Humanizing Genomics
macrogen

Homo sapiens
Whole Exome Sequencing
Report

February 2016

Research use only

Basic Information

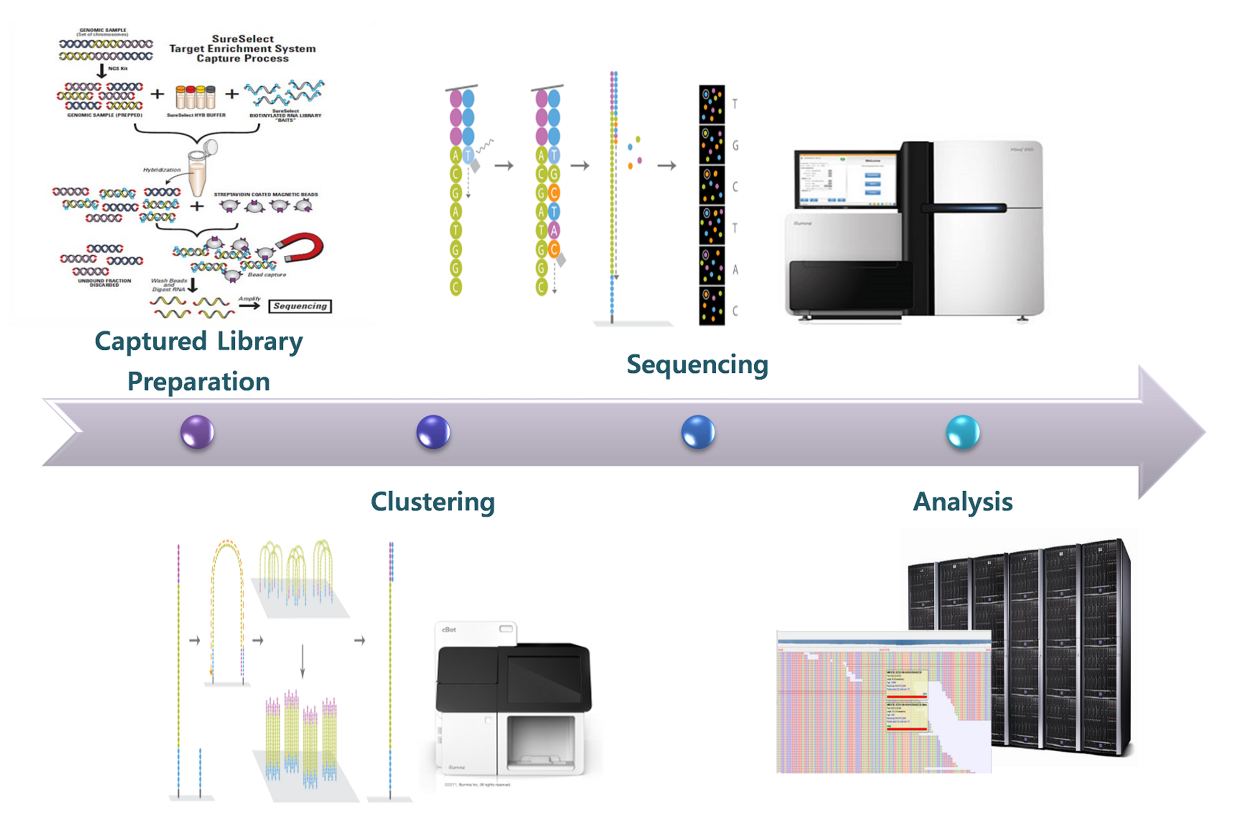
Sample	
Order Number	
Company/Institute	Macrogen Corp. Japan
Client Name	Macrogen Japan
Capture Kit	SureSelect V5-post
Type of Sequencer	HiSeq4000

Research use only

Table of Contents

1. Experimental Methods and Workflow

1. 1. Experiment Overview



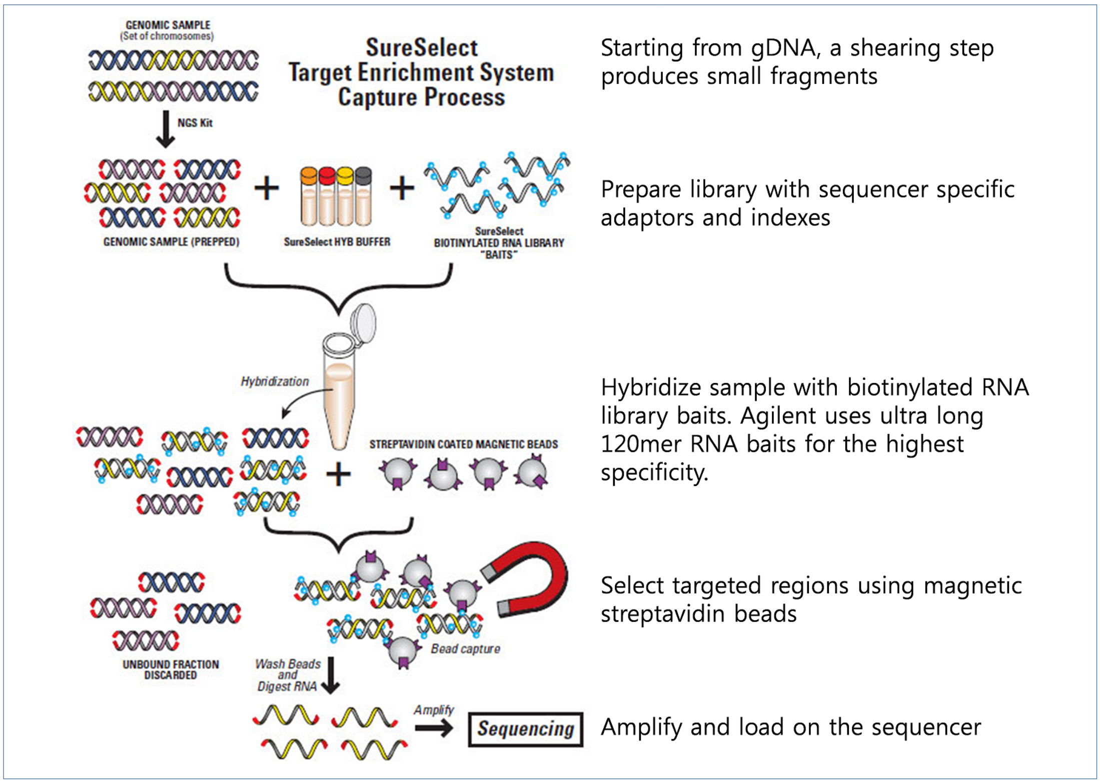
The samples were prepared according to an Agilent SureSelect Target Enrichment Kit preparation guide. The libraries were sequenced with Illumina HiSeq 2000/2500 sequencer.

1. 2. Experiment Procedure

1. 2. 1. Captured Library Construction

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

<SureSelect Target Enrichment Workflow>



Starting from gDNA, a shearing step produces small fragments

Prepare library with sequencer specific adaptors and indexes

Hybridize sample with biotinylated RNA library baits. Agilent uses ultra long 120mer RNA baits for the highest specificity.

Select targeted regions using magnetic streptavidin beads

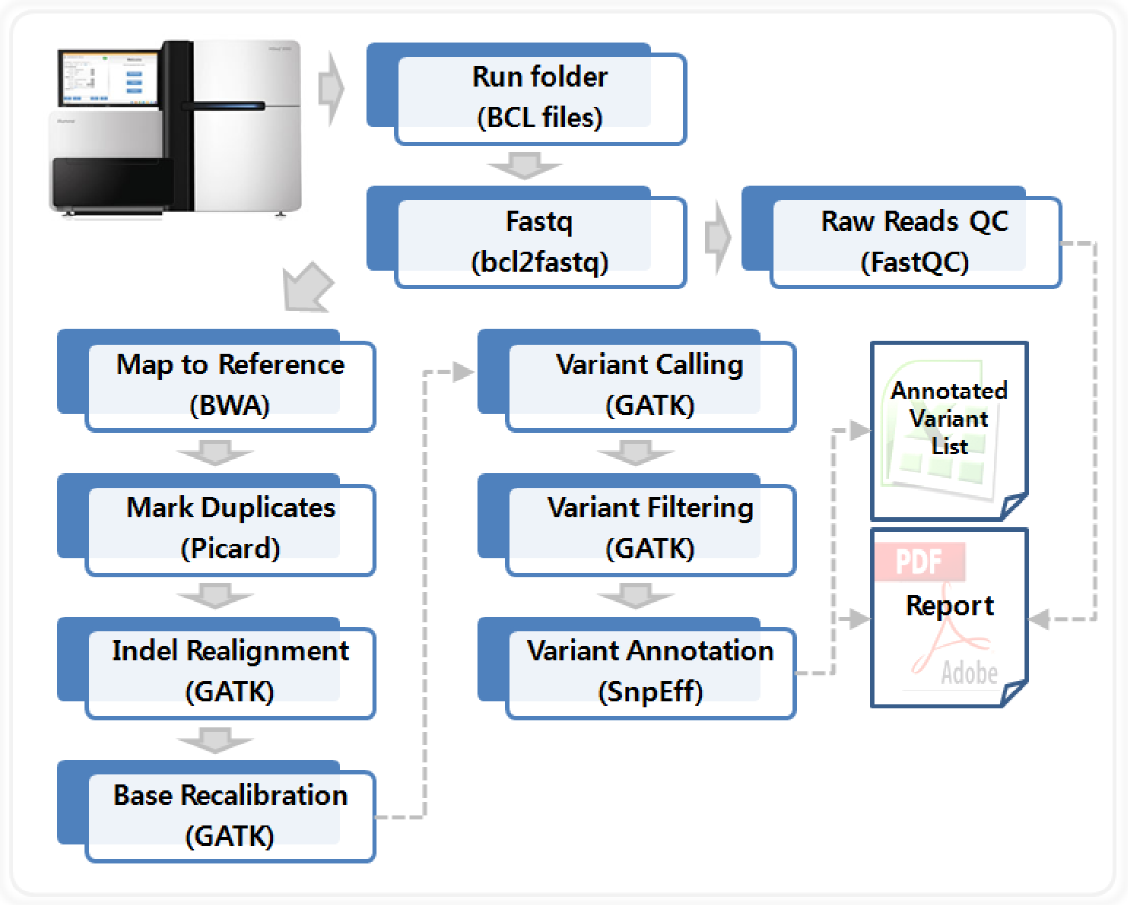
Amplify and load on the sequencer

1. 2. 2. Clustering & Sequencing

Illumina utilizes a unique "bridged" amplification reaction that occurs on the surface of the flow cell. A flow cell containing millions of unique clusters is loaded into the HiSeq 2000/2500 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.

2. Analysis Methods and Workflow

2. 1. Analysis Overview



2. 2. Analysis Software

2. 2. 1. BWA (Burrows-Wheeler Alignment Tool)

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

2. 2. 2. Picard

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

2. 2. 3. GATK (Genome Analysis Toolkit)

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

2. 2. 4. SnpEff

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

2. 3. Resources

2. 3. 1. Mapping Reference

hg19 from UCSC (original GRCh37 from NCBI, Feb. 2009)

2. 3. 2. Software

Software	Version
BWA	bwa-0.7.10
Picard	picard-tools-1.118
GATK	GATK3.v4
SnpEff	SnpEff_v4.1

2. 3. 3. Tuned Parameters

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
	SO	coordinate	Sort order
	REMOVE_DUPLICATES	true	
	AS	true	Assume Sorted
	CREATE_INDEX	true	Create index files
GATK	-T	RealignerTargetCreator	Determine (small) suspicious intervals
		IndelRealigner	Running the realigner
		BaseRecalibrator	Generate the first pass recalibration table file
		HaplotypeCaller	Call SNPs and indels simultaneously via local re-assembly of haplotypes in an active region.
		Selectvariants	Selects variants from a VCF source

		VariantFilteration	Filters variant calls using a number of user-selectable, parameterizable criteria.
		Combinevariants	Combines VCF records from different sources.
	-knownSites	1000G_phase1.indels.hg19.vcf	database of known polymorphic sites
		dbsnp_138.hg19.vcf	
		Mills_and_1000G_gold_standard.indels.hg19.sites.vcf	

* Software not listed in the table uses all default settings

2. 3. 4. Annotation Database

Database	Version
dbSNP	142
1000Genome	Phase3
ClinVar	05/2015
ESP	ESP6500SI_V2

3. Analysis Result

3. 1. Sample & Run information

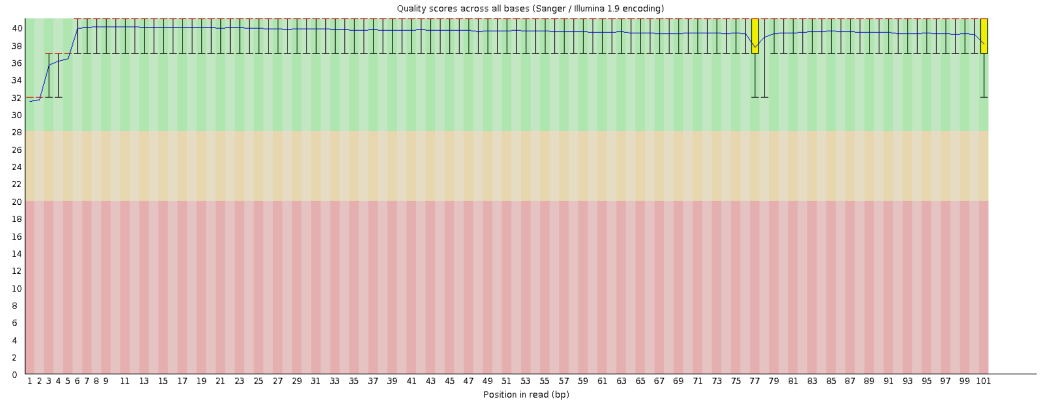
Sample	452
Order Number	1601KHF-0012
Capture Kit	SureSelect V5-post
Type of Sequencer	HiSeq4000

3. 2. Fastq

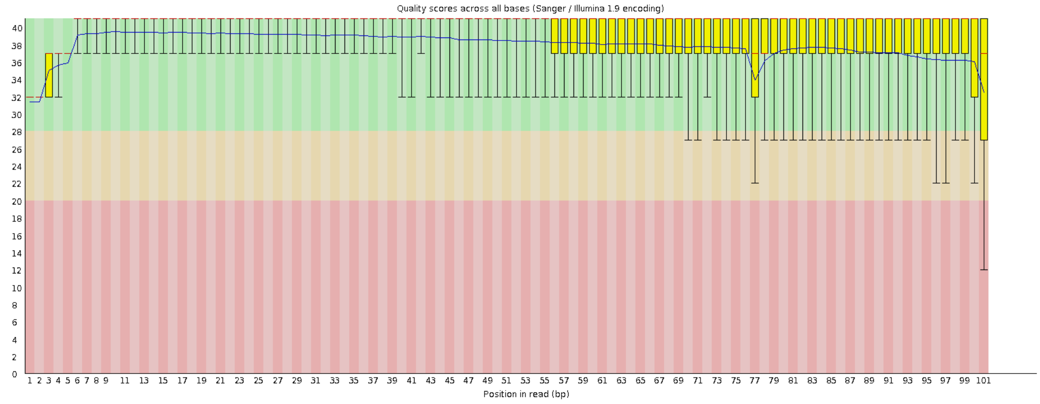
3. 2. 1. Statistics

Sample	Total Read Bases (bp)	Total Reads	GC(%)	Q20(%)	Q30(%)
452	5,857,368,952	57,993,752	48.9	97.	93.7

3. 2. 2. Read1 Quality by Cycle



3. 2. 3. Read2 Quality by Cycle



3. 3. Pre-alignment Statistics

Total Number of Reads	57,993,752
Average Read Length (bp)	101.0
Total Yield (Mbp)	5,857
Target Regions (bp)	50,390,601
Average Throughput Depth of Target regions (X)	116.2

- Total yield: {total number of reads} * {Average read length}
- Target regions : Target region size
- Average throughput depth of target regions (X) : {Total yield} / {Target regions}

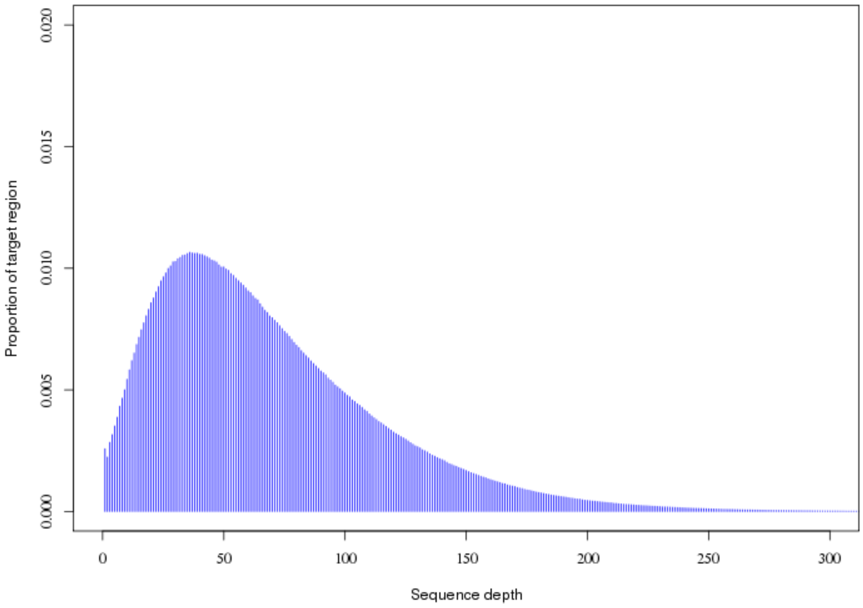
3. 4. Post-alignment Statistics

Initial Mappable Reads	57,967,991
% Initial Mappable Reads	99.9
Non-Redundant Reads	51,221,923
% Non-Redundant Reads	88.3
On-Target Reads	39,705,504
% On-Target Reads	77.5
On-Target Yield (bp)	3,446,163,437
Mean Depth of Target Regions (X)	68.3

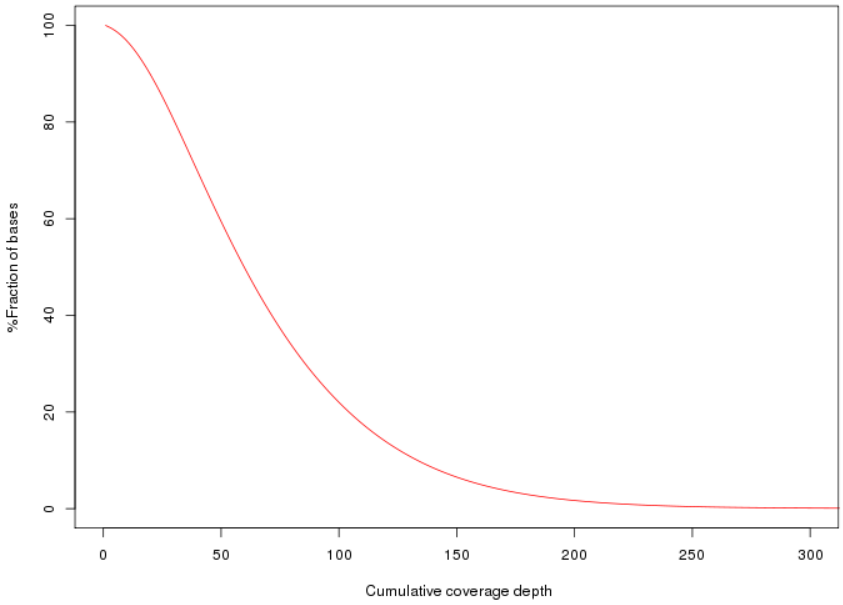
- Initial Mappable Reads : Number of mapped reads to human genome
- % Initial Mappable Reads: 100 * { Initial mappable reads} / {Total reads}
- Non-Redundant Reads : Number of de-duplicate reads from Picard tools
- % Non-Redundant Reads: 100 * {Non-redundant reads } / { Initial mappable reads}
- On-Target 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}

3. 5. Alignment Coverage

3. 5. 1. Histogram of Depth Distribution in Target Regions



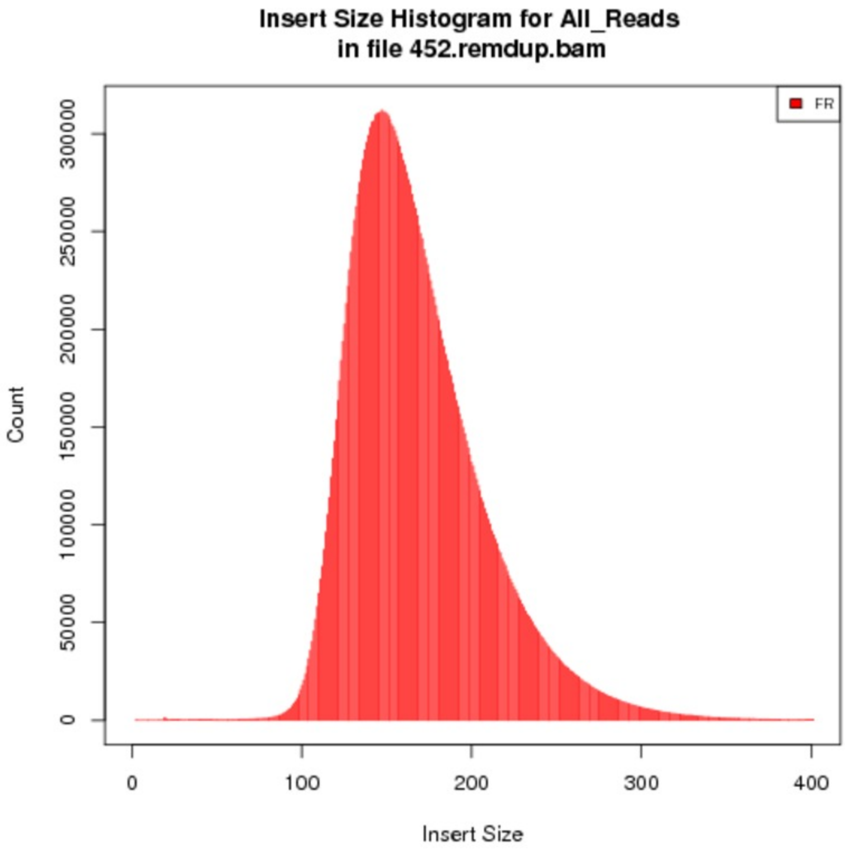
3. 5. 2. Cumulative Depth Distribution in Target Regions



%Coverage	%>1X	%>10X	%>20X	%>30X	%>50X
Value	99.7	96.2	88.9	79.2	58.4

- % Coverage : The percentage of bases in target regions with a depth of coverage or greater

3. 6. Insert Statistics



Fragment Length Median	Standard Deviation
161	39.3

4. SNP & INDEL

# of SNP	75,479
Synonymous Variant	11,244
Missense Variant	10,356
Stop Gained	94
Stop Lost	35
# of INDEL	7,937
Frameshift Variant	277
Inframe Insertion	146
Inframe Deletion	174
% Found in dbSNP142	97.6
Het/Hom Ratio	1.2
Ts/Tv Ratio	2.3

- 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).

5. Data Deliverables

5. 1. Deliverables List

File Type	File Name	Description
Raw Data	Sample1_1.fastq.gz	Raw read1 sequence data
	Sample1_2.fastq.gz	Raw read2 sequence data
Alignment Result	Sample1.recal.bam	Bwa alignment file
	Sample1.recal.bam.bai	Bwa alignment index file
Variant Call Result	Sample1.Filtered.Varinats.vcf	SNP/INDEL file (vcf format)
	Sample1_SNP_Indel_ANNO.xlsx	Annotated variant list file (excel file)
Summary	Sample1.pdf	Analysis Report

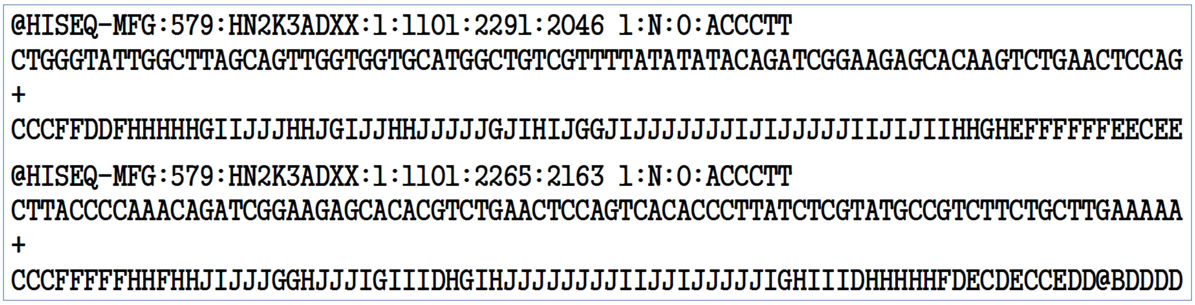
* FASTQ files are compressed in the GNU zip format, an open source data compression program

5. 2. Deliverables File Format

5. 2. 1. FASTQ File

5. 2. 1. 1. FASTQ Format

Example :



```
@HISEQ-MFG:579:HN2K3ADXX:1:1101:2291:2046 1:N:0:ACCCTT
CTGGGTATTGGCTTAGCAGTTGGTGGTGCATGGCTGTCGTTTTATATATACAGATCGGAAGAGCACAAGTCTGAACTCCAG
+
CCCFFDDFHHHHHGIIJJJHHJGIJJHHJJJJGJIHIJGGJIJJJJJJJIJIJJJJJIIJIJIIHHGHEFFFFFFEECEE
@HISEQ-MFG:579:HN2K3ADXX:1:1101:2265:2163 1:N:0:ACCCTT
CTTACCCCAAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACCCTTATCTCGTATGCCGTCTTCTGCTTGAAAAA
+
CCCFFFFFHHFHHJIJJJGGHJJJIGIIIDHGIHJJJJJJJJIIJJIJJJJJIGHIIIDHHHHHFDECDECCCEDD@BDDDD
```

FASTQ File Consists of Four Lines

- Line1 : Sequence identifier
- Line2 : Nucleotide sequences
- Line3 : Quality score identifier line - character '+'
- Line4 : Quality score

5. 2. 1. 2. Phred Scores

$Q = -10 \log_{10}(\text{error rate})$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%
60	1 in 1000000	99.9999%

- Encoding: ASCII Character Code=Phred Quality Value + 33

5. 2. 1. 3. Q-Score Binning (HiSeq4000 only)

HiSeq4000 groups quality scores into specific ranges, or bins, and assigns a value to each range.

For example, the original quality scosres 20-24 may from one bin, and can all be mapped to a new value of 22. Q-score binning significantly reduces storage space requirements without affecting accuracy or performance of downstream applications. Please refer to this table below, Q Scores for HiSeq4000 are binned using the following criteria.

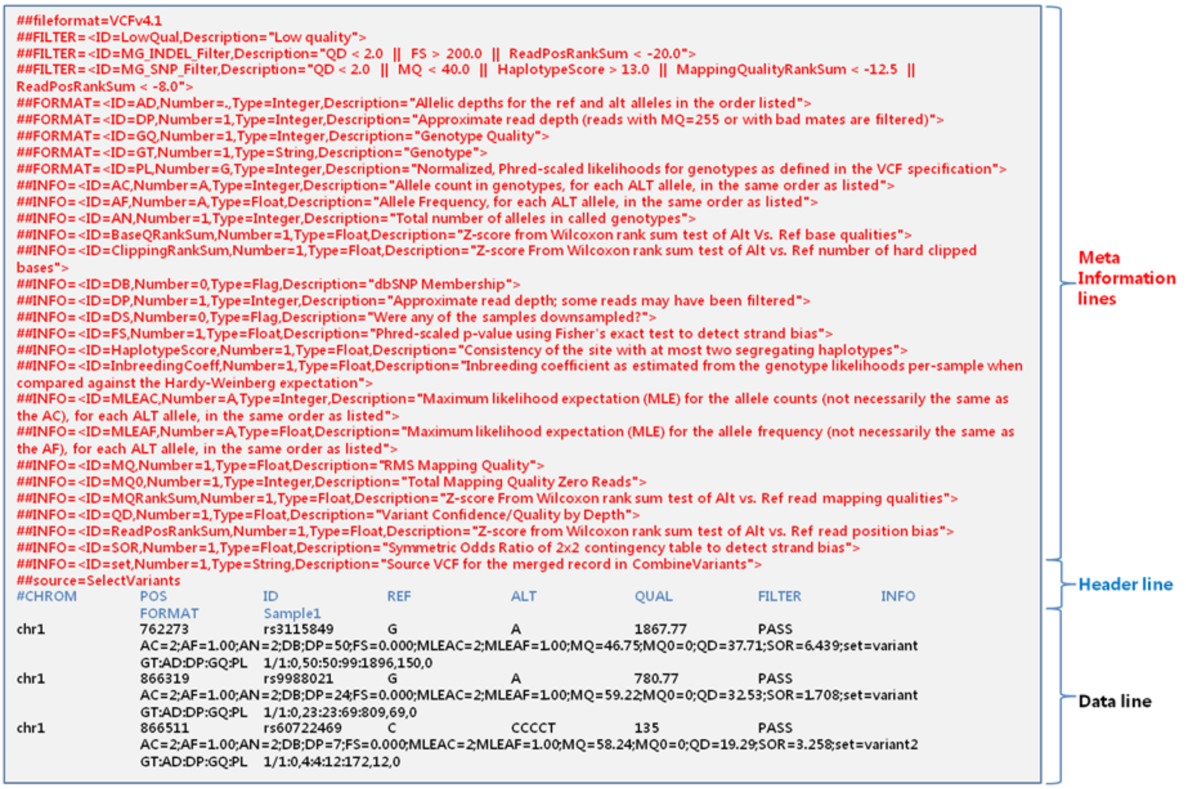
Q-Score Bins	Example of Empirically Mapped Q-Scores
N (no call)	N (no call)
2-9	7
10-19	11
20-24	22
25-29	27
30-34	32
35-39	37
40-45	42

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

5. 2. 2. VCF

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=VCFv4.1
##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  ||  MQ < 40.0  ||  HaplotypeScore > 13.0  ||  MappingQualityRankSum < -12.5  ||
ReadPosRankSum < -8.0">
##FORMAT=<ID=AD,Number=.,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=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##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=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##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=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##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=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##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=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">
##INFO=<ID=set,Number=1,Type=String,Description="Source VCF for the merged record in CombineVariants">
##source=SelectVariants
#CHROM          POS             ID              REF             ALT             QUAL            FILTER          INFO
                FORMAT          Sample1
chr1            762273          rs3115849       G               A               1867.77         PASS
                AC=2;AF=1.00;AN=2;DB;DP=50;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=46.75;MQ0=0;QD=37.71;SOR=6.439;set=variant
                GT:AD:DP:GQ:PL  1/1:0,50:50:99:1896,150,0
chr1            866319          rs9988021       G               A               780.77          PASS
                AC=2;AF=1.00;AN=2;DB;DP=24;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=59.22;MQ0=0;QD=32.53;SOR=1.708;set=variant
                GT:AD:DP:GQ:PL  1/1:0,23:23:69:809,69,0
chr1            866511          rs60722469      C               CCCCT           135             PASS
                AC=2;AF=1.00;AN=2;DB;DP=7;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=58.24;MQ0=0;QD=19.29;SOR=3.258;set=variant2
                GT:AD:DP:GQ:PL  1/1:0,4:4:12:172,12,0
```

5. 2. 2. 1. 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: <key>=<data>. 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.

5. 2. 2. 2. FILTER Tag

Tag	Description
LowQual	Low quality
MG_INDEL_Filter	QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0
MG_SNP_Filter	QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0

5. 2. 2. 3. 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.

5. 2. 2. 4. FORMAT Tag

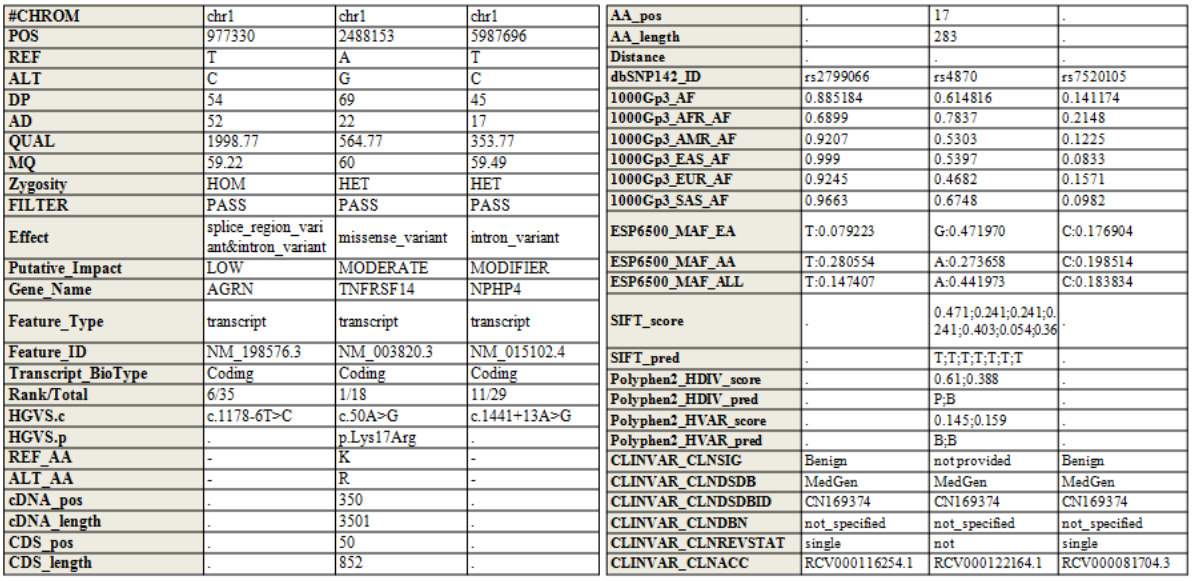
Tag	Description
GT	Genotype 0/0 - the sample is homozygous reference 0/1 - the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles 1/1 - the sample is homozygous alternate
AD	Allelic depths for the ref and alt alleles in the order listed.
DP	Read depth at this position for this 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.

5. 2. 3. Annotated Variant List File (*_SNP_indel_ANNO.xlsx)

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.

The contents will look like following sample sheet (the table is pivoted) :

Example :



#CHROM	chr1	chr1	chr1
POS	977330	2488153	5987696
REF	T	A	T
ALT	C	G	C
DP	54	69	45
AD	52	22	17
QUAL	1998.77	564.77	353.77
MQ	59.22	60	59.49
Zygosity	HOM	HET	HET
FILTER	PASS	PASS	PASS
Effect	splice_region_variant&intron_variant	missense_variant	intron_variant
Putative_Impact	LOW	MODERATE	MODIFIER
Gene_Name	AGRN	TNFRSF14	NPHP4
Feature_Type	transcript	transcript	transcript
Feature_ID	NM_198576.3	NM_003820.3	NM_015102.4
Transcript_BioType	Coding	Coding	Coding
Rank/Total	6/35	1/18	11/29
HGVS.c	c.1178-6T>C	c.50A>G	c.1441+13A>G
HGVS.p	.	p.Lys17Arg	.
REF_AA	-	K	-
ALT_AA	-	R	-
cDNA_pos	.	350	.
cDNA_length	.	3501	.
CDS_pos	.	50	.
CDS_length	.	852	.

AA_pos	.	17	.
AA_length	.	283	.
Distance	.	.	.
dbSNP142_ID	rs2799066	rs4870	rs7520105
1000Gp3_AF	0.885184	0.614816	0.141174
1000Gp3_AFR_AF	0.6899	0.7837	0.2148
1000Gp3_AMR_AF	0.9207	0.5303	0.1225
1000Gp3_EAS_AF	0.999	0.5397	0.0833
1000Gp3_EUR_AF	0.9245	0.4682	0.1571
1000Gp3_SAS_AF	0.9663	0.6748	0.0982
ESP6500_MAF_EA	T:0.079223	G:0.471970	C:0.176904
ESP6500_MAF_AA	T:0.280554	A:0.273658	C:0.198514
ESP6500_MAF_ALL	T:0.147407	A:0.441973	C:0.183834
SIFT_score	.	0.471;0.241;0.241;0.241;0.403;0.054;0.36	.
SIFT_pred	.	T;T;T;T;T;T;T	.
Polyphen2_HDIV_score	.	0.61;0.388	.
Polyphen2_HDIV_pred	.	P;B	.
Polyphen2_HVAR_score	.	0.145;0.159	.
Polyphen2_HVAR_pred	.	B;B	.
CLINVAR_CLNSIG	Benign	not provided	Benign
CLINVAR_CLNDSDB	MedGen	MedGen	MedGen
CLINVAR_CLNDSDBID	CN169374	CN169374	CN169374
CLINVAR_CLNDBN	not_specified	not_specified	not_specified
CLINVAR_CLNREVSTAT	single	not	single
CLINVAR_CLNACC	RCV000116254.1	RCV000122164.1	RCV000081704.3

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[1]	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_Len	Trancript's cDNA length
CDS_Pos	Position of coding bases (one based includes START and STOP codons).
CDS_Len	Number of coding bases (one based includes START and STOP codons).
AA_Pos	Position of AA (one based, including START, but not STOP).
AA_Len	Number of AA (one based includes START and STOP codons).

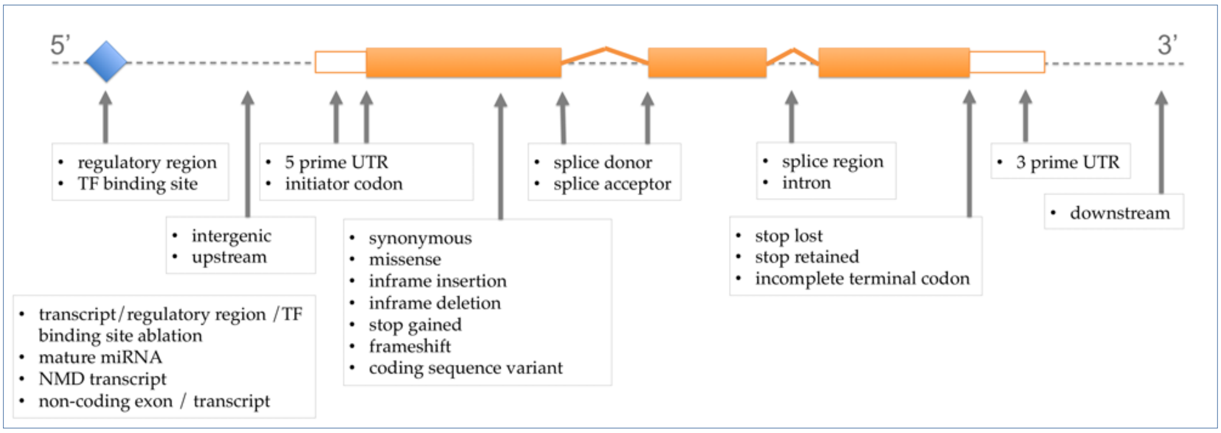
Distance	All items in this field are options, so the field could be empty. - Up/Downstream: Distance to first / last codon - Intergenic: Distance to closest gene - Distance to closest Intron boundary in exon (+/up/downstream). If same, use positive number. - Distance to closest exon boundary in Intron (+/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)
dbSNP142_ID	dbSNP rsNo.
1000G_AF	Non-reference allele frequency of existing variation in 1000 Genomes
1000G_AFR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined African population
1000G_AMR_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined American population
1000G_EAS_AF	Non-reference allele frequency of existing variation in 1000 Genomes combined East Asian population
1000G_EUR_MAF	Non-reference allele frequency of existing variation in 1000 Genomes combined European population
1000G_SAS_MAF	Non-reference allele frequency of existing variation in 1000 Genomes combined South Asian population
ESP6500[2]_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)
SIFT[3]_Score	SIFT score (SIFTori).. Scores range from 0 to 1. The smaller the score the more likely the SNP has damaging effect. Multiple scores separated by ";".
SIFT_Pred	If SIFTori is smaller than 0.05 (rankscore>0.55) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". Multiple predictions separated by ";".
PolyPhen2[4]_HDIV_Score	Polyphen2 score based on HumDiv, i.e. hdiv_prob. The score ranges from 0 to 1. Multiple entries separated by ";".

PolyPhen2_HDIV_Pred	Polyphen2 prediction based on HumDiv, "D" ("porobably damaging", HDIV score in [0.957,1] or rankscore in [0.52996,0.89917]), "P" ("possibly damaging", HDIV score in [0.453,0.956] or rankscore in [0.34412,0.52842]) and "B" ("benign", HDIV score in [0,0.452] or rankscore in [0.02656,0.34399]). Score cutoff for binary classification is 0.5 for HDIV score or 0.35411 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5 (rankscore is smaller than 0.35411), and "deleterious" if the HDIV score is larger than 0.5 (rankscore is larger than 0.35411). Multiple entries are separated by ";".
PolyPhen2_HVAR_Score	Polyphen2 score based on HumVar, i.e. hvar_prob. The score ranges from 0 to 1. Multiple entries separated by ";".
PolyPhen2_HVAR_Pred	Polyphen2 prediction based on HumVar, "D" ("porobably damaging", HVAR score in [0.909,1] or rankscore in [0.62955,0.9711]), "P" ("possibly damaging", HVAR in [0.447,0.908] or rankscore in [0.44359,0.62885]) and "B" ("benign", HVAR score in [0,0.446] or rankscore in [0.01281,0.44315]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45998 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5 (rankscore is smaller than 0.45998), and "deleterious" if the HVAR score is larger than 0.5 (rankscore is larger than 0.45998). Multiple entries are separated by ";".
CLINVAR(5)_CLNSIG	Variant Clinical Significance, 0 - Uncertain significance, 1 - not provided, 2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - drug response, 7 - histocompatibility, 255 - other
CLINVAR_CLNDSDB	Variant disease database name
CLINVAR_CLNDSDBID	Variant disease database ID
CLINVAR_CLNDBN	Variant disease name
CLINVAR_CLNREVSTAT	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
CLINVAR_CLNACC	Variant Accession and Versions

(1) 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:



REFERENCE http://asia.ensembl.org/info/genome/variation/consequences.jpg

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

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
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

(2) 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.

(3) 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'.

REFERENCE 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)
LINK doi:10.1038/nprot.2009.86

(4) PolyPhen2

PolyPhen-2(**Poly**morphism **Phen**otyping v**2**) predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam annotations, 3D structures from PDB where available, and a number of other databases and tools (including DSSP, ncoils etc.). The PolyPhen score represents the probability that a substitution is damaging, so values nearer to 1 are more confidently predicted to be deleterious (note that this the opposite to SIFT). The qualitative prediction is based on the False Positive Rate of the classifier model used to make the predictions.

REFERENCE Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR.
A method and server for predicting damaging missense mutations
Nature Methods 7(4):248-249 (2010)
LINK doi:10.1038/nmeth0410-248

(5) CLINVAR

ClinVar is a freely accessible, public 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.

Humanizing Genomics
macrogen

Macrogen Korea

10F, 254 Beotkkot-ro,
Geumcheon-gu, Seoul
Rep. of Korea
Phone : +82-2-2113-7100

Contact

Web : www.macrogen.com
Lims : http://dna.macrogen.com