExAC r0.3 nonTCGA subset
ExAC_nonTCGA_pRec	'the probability of being intolerant of homozygous, but not heterozygous lof variants' based on ExAC r0.3 nonTCGA subset
ExAC_nonTCGA_pNull	'the probability of being tolerant of both heterozygous and homozygous lof variants' based on ExAC r0.3 nonTCGA subset

Column	Description
ExAC_nonpsych_pLI	'the probability of being loss-of-function intolerant (intolerant of both heterozygous and homozygous lof variants)' based on ExAC r0.3 nonpsych subset
ExAC_nonpsych_pRec	'the probability of being intolerant of homozygous, but not heterozygous lof variants' based on ExAC r0.3 nonpsych subset
ExAC_nonpsych_pNull	'the probability of being tolerant of both heterozygous and homozygous lof variants' based on ExAC r0.3 nonpsych subset
ExAC_del.score	'Winsorised deletion intolerance z-score' based on ExAC r0.3.1 CNV data
ExAC_dup.score	'Winsorised duplication intolerance z-score' based on ExAC r0.3.1 CNV data
ExAC_cnv.score	'Winsorised cnv intolerance z-score' based on ExAC r0.3.1 CNV data
ExAC_cnv_flag	'Gene is in a known region of recurrent CNVs mediated by tandem segmental duplications and intolerance scores are more likely to be biased or noisy.' from ExAC r0.3.1 CNV release
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

Column	Description
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
LoFtool_score	a percential score for gene intolerance to functional change. The lower the score the higher gene intolerance to functional change. For details please contact Dr. Joao Fadista(joao.fadista@med.lu.se)
SORVA_LOF_MAF0.005_HetOrHom	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Heterozygote or Homozygote of LOF SNVs whose MAF<0.005. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOF_MAF0.005_HomOrCompoundHet	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Compound Heterozygote or Homozygote of LOF SNVs whose MAF<0.005. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOF_MAF0.001_HetOrHom	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Heterozygote or Homozygote of LOF SNVs whose MAF<0.001. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOF_MAF0.001_HomOrCompoundHet	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Compound Heterozygote or Homozygote of LOF SNVs whose MAF<0.001. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.

Column	Description
SORVA_LOForMissense_MAF0.005_HetOrHom	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Heterozygote or Homozygote of LOF or missense SNVs whose MAF<0.005. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOForMissense_MAF0.005_HomOrCompoundHet	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Compound Heterozygote or Homozygote of LOF or missense SNVs whose MAF<0.005. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOForMissense_MAF0.001_HetOrHom	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Heterozygote or Homozygote of LOF or missense SNVs whose MAF<0.001. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
SORVA_LOForMissense_MAF0.001_HomOrCompoundHet	the fraction of individuals in the 1000 Genomes Project data (N=2504) who are either Compound Heterozygote or Homozygote of LOF or missense SNVs whose MAF<0.001. This fraction is from a method for ranking genes based on mutational burden called SORVA (Significance Of Rare VAriants). Please see doi: 10.1101/103218 for details.
Essential_gene	Essential (E) or Non-essential phenotype-changing (N) based on Mouse Genome Informatics database. from doi:10.1371/journal.pgen.1003484
MGI_mouse_gene	Homolog mouse gene name from MGI
MGI_mouse_phenotype	Phenotype description for the homolog mouse gene from MGI
ZFIN_zebrafish_gene	Homolog zebrafish gene name from ZFIN
ZFIN_zebrafish_structure	Affected structure of the homolog zebrafish gene from ZFIN
ZFIN_zebrafish_phenotype_quality	Phenotype description for the homolog zebrafish gene from ZFIN
ZFIN_zebrafish_phenotype_tag	Phenotype tag for the homolog zebrafish gene from ZFIN

- dbSNP : dbSNP138, dbSNP154
- Prediction scores : SIFT, LRT, MutationTaster, FATHMM, PROBEN, MetaSVM, MetaLR, M-AP, MutaPred, fathmmMKL, Eigen, GenoCanyon, fitCons
- Conservation scores : GERP++, phyloP, phastCons, SiPhy
- Gene information : GTEx, CCDS, Refseq, MIM
- Pathway : Uniprot, BioCarts, KEGG

Analysis Tools

BWA (Burrows-Wheeler Alignment Tool) 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 2.18.2

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) 4.0.5.1

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 (Annotation Tool) 5.0e

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). Using this tool, we follow the annotation cascade shown below.
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

Control-FREEC (Copy Number Variant Caller) 11.5

Control-FREEC is a tool which enables automatic calculation of copy number and allelic content profiles, and consequently predicts regions of genomic alterations such as gains and losses. It accurately calls genotype status even when no control experiment is available. It also corrects for GC-content mappability biases of the polyploid genomes.

More information can be found here:

https://academic.oup.com/bioinformatics/article/28/3/423/189142

PennCNV (Annotation Tool) 1.0.5

PennCNV is a tool which identifies overlapping or neighboring genes for Copy Number Variation(CNV) annotation. The scan_region.pl program, one of the pennCNV packages efficiently searches for overlapping CNV calls with UCSC known gene annotation. The output file contains two additional columns representing the gene symbols and the distance between CNV and gene.

More information can be found here:

http://penncnv.openbioinformatics.org/en/latest/user-guide/annotation/

Manta (Structural Variant Caller) 1.5.0

Manta is a tool to call structural variants and indels from short paired-end sequencing reads. It combines paired-end and split read evidence during SV discovery and scoring to improve performance. However, it does not require split reads or successful breakpoint assemblies to report a variant in cases where there is strong evidence of an imprecise variant. It provides genotypes and quality scores for variants in single diploid samples, and will also call somatic variants when a matched tumor sample is specified. Manta can detect all classes of structural variants which can be identified in the absence of copy number analysis and large-scale assembly.

More information can be found here:

https://github.com/StructuralVariants/manta

Circos 0.69-6

Circos is a tool for visualizing data into a circular layout. The number and types of variants or relationship between chromosomes can be represented by tracks.

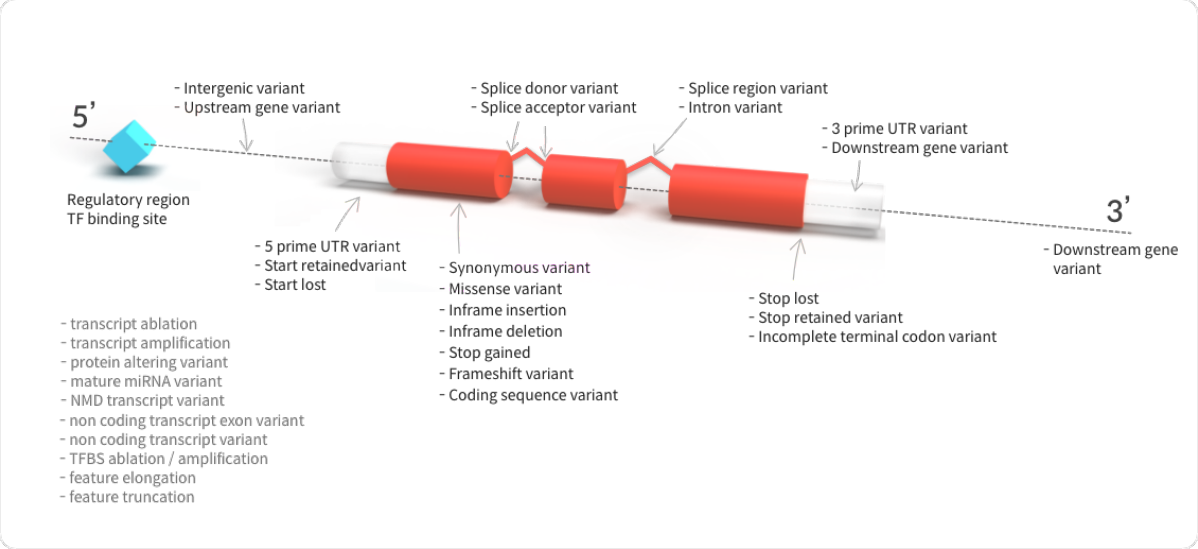
More information can be found here:

https://pubmed.ncbi.nlm.nih.gov/19541911/

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
frameshift_variant	Insertion or deletion causes a frame shift e.g.: An indel size is not multple of 3.	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
inframe_insertion	One or many codons are inserted. e.g.: An insert multiple of three in a codon boundary	SO:0001821
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
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence.	SO:0001822
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
missense_variant	Variant causes a codon that produces a different amino acid. e.g.: Tgg/Cgg, W/R	SO:0001583
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	Variant causes start codon to be mutated into another start codon (the new codon produces a different AA). e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons)	SO:0001582
synonymous_variant	Variant causes a codon that produces the same amino acid. e.g.: Ttg/Ctg, L/L	SO:0001819

SO Term	SO Description	SO Accession
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
coding_sequence_variant	The variant hits a CDS.	SO:0001580
5_prime_UTR_variant	Variant hits 5'UTR region.	SO:0001623
3_prime_UTR_variant	Variant hits 3'UTR region.	SO:0001624
intron_variant	Variant hits and intron. Technically, hits no exon in the transcript.	SO:0001627
non_coding_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript.	SO:0001792
upstream_gene_variant	Upstream of a gene (default length: 5K bases).	SO:0001631
downstream_gene_variant	Downstream of a gene (default length: 5K bases).	SO:0001632
TF_binding_site_variant	A sequence variant located within a transcription factor binding site.	SO:0001782
regulatory_region_variant	The variant hits a known regulatory feature (non-coding).	SO:0001566
intergenic_variant	A sequence variant located in the intergenic region, between genes.	SO:0001628
bidirectional_gene_fusion	Fusion of two genes in opposite directions.	SO:0002086
chromosome_number_variation	A kind of chromosome variation where the chromosome complement is not an exact multiple of the haploid number.	SO:1000182
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
duplication	Duplication of a large chromosome segment (over 1% or 1,000,000 bases).	SO:1000035
exon_loss_variant	A deletion removes the whole exon.	SO:0001572
exon_region	The variant is in an exonic region.	SO:0000852
feature_ablation	Deletion of a gene.	SO:0001879
feature_fusion	A sequence variant, caused by an alteration of the genomic sequence, where a deletion fuses genomic features.	SO:0001882
gene_fusion	A sequence variant whereby a two genes have become joined.	SO:0001565
intergenic_region	The variant is in an intergenic region.	SO:0000605

SO Term	SO Description	SO Accession
intragenic_variant	The variant hits a gene, but no transcripts within the gene.	SO:0002011
inversion	Inversion of a large chromosome segment (over 1% or 1,000,000 bases).	SO:1000036
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript.	SO:0001792
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature.	SO:0001893

• Click on the “So Accession” link for more information.

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.

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.

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
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://sift.bii.a-star.edu.sg/

gnomAD

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

More information can be found here:

http://gnomad.broadinstitute.org

ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

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