Research use only

Reference
Whole Genome Resequencing

Report

November 2019

Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen
Company/Institution	Macrogen
Order Number	HN00000000
Species	*Reference*
Reference	NCBI
Library Kit	DNA Kit
Type of Read	Paired-ends
Read Length	101
Number of Samples	3
Type of Analysis	Whole Genome Resequencing
Type of Sequencer	Illumina platform
Comment	Reference 26695 Reference Download URL :

Summary of Project Result

In this study, whole genome resequencing of *Reference* was performed in order to identify variants, and perform gene annotation on useful genes based on database information.

Analyses were successfully performed on all 3 paired-end samples. Figure 1 shows the throughput of raw data. Figure 2 shows the Q20 and Q30 percentage (% of bases with quality over phred score 20, 30) of each sample's raw data.

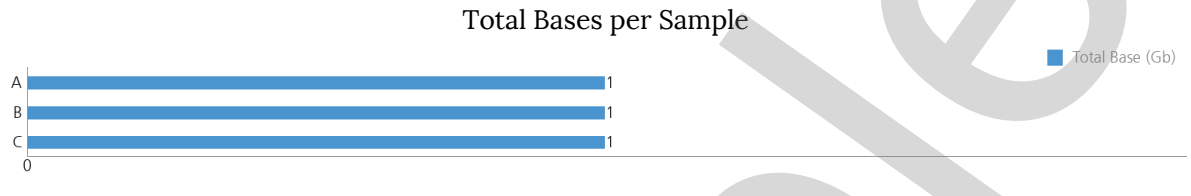


Figure 1.Throughput of Raw data

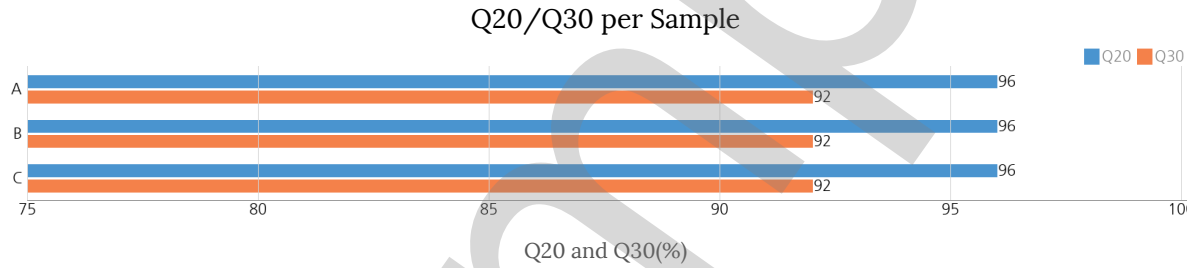


Figure 2. Q20/Q30 scores of Raw data

After filtering the data, reads in normal range were mapped to a reference genome using BWA. Figure 3 shows the overall read mapping ratio which is the ratio of mapped reads to total reads of the sample.

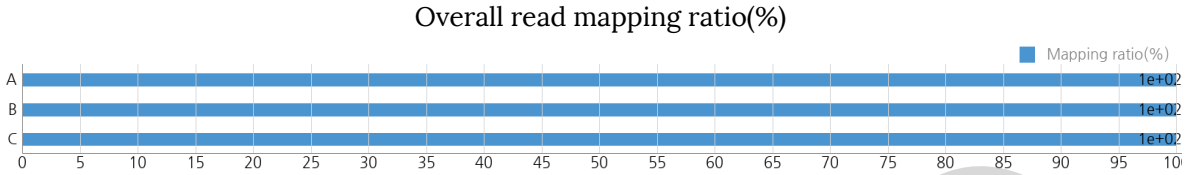


Figure 3. Overall read mapping ratio (%)

After mapping, Sambamba and SAMTools were respectively used to remove duplicated reads and identify variants.

Table 1. Project result summary

Library name	Ref.Length	Mapped site	Total read	Mapped read	Variant
A	1,000,000	1,000,000	1,000,000	1,000,000	61,000
B	1,000,000	1,000,000	1,000,000	1,000,000	61,000
C	1,000,000	1,000,000	1,000,000	1,000,000	61,000

- Library name : Sample name.
- Ref.Length : Length of reference genome.
- Mapped site : Length of mapped site in reference genome.
- Total read : Total number of reads.
- Mapped read : Total number of reads mapped to the reference genome.
- Variant : Number of variants (Insertions, deletions, SNPs) compared to reference genome.

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis results

Download link	File size	md5sum
A_1.fastq.gz	236M	84dc8c4dda0dd952f1ef6bd62ba4b5a0
A_2.fastq.gz	253M	bc5addb5603a833268e58bad6bc284ba
B_1.fastq.gz	490M	325df54516cee5c71e640175288e75bf
B_2.fastq.gz	574M	28c53cafcf3480a26aa4d5430a260831
B_filtered_1.fastq.gz	391M	cf3b00e54060ec828e79cf2c1e7a060a
B_filtered_2.fastq.gz	420M	356ca634c7761927d85b85c23af067a8
C_1.fastq.gz	276M	007a8c2e0820ed8811705afb6cfca0a1
C_2.fastq.gz	317M	85d11972128c8935276ec3f1ca55788d
C_filtered_1.fastq.gz	225M	af78796aa7392e2c9e4cf700732f3d4c
C_filtered_2.fastq.gz	241M	ab4e8bcb9997afbfa745ac6ca426d547
A_Analysis_Result.zip	379M	14b6fc69941be7d37aeff97dcb8b47cd
B_Analysis_Result.zip	640M	99270ef74d8021735e73acdd8feda5ea
C_Analysis_Result.zip	369M	cb457d91daf921911c2655025d0b4a61

fastq.gz - Compressed file of raw data.
filtered_1/2.fastq.gz - Compressed file of adapter trimmed data used in analysis.
md5sum - In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

1. 2. Details of File Extensions

Raw Data

File extensions	Details
*.fastq	File format typically used in NGS technology. Includes ID, sequence and quality value.

Alignment and Annotation results

File extensions	Details
*.xlsx	Result of variant calling.
*.bam	Binary format for a SAM(Sequence Alignment Map).
*.bai	Index format for BAM files
*_filtered.vcf	Text file format that contains filtered variants found at specific positions.
*_annotated.vcf	Text file format that contains filtered variants found at specific positions with SnpEff annotation.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please email (ngskr@macrogen.com) or contact our sales team.

2. Experimental Methods and Workflow

2. 1. Overview

Figure 4. Workflow overview

Sequencing

1) Sample Prep. (Sample Preparation)

For library construction, DNA/RNA is extracted from a sample. After performing quality control(QC), passed sample is proceeded with the library construction.

2) Library Construction

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

3) 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.

4) Raw data

Sequencing data is converted into raw data for the analysis.

Preprocessing

1) Quality Control

After sequencing, the raw reads undergo quality control. Overall quality of reads generated, total number of bases, reads, GC content and basic statistics are calculated.

2) Preprocessing

In order to reduce biases in analysis, adapter trimming and quality filtering are performed. The quality of filtered reads, total bases, total reads, GC (%) and basic statistics are calculated again.

Analysis

1) Mapping

The filtered reads are mapped to a reference genome. In this process, sufficient read depth is required for a more accurate analysis. After mapping, duplicated reads are removed.

2) Variant Analysis

Variants (SNPs and short indels) are identified by analyzing the information taken from aligned reads.

3) Annotation

The variants are classified by each chromosomes or scaffolds, and the information of the location is marked.

3. Summary of Produced Data

3. 1. Raw data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) were calculated for the 3 samples. For example, in A, 1,000,000 reads were produced, and total read bases are 1000.0M bp. The GC content is 40.00% and Q30 is 92.00%.

Table 2. Raw data Stats

Library name	Total read bases (bp)	Total reads	GC(%)	Q20(%)	Q30(%)
A	1,000,000,000	1,000,000	40.00	96.00	92.00
B	1,000,000,000	1,000,000	40.00	96.00	92.00
C	1,000,000,000	1,000,000	40.00	96.00	92.00

- Library name : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. In Illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- GC(%) : GC 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.

3. 2. Filtered data Statistics

Trimmomatic was used to remove adapter sequences and low quality reads in order to reduce biases in analysis. The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) were calculated for the 3 samples after filtering.

Table 3. Filtered data Stats

Library name	Total read bases (bp)	Total reads	GC(%)	Q20(%)	Q30(%)
A	100,000,000	1,000,000	40.00	99.00	96.00
B	100,000,000	1,000,000	40.00	99.00	96.00
C	100,000,000	1,000,000	40.00	99.00	96.00

- Library name : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. In Illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- GC(%) : GC 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.

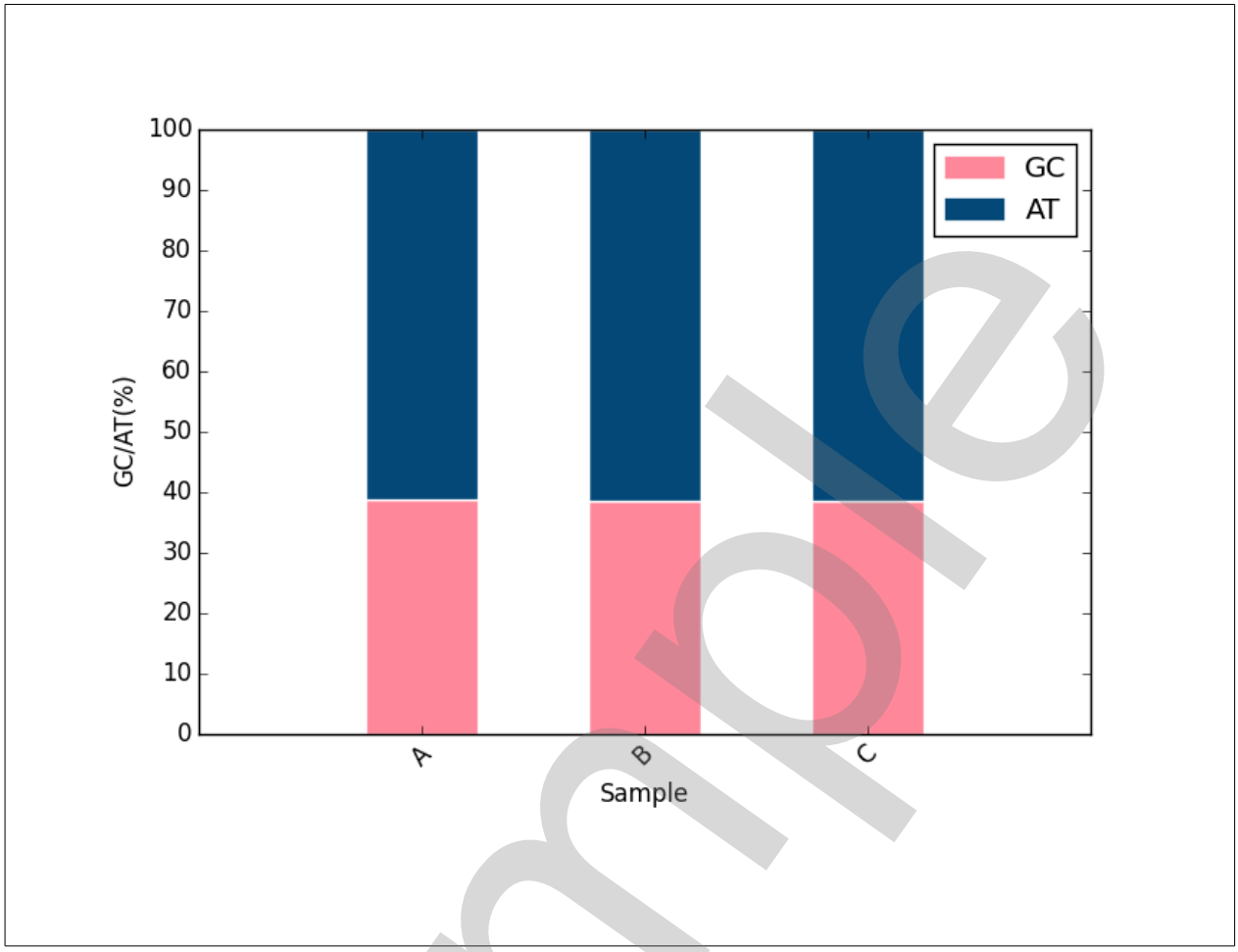
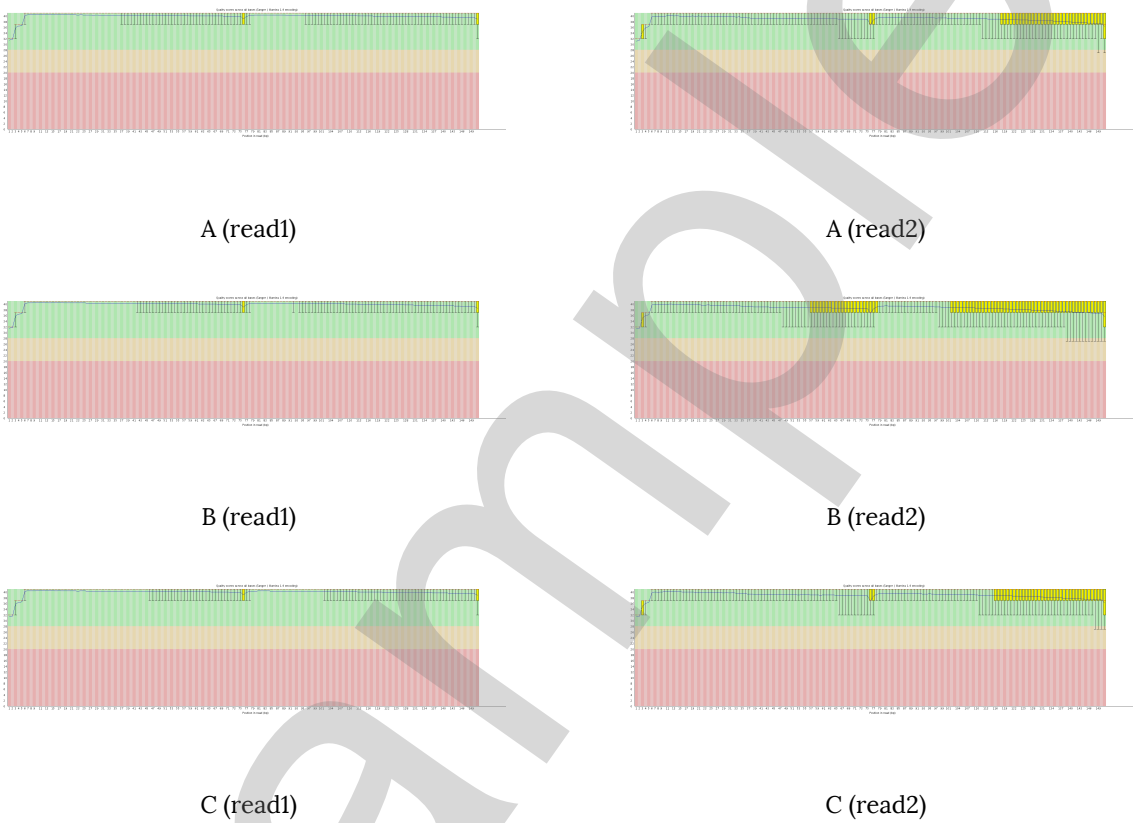


Figure 5. GC content (%)

3. 3. Average Base Quality at Each Cycle

The 'per base sequence quality' plot generated by FastQC was used to check the overall quality of the produced data. This plot shows the average quality at each cycle.

The x-axis and y-axis are respectively the number of cycles, and phred quality score. Phred quality score of 20 means 99% accuracy and reads with quality score over 20 are generally accepted as good quality reads.



A (read1) A (read2)

B (read1) B (read2)

C (read1) C (read2)

Figure 6. Read quality at each cycle of samples

- Yellow box : Interquartile range (25-75%) of phred score at each cycle.
- Red line : Median phred score at each cycle.
- Blue line : Average phred score at each cycle.
- Upper & Lower whiskers : Point of 10% and 90%.
- Green background : Good quality.
- Orange background : Acceptable quality.
- Red background : Bad quality.

4. Reference Mapping Results

4. 1. Mapping Data Statistics

In order to map the reads obtained from sequencing, *Reference* was used as a reference genome. Table 4 shows the statistic obtained from BWA. You can check the number of mapped sites per sample, mapping coverage, the total number of reads, number of mapped reads, overall mapping ratio, number of mapped bases, and the average alignment depth.

Table 4. Mapped data Stats

Library name	Ref.Length	Mapped Sites (>=1x)	Total Reads	Mapped Reads	Mapped Bases	Mean Depth
A	1,000,000	1,000,000 (100.00%)	100,000	10,000 (100.00%)	10,000,000	100
B	1,000,000	1,000,000 (100.00%)	100,000	10,000 (100.00%)	10,000,000	100
C	1,000,000	1,000,000 (100.00%)	100,000	10,000 (100.00%)	10,000,000	100

- Library name : Sample name.
- Ref.Length : Length of reference genome.
- Mapped Sites : Length of mapped site.
- Total Reads : Number of total read.
- Mapped Reads : Number of reads mapped to the reference.
- Mapped Bases : Number of bases in reads mapped to the reference.
- Mean Depth : Average alignment depth.

4. 2. SNP and Indel Discovery

4. 2. 1. Variants Count

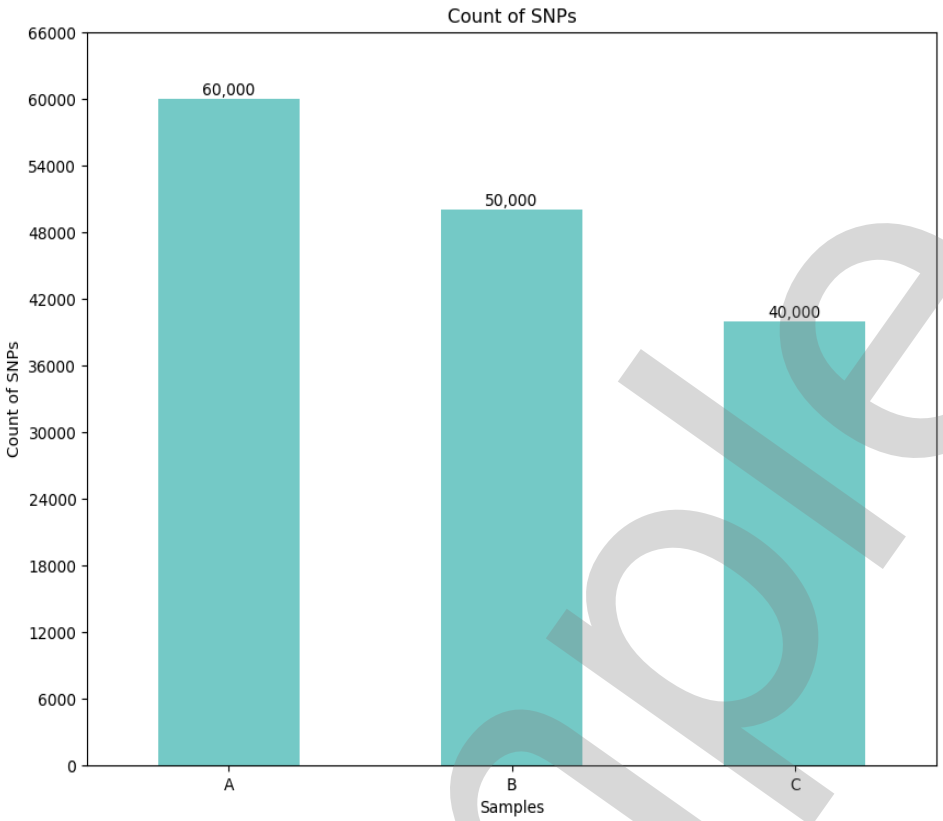
Produced mass sequence data were used to search for genetic variation. In this analysis, the reference genome is based on NCBI.

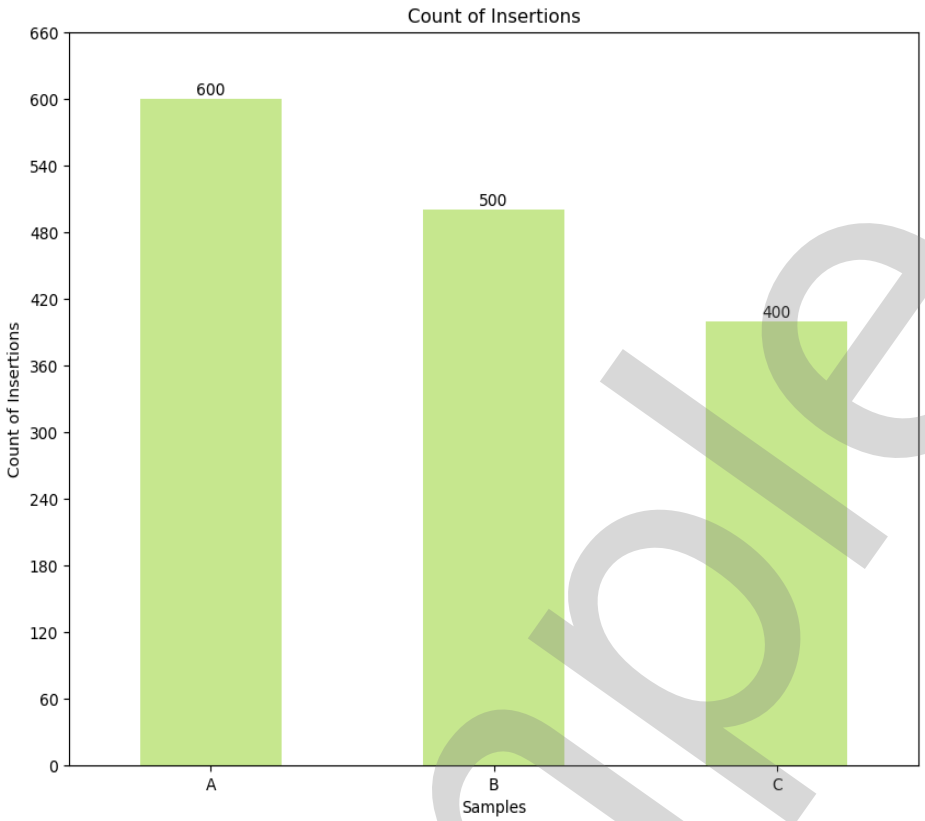
After removing duplicates with Sambamba and identifying variants with SAMTools, information of each variants were gathered and classified by chromosomes or scaffolds. Table 5 shows the summary of variant calling for the 3 samples.

Table 5. Summary of Variant Calling

Library name	Number of SNPs	Number of insertions	Number of deletions
A	60,000	600	600
B	50,000	500	500
C	40,000	400	400

- Library name : Sample name.
- Number of SNPs : Number of SNPs in sample.
- Number of insertions : Number of insertions in sample.
- Number of deletions : Number of deletions in sample.


Count of SNPs
66000
60000
54000
48000
42000
36000
30000
24000
18000
12000
6000
0
Count of SNPs
60,000
50,000
40,000
A
B
C
Samples


Count of Insertions
660
600
540
480
420
360
300
240
180
120
60
0
Count of Insertions
600
500
400
A
B
C
Samples

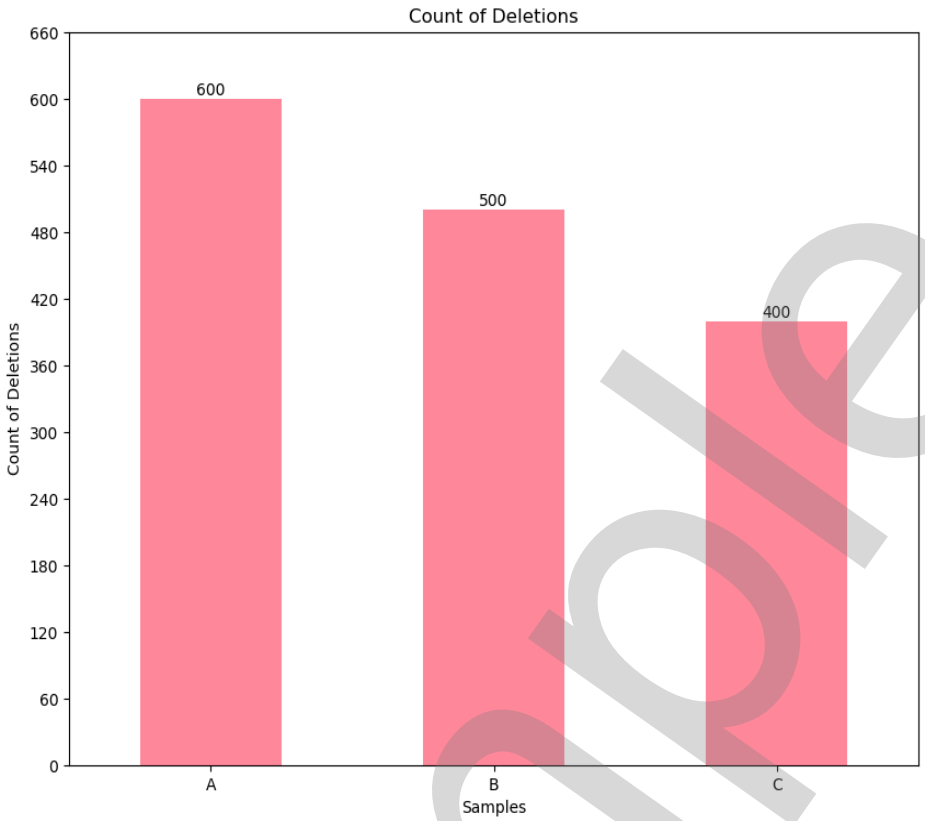


Figure 7. SNP/Insertion/Deletion Count

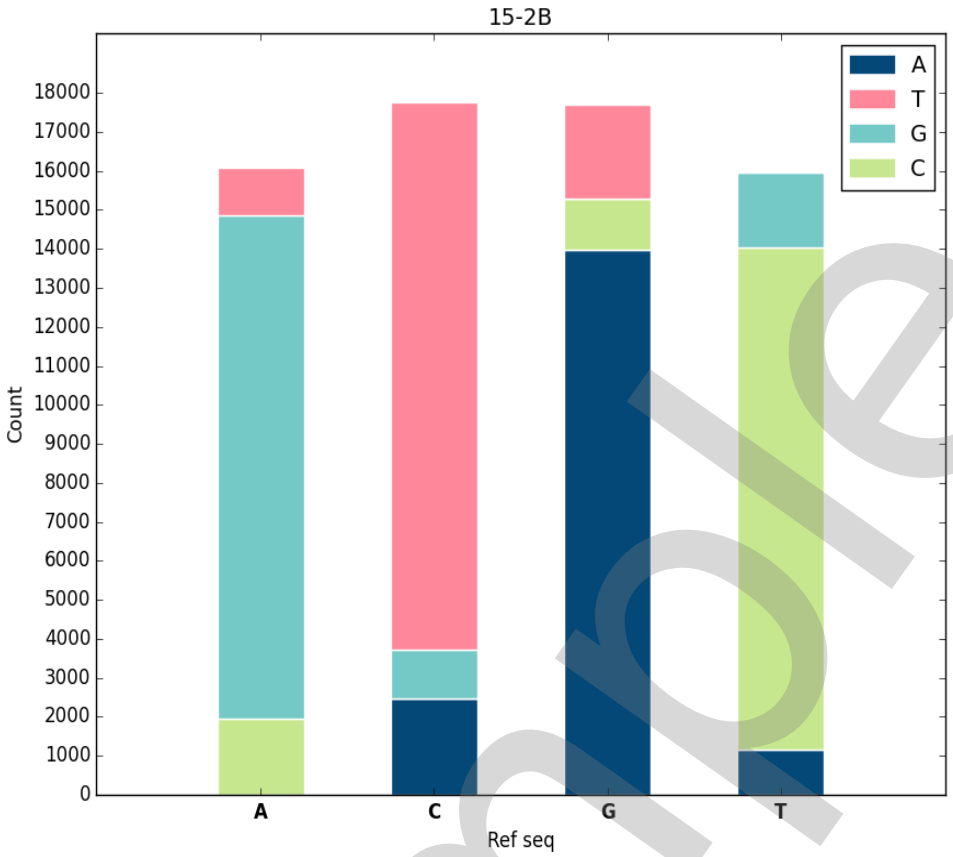
4. 2. 2. Base Change Count

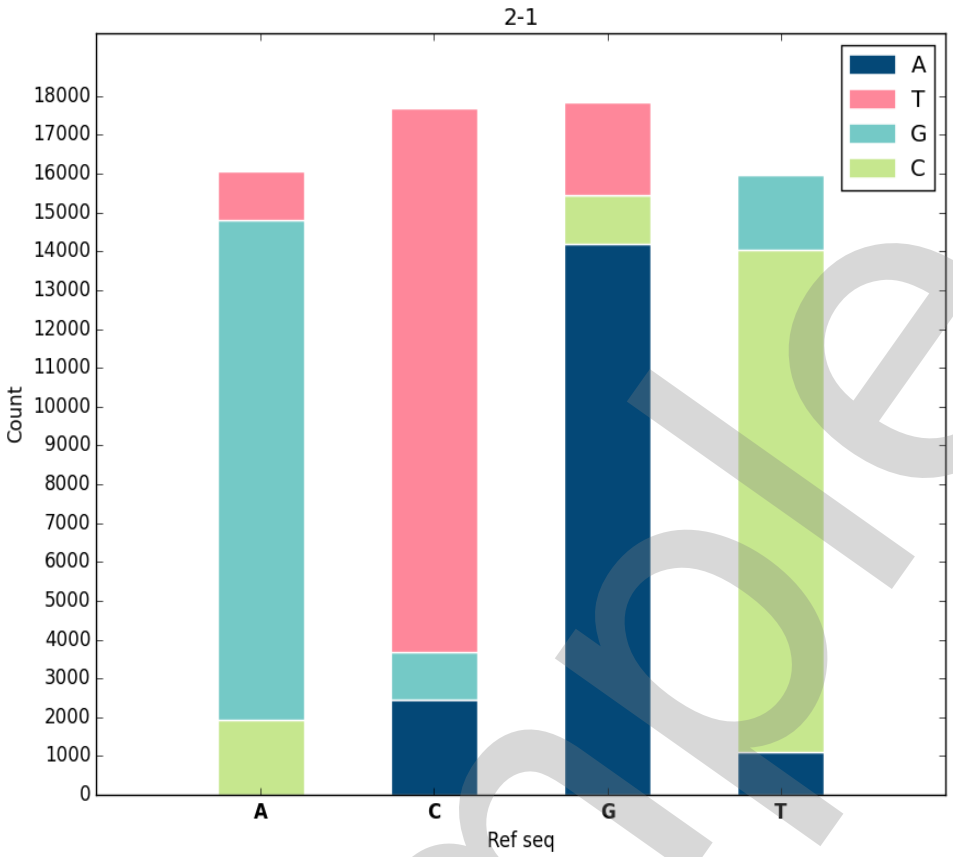
Table 6 shows the base change count on every SNPs. And the results are visualized in Figure 8.

Table 6. Base change count

Library name	Ref	A			C		
	Alt	T	G	C	A	T	G
A		1,200	13,000	2,000	2,500	14,000	1,200
B		1,200	13,000	2,000	2,500	14,000	1,200
C		1,200	13,000	2,000	2,500	14,000	1,200

Library name	Ref	G			T		
	Alt	A	T	C	A	G	C
A		14,000	2,500	1,300	1000	2,000	13,000
B		14,000	2,500	1,300	1000	2,000	13,000
C		14,000	2,500	1,300	1000	2,000	13,000


15-2B
A
T
G
C
Count
18000
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
A
C
G
T
Ref seq


2-1
A
T
G
C
18000
17000
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Count
A
C
G
T
Ref seq

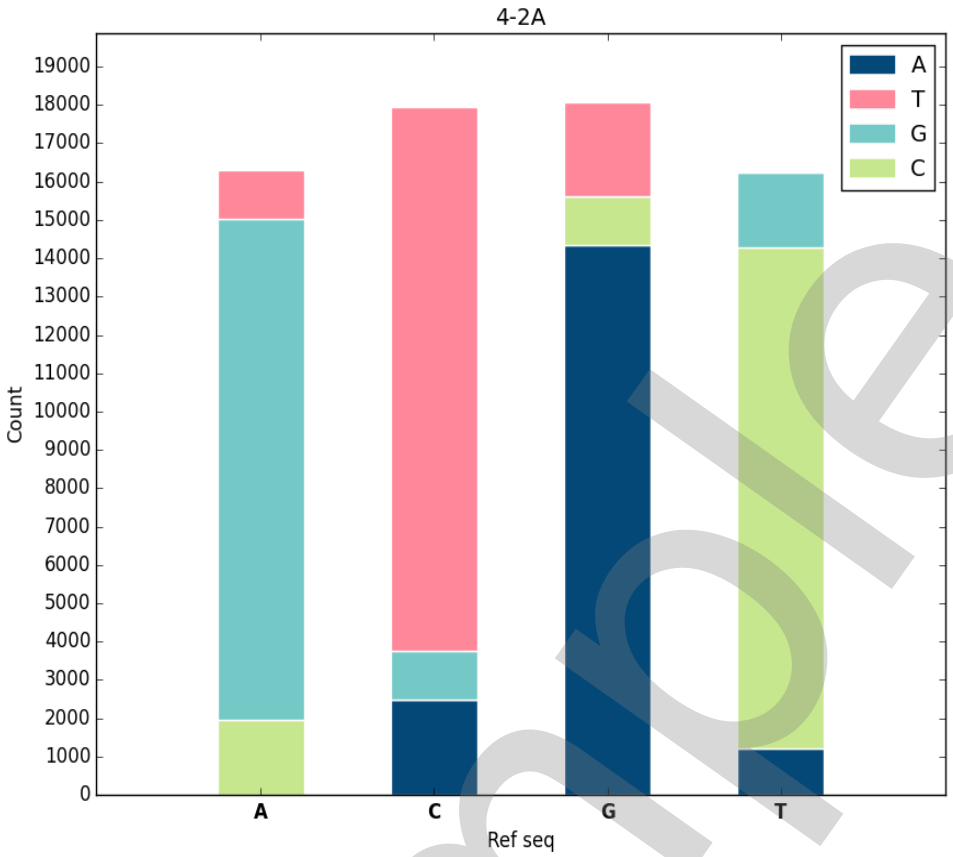


Figure 8. Base change count of each sample

4. 2. 3. Transition and Transversion Information

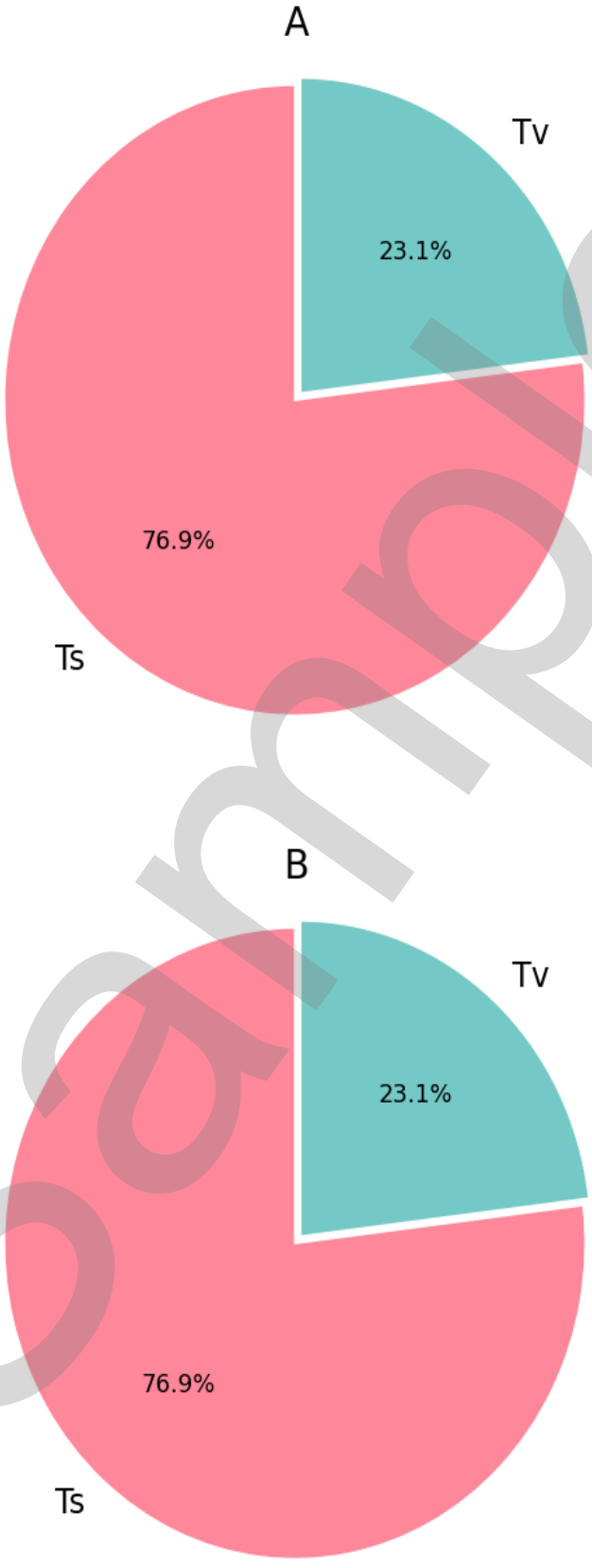
The number of transition (Ts) and transversion (Tv), and the Ts/Tv ratio were calculated using the base change count.

Base changes (DNA substitution) are of two types. Interchanges of purines (A <-> G), or of pyrimidines (C <-> T) are transitions, while interchanges of purine for pyrimidine bases, and vice versa, are transversions. Although there are twice as many possible transversions, transitions are more common than transversions due to difference in structural characteristics.

Generally, transversions are more likely to cause amino acid sequence changes.

Table 7. Transition, Transversion information table

Library name	Total SNP count	Transition	Transversion	Ts/Tv
A	30,000	20,000	6,000	4.00 %
A	30,000	20,000	6,000	4.00 %
A	30,000	20,000	6,000	4.00 %



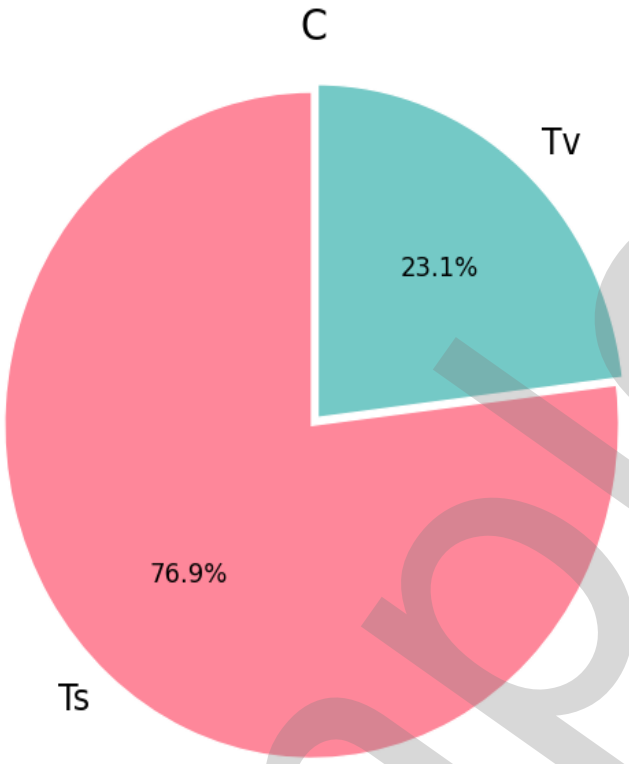


Figure 9. Transition, Transversion proportion

4. 3. Variant Annotation

In order to find out the annotation information such as amino acid changes by variants, SnpEff was used. Because genes usually have multiple transcripts, a single variant can have different effects on different transcripts. Table 8 and 9 shows the number of variants per type (based on the representative transcript), and brief explanations about the variant type, respectively. In Table 8, top 10 types of annotations are shown.

Table 8. Annotation type count

Library name	Type of annotation	Count	Ratio
A	synonymous_variant	500	50.0%
	missense_variant	300	30.19%
	frameshift_variant	200	20.31%
	stop_gained	20	5.22%
	splice_region_variant & stop_retained_variant	5	1.12%
	conservative_inframe_insertion	1	0.1%
	intragenic_variant	1	0.09%
	non_coding_transcript_exon_variant	1	0.06%
	disruptive_inframe_insertion	1	0.06%
	downstream_gene_variant	1	0.05%
B	synonymous_variant	500	50.0%
	missense_variant	300	30.19%
	frameshift_variant	200	20.31%
	stop_gained	20	5.22%
	splice_region_variant & stop_retained_variant	5	1.12%
	conservative_inframe_insertion	1	0.1%
	intragenic_variant	1	0.09%
	non_coding_transcript_exon_variant	1	0.06%
	disruptive_inframe_insertion	1	0.06%
	downstream_gene_variant	1	0.05%

C	synonymous_variant	500	50.0%
	missense_variant	300	30.19%
	frameshift_variant	200	20.31%
	stop_gained	20	5.22%
	splice_region_variant & stop_retained_variant	5	1.12%
	conservative_inframe_insertion	1	0.1%
	intragenic_variant	1	0.09%
	non_coding_transcript_exon_variant	1	0.06%
	disruptive_inframe_insertion	1	0.06%
	downstream_gene_variant	1	0.05%

Table 9. Annotation type information

Type of annotation	Description	Impact
coding_sequence_variant	The variant hits a CDS.	MODIFIER
chromosome	A large part (over 1% or 1,000,000 bases) of the chromosome was deleted.	HIGH
duplication	Duplication of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
inversion	Inversion of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
coding_sequence_variant	One or many codons are changed.	LOW
inframe_insertion	One or many codons are inserted (e.g.: An insert multiple of three in a codon boundary).	MODERATE
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).	MODERATE
inframe_deletion	One or many codons are deleted (e.g.: A deletion multiple of three at codon boundary).	MODERATE
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).	MODERATE
downstream_gene_variant	Downstream of a gene (default length: 5K bases).	MODIFIER
exon_variant	The variant hits an exon (from a non-coding transcript) or a retained intron.	MODIFIER
exon_loss_variant	A deletion removes the whole exon.	HIGH
exon_loss_variant	Deletion affecting part of an exon.	HIGH
duplication	Duplication of an exon.	HIGH
duplication	Duplication affecting part of an exon.	HIGH
inversion	Inversion of an exon.	HIGH
inversion	Inversion affecting part of an exon.	HIGH
frameshift_variant	Insertion or deletion causes a frame shift (e.g.: An indel size is not multple of 3).	HIGH
gene_variant	The variant hits a gene.	MODIFIER
feature_ablation	Deletion of a gene.	HIGH
duplication	Duplication of a gene.	MODERATE
gene_fusion	Fusion of two genes.	HIGH
gene_fusion	Fusion of one gene and an intergenic region.	HIGH
bidirectional_gene_fusion	Fusion of two genes in opposite directions.	HIGH
rearranged_at_DNA_level	Rearrengment affecting one or more genes.	HIGH
intergenic_region	The variant is in an intergenic region.	MODIFIER

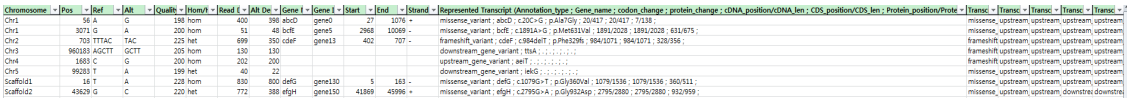
conserved_intergenic_variant	The variant is in a highly conserved intergenic region.	MODIFIER
intragenic_variant	The variant hits a gene, but no transcripts within the gene.	MODIFIER
intron_variant	Variant hits and intron. Technically, hits no exon in the transcript.	MODIFIER
conserved_intron_variant	The variant is in a highly conserved intronic region.	MODIFIER
miRNA	Variant affects an miRNA.	MODIFIER
missense_variant	Variant causes a codon that produces a different amino acid (e.g.: Tgg/Cgg, W/R).	MODERATE
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))	LOW
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon (the new codon produces a different AA). (e.g.: Atg/Ctg, M/L (ATG and CTG can be START codons))	LOW
protein_protein_contact	Protein-Protein interacion loci.	HIGH
structural_interaction_variant	Within protein interacion loci (e.g. two AA that are in contact within the same protein, prossibly helping structural conformation).	HIGH
rare_amino_acid_variant	The variant hits a rare amino acid thus is likely to produce protein loss of function..	HIGH
splice_acceptor_variant	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	HIGH
splice_donor_variant	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	HIGH
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.	LOW
splice_region_variant	A variant affective putative (Lariat) branch point, located in the intron.	LOW
splice_region_variant	A variant affective putative (Lariat) branch point from U12 splicing machinery, located in the intron.	MODERATE
stop_lost	Variant causes stop codon to be mutated into a non-stop codon (e.g.: Tga/Cga, */R).	HIGH
5_prime_UTR_premature start_codon_gain_variant	A variant in 5'UTR region produces a three base sequence that can be a START codon.	LOW

start_lost	Variant causes start codon to be mutated into a non-start codon (e.g.: aTg/aGg, M/R).	HIGH
stop_gained	Variant causes a STOP codon (e.g.: Cag/Tag, Q/*).	HIGH
synonymous_variant	Variant causes a codon that produces the same amino acid (e.g.: Ttg/Ctg, L/L).	LOW
start_retained	Variant causes start codon to be mutated into another start codon (e.g.: Ttg/Ctg, L/L (TTG and CTG can be START codons)).	LOW
stop_retained_variant	Variant causes stop codon to be mutated into another stop codon (e.g.: taA/taG, */*).	LOW
transcript_variant	The variant hits a transcript.	MODIFIER
feature_ablation	Deletion of a transcript.	HIGH
regulatory_region_variant	The variant hits a known regulatory feature (non-coding).	MODIFIER
upstream_gene_variant	Upstream of a gene (default length: 5K bases).	MODIFIER
3_prime_UTR_variant	Variant hits 3'UTR region.	MODIFIER
3_prime_UTR_truncation + exon_loss	The variant deletes an exon which is in the 3'UTR of the transcript.	MODERATE
5_prime_UTR_variant	Variant hits 5'UTR region.	MODIFIER
5_prime_UTR_truncation + exon_loss_variant	The variant deletes an exon which is in the 5'UTR of the transcript.	MODERATE

* SnpEff reports putative variant impact in order to make it easier and faster to categorize and prioritize variants. However, impact categories must be used with care as they were created only to help and simplify the filtering process. Obviously, there is no way to predict whether a HIGH impact or a LOW impact variant is the one producing a phenotype of interest.

- Type of annotation : Sequence ontology which allows to standardize terminology used for assessing sequence changes and impact.
- Description : Detailed description of the effect (annotation).
- Impact : Effects are categorized by 'impact': {High, Moderate, Low, Modifier}. These are pre-defined categories to help users find more significant variants.
  - HIGH : The variant is assumed to have high (disruptive) impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.
  - MODERATE : A non-disruptive variant that might change protein effectiveness.
  - LOW : Assumed to be mostly harmless or unlikely to change protein behavior.
  - MODIFIER : Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.

The results of variant calling are provided in excel file (See below for example).



Chromosome	Pos	Ref	Alt	Qualit	Hom/t	Read E	Alt De	Gene I	Gene I	Start	End	Strand	Represented Transcript (Annotation_type ; Gene_name ; codon_change ; protein_change ; cDNA_position/cDNA_len ; CDS_position/CDS_len ; Protein_position/Prote	Transc	Transc	Transci	Transc	Transc
Chr1	56	A	G	198	hom	400	398	abcD	gene0	27	1076	+	missense_variant ; abcD ; c.20C>G ; p.Ala7Gly ; 20/417 ; 20/417 ; 7/138 ;	missense_ upstream; upstream, upstream, upstream				
Chr1	3071	G	A	200	hom	51	48	bcfE	gene5	2968	10069	-	missense_variant ; bcfE ; c.1891A>G ; p.Met631Val ; 1891/2028 ; 1891/2028 ; 631/675 ;	missense_ upstream; upstream, upstream, upstream				
Chr2	703	TTTAC	TAC	225	het	699	350	cdeF	gene13	402	707	-	frameshift_variant ; cdeF ; c.984delT ; p.Phe329fs ; 984/1071 ; 984/1071 ; 328/356 ;	frameshift upstream; upstream, upstream, upstream				
Chr3	960183	AGCTT	GCTT	205	hom	130	130						downstream_gene_variant ; ttsA ; . ; . ; . ; . ; .	frameshift upstream; upstream, upstream, upstream				
Chr4	1683	C	G	200	hom	202	200						upstream_gene_variant ; aeiT ; . ; . ; . ; . ; .	frameshift upstream; upstream, upstream, upstream				
Chr5	99283	T	A	199	het	40	22						downstream_gene_variant ; ntkG ; . ; . ; . ; . ; .	missense_ upstream; upstream, upstream, upstream				
Scaffold1	16	T	A	228	hom	830	800	defG	gene130	5	163	-	missense_variant ; defG ; c.1079G>T ; p.Gly360Val ; 1079/1536 ; 1079/1536 ; 360/511 ;	missense_ upstream; upstream, upstream, upstream				
Scaffold2	43629	G	C	220	het	772	388	efgH	gene150	41869	45996	+	missense_variant ; efgH ; c.2795G>A ; p.Gly932Asp ; 2795/2880 ; 2795/2880 ; 932/959 ;	missense_ upstream; upstream, downstrea downstrea				

Figure 10. Example of variant calling result

- Chromosome : Chromosome name.
- Pos : Position information of target variant.
- Ref : Reference sequence regarding specific position.
- Alt : DNA sequence of the sample.
- Quality : Phred-scaled probability of all samples being homozygous reference. The value is in -log. The smaller the value, the more likely ALT is wrong.
- Hom/Het : Indicates the genotype. "hom" refers to non-reference homozygote, while "het" refers to heterozygote.
    - Homozygous : The circumstances when there are mutations on most reads that are mapped to certain region.
    - Heterozygous : The circumstances when there are mutations on some reads that are mapped to certain region.
- Read Depth : Total read depth.
- Alt Depth : Allelic depths for the ref and alt alleles in the order listed.
- Gene Name, GeneID : Gene name and gene symbol.
- Start, End : Position information of target gene.
- Strand : Strand information of target gene.
- Transcript : The results of functional annotation by transcripts. Type of variant (syn/nonsyn), protein change, and etc. can be ascertained in this section. A representative transcript is chosen by the gene name obtained from variant calling analysis. Other transcripts are chosen by information of neighboring genes which are close enough.
    - It is not uncommon for a gene to have more than one transcript. A variant might affect different transcripts in different ways, as a result of different reading frames.

5. Appendix

5. 1. FAQ

Q: I want to see the produced data. How can I open the files?

A: As the large size zip files provided by our company are hard to process in the Windows environment, we highly recommend using Linux environment for a smoother operation.

5. 2. FASTQ File

Example of FASTQ

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNNTNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIIII#3AC##########################
```

FASTQ file is composed of four lines.
Line 1 : ID line includes information such as flow cell lane information.
Line 2 : Sequences line.
Line 3 : Separator line (+ mark).
Line 4 : Quality values line about sequences.

5. 3. Phred Quality Score Chart

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.

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

Encoding: Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

5. 4. Programs used in Analysis

5. 4. 1. FastQC

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC (V0.11.6) is a quality checking tool for high throughput sequencing data. The condition of the data can be checked through the various modules provided by the tool. Among the modules, 'Per base sequence quality' and 'Per tile sequence quality 'modules are commonly used to validate whether the data can be used for analysis.

5. 4. 2. Trimmomatic

LINK http://www.usadellab.org/cms/?page=trimmomatic

Incomplete removal of adapter sequences from NGS data can ultimately affect the accuracy of analysis. To avoid this, Trimmomatic (v0.36) is used to remove adapter sequences. Depending on the library type used for data production, appropriate adapter sequence is used to remove the said sequences from the data. In addition to removing adapter sequences, Trimmomatic trims out bases of low quality.

5. 4. 3. BWA (Burrows-Wheeler Aligner)

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

BWA (v0.7.17) is a tool which uses BWT (Burrows-Wheeler Transform) algorithm for aligning reads to references sequences. Three algorithms are present in BWA, namely: BWA-backtrack, BWA-SW, BWA-MEM. Among these Macrogen uses BWA-MEM for alignment. BWA-MEM is optimized for aligning reads greater than 70bp, and has an advantage of providing split alignment. The said characteristics can be found in BWA-SW as well, however BWA-MEM algorithm is better in terms of speed and accuracy.

5. 4. 4. Sambamba

LINK http://lomereiter.github.io/sambamba/

Sambamba (v0.6.7) is a tool written in D programming language, which makes use of SAM/BAM files for various analyses. During mapping, duplicated reads can falsely cause erroneous data to stand out. To prevent this Sambamba is used to remove duplicated reads. Duplicated reads are identified using mapping information such as start position, and CIGAR string.

5. 4. 5. SAMTools

LINK http://samtools.sourceforge.net/

SAMTools is used to manipulate the SAM/BAM files that come out as a result from mapping. In resequencing analysis, it is especially used for finding out variant information by calculating genotype likelihood from every position within the sample of analysis.

5. 4. 6. SnpEff

LINK http://snpeff.sourceforge.net/

SnpEff (v4.3t) is a tool for annotating possible effects (on genes) that can be caused by variants identified through mapping. Not only does it have the advantage of predicting whether the variant is synonymous or nonsynonymous based on information taken from the reference sequence, it can also predict changes in amino acids caused by the variant.

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)

HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

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

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 31, 1105 AZ, 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

3F Kyoto University International Science Innovation Bldg.
36-1 Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
Tel: +81-75-746-2773
Email: customer@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

Humanizing Genomics
macrogen