eurofins Genomics

Data Analysis Report: Variant Analysis v1.2

Project / Study: GATC-Demo

Date: February 28, 2018

Table of Contents

1 Analysis workflow

The schematic diagram of the data analysis steps that have been performed is shown in figure 1.

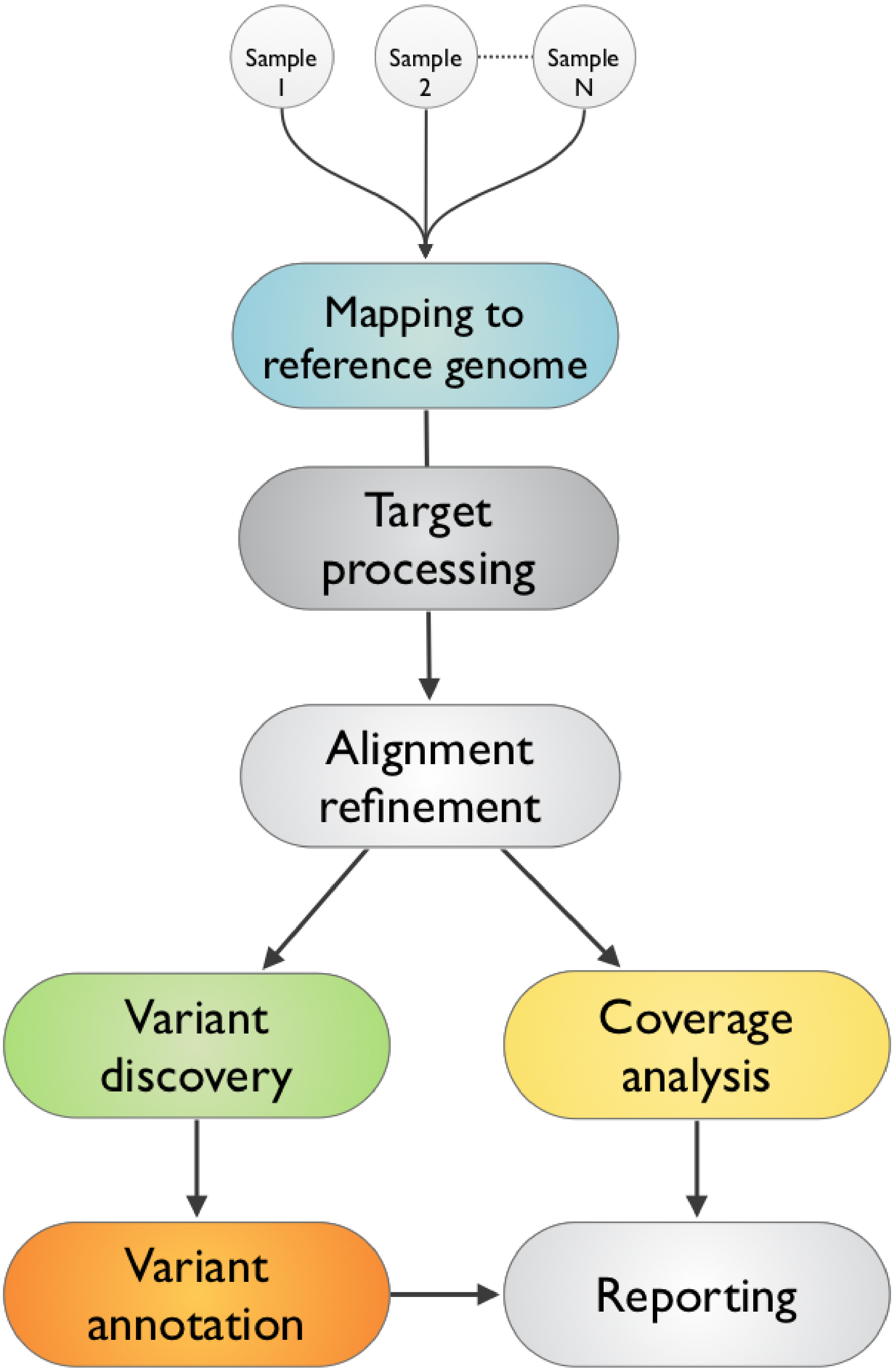

Figure 1: Variant Analysis v1.2 Workflow

2 Samples Analysed

NA12878_v6_R1.

3 Reference Database

Table 1: Information about the Homo sapiens Reference Database.

Tag	Description
Name	Homo sapiens
Version	hg19.chronly
Source	UCSC
Size (bp)	3.095 GB
Sequences	23

Table 2: Information about additional reference data used.

Type	Version	Source
Annotation	19	GENCODE
dbSNP	138	NCBI

Table 3: Information about the target region used.

Tag	Description
Name	SureSelect All Exon V6
Size (bp)	60,700,153
Source	Agilent

4 Results

4.1 Sequence Quality Metrics

The base quality of each sequence read is inspected. Low quality calls are removed before proceeding with further processing. Using a sliding window approach, bases with low quality are removed from the 3' and 5' ends. Bases are removed if the average phred quality is below 15. Finally only mate pairs (forward and reverse read) were used for the next analysis step. The total amount of raw sequence data and the results of the quality filtering is collected and reported in the following table.

Table 4: Sequence quality metrics per sample

Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
NA12878_v6_R1	131,879,166	1,897,496 (1.4%)	1,620,320 (1.2%)	128,361,350 (97.3%)

Total Reads: Total number of sequence reads analysed for each sample.

LQ Reads: Number (percentage) of low quality reads.

Single Reads: High quality reads without mates (2nd read). These are not included for further analysis.

HQ Reads: Number (percentage) of high quality reads used for further analysis.

4.2 Alignment Metrics

Mapping to the reference sequence/database is done using BWA[1] with default parameters. The following table contains the number of reads mapped to the reference for each sample. Please note that the mapping efficiency depends on the accuracy of the reference and the quality of sequence reads.

Table 5: Mapped read metrics observed per sample

Sample Name	HQ Reads	Mapped Reads
NA12878_v6_R1	128,361,350	128,067,005 (99.77%)

4.3 Alignment Classification

The alignment classification table includes the following read categories:

- Mapped: Reads mapped to reference.
- Unique: Reads mapped to exactly one site on the reference.
- Non-unique: Reads mapped to more than one site on the reference.
- Singletons: Mapped reads without mates (read not paired).
- Cross-Contig: Read pairs with the mate mapped to a different contig.
- On target: Reads mapped to target +/- 100 bp extension.

Percentage of reads in categories **Non-unique, Unique, Singletons, Cross-Contig** are calculated based on the number of reads mapping to entire reference.
Percentage of reads in category **On target** is calculated based on the number of reads mapped uniquely (excluding **Singletons** and **Cross-Contig** - if any).

Table 6: Read metrics for NA12878_v6_R1.

Read category	NA12878_v6_R1
Mapped	128,067,005
Unique	121,902,847 (95.19%)
Non-unique	6,164,158 (4.81%)
Singletons	17,155 (0.01%)
Cross-Contig	3,073,997 (2.40%)
On target	106,482,872 (89.62%)

Reads in categorie(s) **Non-unique** , **Singletons** and **Cross-Contig** are excluded from analysis.

4.4 Alignment Refinement Metrics

The removal of PCR duplicates is done using Picard[2] in order to remove the artificial coverage brought on by the PCR amplification step during the library preparation. If a read maps to the same genomic location and has same orientation as the read already mapped it is considered as duplicated. For paired-end, both reads should fulfill the criteria in order to designate as PCR duplicate. One copy of the duplicate read pair is kept in the alignment.

Local realignment serves to transform regions with misalignments due to indels into clean reads containing a consensus indel suitable for standard variant discovery approaches. GATK is used for this purpose.

The goal of Base Quality Recalibration is to improve the base quality score of reads for downstream processing and also correct for error covariates like machine cycle and dinucleotide context. A base quality score represents the probability of a particular base mismatching the reference genome. After recalibration quality scores are more accurate in that they are closer to the true probability of mismatch. This process is achieved by analyzing the covariation among several different features of a base. The reported quality score, sequencing cycle, and sequencing context are considered for this step. GATK modules are used for achieving this.

The following table contains the number of high-quality reads after read mapping, alignment and refinement.

Table 7: HQAligned reads per sample

Sample Name	Input Reads	Duplicate Reads	HQ Reads
NA12878_v6_R1	106,482,872	16,218,132 (15.23%)	90,264,740 (84.77%)

4.5 Coverage Report

The coverage plot showing the base coverage distribution from the HQ aligned data. Depth of coverage is plotted on X-axis and the percentage of the respective reference covered is plotted on Y-axis. The coverage plot is restricted to the target region without extension. The shape of the curve defines the uniformity of the reference coverage in the samples analysed.

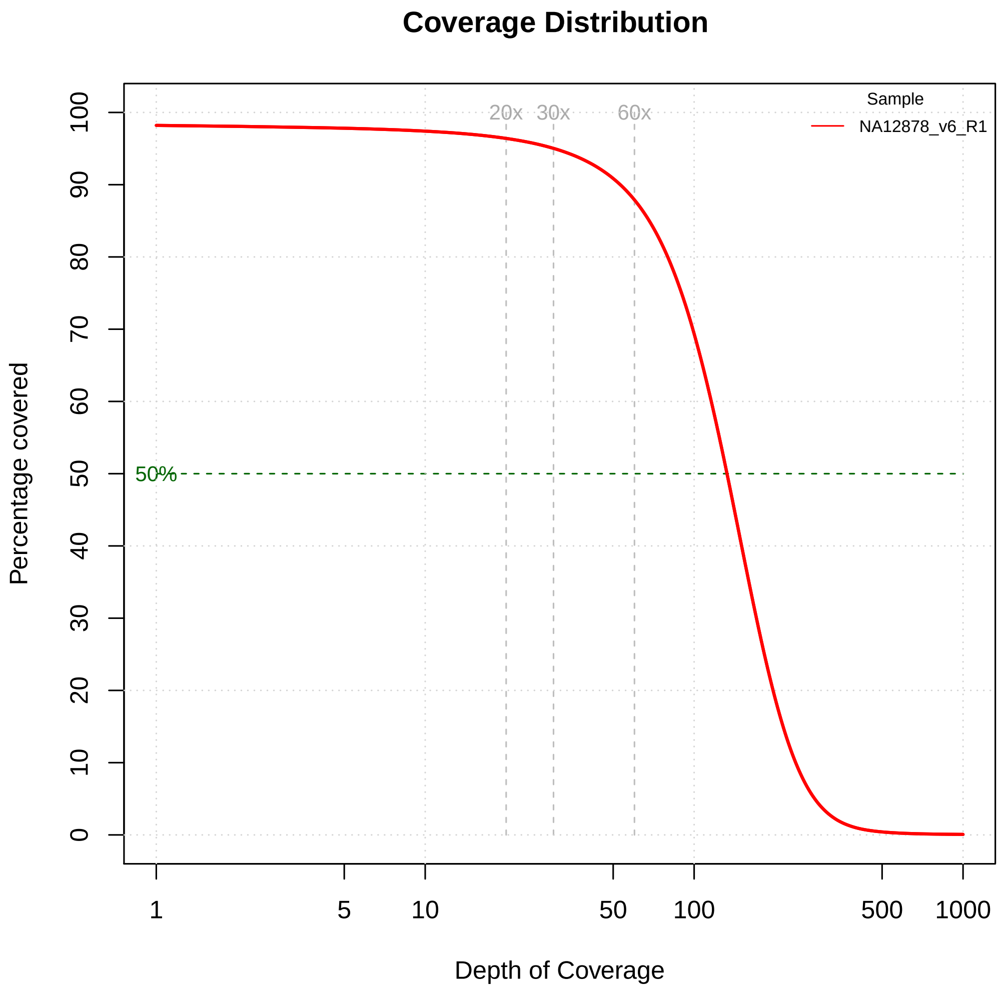

Figure 2: Coverage plot (excluding duplicated fragments).

Table 8: Depth of coverage summary (excluding duplicated fragments).

	target coverage		% of target covered with at least							
sample	total bases	average (x)	2x	5x	10x	20x	30x	60x	90x	120x
NA12878_v6_R1	8.66 GB	142.69	98.1	97.8	97.4	96.4	95	87.9	74.9	57.5

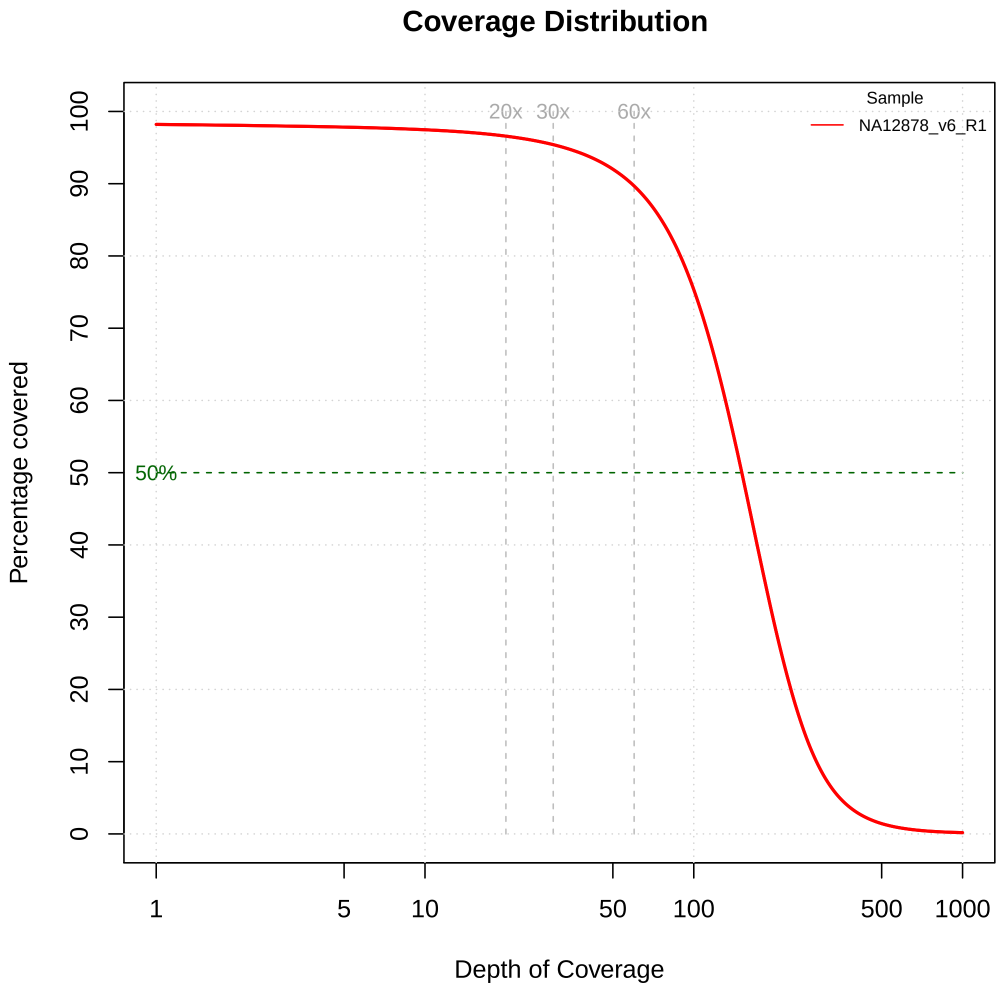

Figure 3: Coverage plot (including duplicated fragments).

Table 9: Depth of coverage summary (including duplicated fragments).

	target coverage		**% of target covered with at least**							
sample	total bases	average (x)	2x	5x	10x	20x	30x	60x	90x	120x
NA12878_v6_R1	10.24 GB	168.78	98.1	97.8	97.5	96.6	95.4	89.7	79.7	65.7

4.6 Library Report

Fragment insert size histogram of the paired-end library observed from all the samples analysed. The insert size is determined by mapping individual read pairs on the reference sequence. The distance between 5'prime ends of both sequenced reads in a pair that are mapped to the reference is the observed length of the sequenced fragment. By performing this operation for all mapped reads the distribution can be generated. X-axis shows the insert size in bp and Y-axis shows the number of fragments with the observed fragment insert sizes.

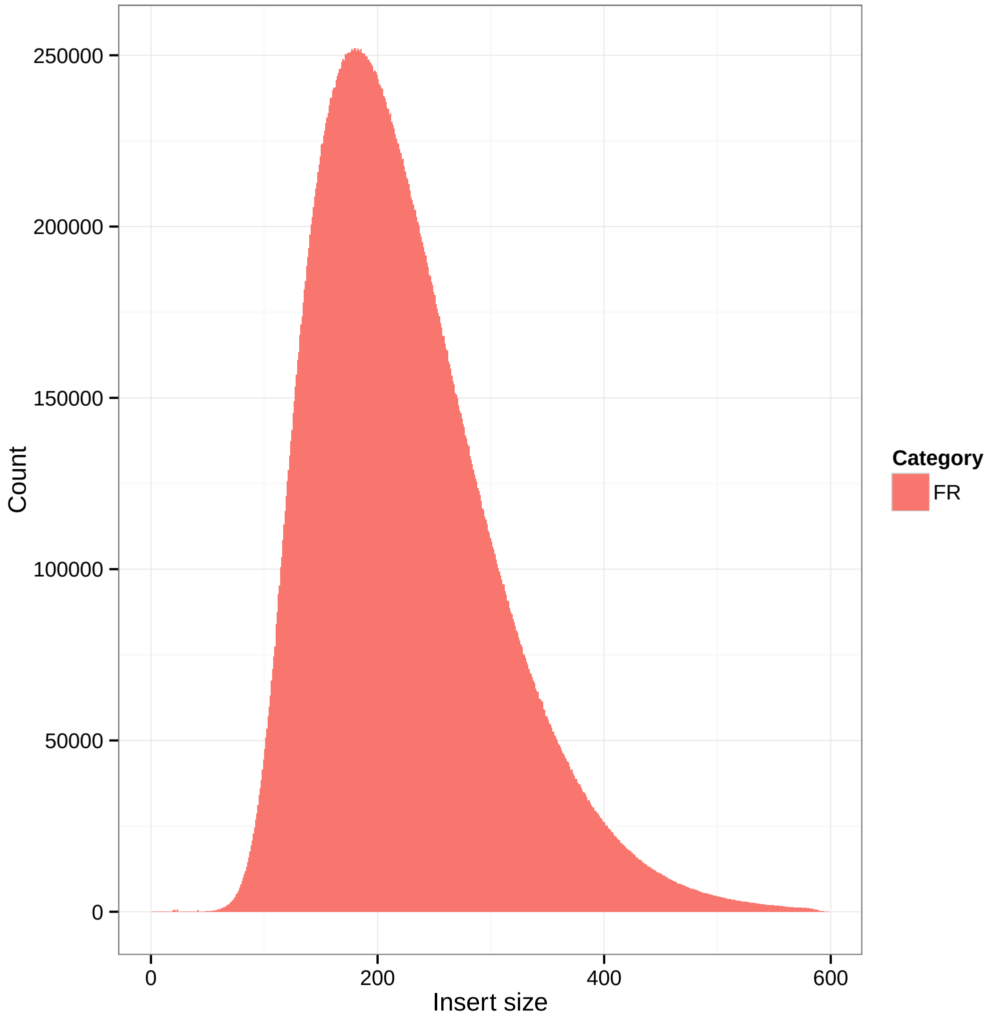

Figure 4: NA12878_v6_R1 .

Table 10: Sample wise insert size metrics for HQ aligned reads. The mean insert size (Mean) and its standard deviation (Stddev) is given in base pairs.

Sample	Pair orientation	Mean	Stddev	# Read pairs
NA12878_v6_R1	FR	224	79	45,091,258

4.7 Variant Analysis

The SNP and InDel calling is done using GATK's Haplotype Caller [3, 4].

Variants detected are annotated based on their gene context using snpEff [5]. The available annotations and their description is described in the tables 20 and 21. Several metrics, that are used to evalutate the quality of a variant, are annotated using GATK's VariantAnnotator module.

Customised filters are applied to the variants to filter false positive variants using GATK's VariantFiltration module. Filters used are described in tables 23 and 24.

Please note the variants reported are NOT VALIDATED and provided as it is reported from the programs mentioned above. Therefore it is highly recommended to inspect the variants thoroughly and validate using alternative methods.

The complete list of variants, stratified in single or few nucleotide(s) variants and insertion/deletion (InDel) variants is contained in the delivery package (see chapter 5) in the corresponding VCF and TSV files. The TSV format is described in tables 16 and 17 (fixed fields) and tables 18 and 19 (sample wise fields). The variants (SNV and InDels) detected are summarized in the following table(s).

Table 11: Variant metrics for NA12878_v6_R1

Variant Type	Feature	NA12878 _v6_R1
ALL [1]	TOTAL	156926
SNV	TOTAL	138409
	KNOWN	133394
	UNKNOWN	5015
	MISSENSE	13203
	NONSENSE	159
	SILENT	12589
	NONE	106258
	PASSED	95446
	FAILED	42963
	PASSED KNOWN	93227
	PASSED UNKNOWN	2219
	PASSED MISSENSE	12127
	PASSED NONSENSE	144
	PASSED SILENT	11953
	PASSED NONE	67647
INDEL	TOTAL	18517
	INS TOTAL	8459
	DEL TOTAL	10058
	KNOWN	15776
	UNKNOWN	2741
	INS MAX SIZE	180
	DEL MAX SIZE	167
	PASSED	17995
	FAILED	522
	PASSED KNOWN	15528
	PASSED UNKNOWN	2467

[1] Excluding complex sites (i.e. multiallelic calls).

5 Deliverables

Table 12: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
SAMPLE.snps.bed	BED	UCSC Genome Browser
SAMPLE.indels.bed	BED	UCSC Genome Browser
SAMPLE.snps.tsv	TSV	Spreadsheet Editor
SAMPLE.indels.tsv	TSV	Spreadsheet Editor
SAMPLE.snps.vcf	VCF	Text Editor
SAMPLE.indels.vcf	VCF	Text Editor
SAMPLE.alignment.bam	BAM	IGV, Tablet
SAMPLE.alignment.bam.bai	BAI	None
SAMPLE.unmapped.fastq	FASTQ	Text Editor

6 Formats

Table 13: References and descriptions of file format.

Format	Description
TSV	Tab separated table style text file. This can be imported into spreadsheet processing software like MS OFFICE Excel.
FASTQ[6]	Text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
BAM[7]	Compressed binary version of the Sequence Alignment/Mapping (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
VCF[8]	Variant Call Format (VCF) is a format to describe and report the variants.
BED	Browser Extensible Data (BED) is a text file compatible with genome browsers.

7 FAQ

Q: How can I open a TSV file in Excel?
A: Start Excel and click File -> Open and select the TSV file you want to open. Next an assistant dialog should show up. Make sure that you select tab as separator. Set the format of all rows without numbers to text. The TSV files use the dot as decimal separator and comma as thousands separator. Make sure that you set both correctly.

Bibliography

[1] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[2] Picard. http://picard.sourceforge.net.

[3] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[4] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43:491–498, 2011.

[5] Pablo Cingolani. "snpEff: Variant effect prediction". http://snpeff.sourceforge.net, 2012.

[6] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[7] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[8] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[9] Mary Kate Wing. "bamUtil is a repository that contains several programs that perform operations on SAM/BAM files.". http://genome.sph.umich.edu/wiki/BamUtil, 2015.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[11] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[12] Marc Lohse, Anthony M. Bolger, Axel Nagel, Alisdair R. Fernie, John E. Lunn, Mark Stitt, and Björn Usadel. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, July 2012.

[13] Derek Barnett, Erik Garrison, Aaron Quinlan, Michael Strömberg, and Gabor Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):btr174–1692, April 2011.

[14] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[15] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, February 2015.

A Sequence Data Used

Table 14: Analysed samples (SE = single end, PE = paired end).

Sample	Read Type	File Name
NA12878_v6_R1	PE	GATC-Demo_NA12878_v6_R1_lib00000_1.fastq GATC-Demo_NA12878_v6_R1_lib00000_2.fastq

B Relevant Programs

Table 15: Name, version and description of relevant programs.

Program	Version	Description
BWA[1]	0.7.15	BWA is a software package for mapping low-divergent sequences against a large reference genome
BamUtil[9]	1.0.10	BamUtil is a repository that contains several programs that perform operations on SAM/BAM files
GATK[3, 4]	3.7	GATK is a java-based command-line toolkit that process SAM / BAM / VCF files.
Picard[2]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[10]	2.15.3	R is a programming language and environment for statistical computing.
SAMTools[11]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
Trimmomatic[12]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.
bamtools[13]	2.3.0	BamTools provides a small, but powerful suite of command-line utility programs for manipulating and querying BAM files for data.
bedtools[14]	2.26.0	Bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-usedgenomic file formats such as BAM, BED, GFF/GTF, VCF
sambamba[15]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
snpEff[5]	4.3	snpEff is a variant annotation and effect prediction tool.

C Tables

Table 16: Examples of fixed fields of the tab delimited variant report table.

CHROMOSOME	POSITION	DBSNP ID	REFERENCE BASE	OBSERVED BASE	QUALITY SCORE	FILTER	COVERAGE
chr3	46399798	rs1799865	T	C	9551.17	PASS	254
chr3	164777677	rs9290259	T	G	9919.08	PASS	306
chr11	55873024	rs2449148	A	G	9104.32	PASS	240
chr12	55945119	rs7313899	A	G	9616.99	PASS	281
chr12	10570965	rs2682495	C	G	9476.45	PASS	278
chr17	66039350	rs4638	A	G	9077.84	PASS	253
chr19	53911973	rs10425136	A	G	9853.53	PASS	252
chr19	55378008	rs3745902	C	T	9066.27	PASS	297

Table 17: Defintion of fixed fields of the tab delimited variant report table

Name	Meaning
CHROMOSOME	Name of reference contig or chromosome where the variant occurs
POSITION	Position of reference contig or chromosome where the variant occurs
DBSNP ID	The dbSNP rs identifier of the SNP based on the contig or chromosome position of the call. If there is an entry in the dbSNP then the respective rs id will be displayed. Dot ('.') indicates no entry in the dbSNP.
REFERENCE BASE	The reference base at the variant site
OBSERVED BASE	Alternative (observed) base in the samples in general [VARIANT]
QUALITY SCORE	The Phred scaled probability of OBSERVED BASE is correct at this site given sequencing data. The value is computed based on error models designed by Broad Institute. Since 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. The higher the value the more accurate is the variant call.
FILTER	In addition to quality score, several filters can be defined to filter the SNPs by considering factors other than quality score alone. For e.g., SNP with low quality score threshold of < 30 could be tagged as LowQual SNPs and the ones which pass this filter will be tagged as PASS. More than one filter can be defined and applied to the variant calls. Default filters are SnpCluster (more than 2 SNPs found in cluster of size=10), LowQual (SNP with quality score < 30), LowCov (SNP with coverage < 20), Mask (SNP is at least 10 base near to indel location) and HardToValidate (Not enough evidence to validate). Variant passing the default filters will be tagged "PASS"
COVERAGE	Sequencing depth or coverage at the variant position. More accurate is to see the SAMPLE:COVERAGE

Table 18: Example of sample wise fields.

SAMPLE: GENOTYPE	SAMPLE:GQ	SAMPLE: ALLELE DEPTH	SAMPLE: ALELLE BALANCE	SAMPLE: COVERAGE
1/1	96.11	0.68		65
0/1	99	86.26	0.77	110
0/0	48.35	72.50	.	77

Table 19: Definition of sample wise fields.

Name	Meaning
SAMPLE:GENOTYPE	The genotype of the sample. For a diploid genome, the GENOTYPE indicates the two alleles carried by the sample, encoded by a 0 for the REFERENCE allele, 1 for the first ALTERNATIVE (OBSERVED) allele. Possible GENOTYPEs are 0/0 (the sample is homozygous to reference), 0/1 (the sample is heterozygous, carrying 1 copy of each of the REFERENCE and ALTERNATIVE alleles) and 1/1 (the sample is homozygous alternate i.e., completely opposite to the REFERENCE)
SAMPLE:GQ	The phred scaled genotype quality.
SAMPLE:ALLELE DEPTH	The allele depth, one for each REFERENCE and ALTERNATIVE (OBSERVED), is the count of all reads that carried with them the respective alleles. The read counts also include the poor mapping quality reads, unlike the COVERAGE counts.
SAMPLE:ALELLE BALANCE	Allele balance is a ratio of the REFERENCE bases to the total bases observed in the give position. This applies for only heterozygous calls and value ranges from > 0.0 to < 1.0
SAMPLE:COVERAGE	The total depth of the reads that passed the internal quality control metrics (for eg., mapping quality >17) from all reads present at this site.

Table 20: Examples of genomic annotations as produced by snpEff.

AMINO ACID CHANGE	CODON CHANGE	EFFECT	EXON ID	FUNCTIONAL CLASS	GENE NAME	IMPACT	TRANSCRIPT
R44S	agG/agT	NON SYNONYMOUS CODING	exon_1_935072_935552	MISSENSE	HES4	MODERATE	ENST00000428771
L615	Ctg/Ttg	SYNONYMOUS CODING	exon_1_881553_881666	SILENT	NOC2L	LOW	ENST00000327044
		FRAME_SHIFT	exon_1_877939_878438	NONE	SAMD11	HIGH	ENST00000342066
P605PG	cca/ccCGGa	CODON CHANGE PLUS CODON INSERTION	exon_1_35653574_35653691	NONE	SFPQ	MODERATE	ENST00000357214
-409G	-/GGG	CODON INSERTION	exon_1_1683910_1684499	NONE	NADK	MODERATE	ENST00000342348
Y205*	taT/taG	STOP GAINED	exon_1_25167264_25170815	NONSENSE	CLIC4	HIGH	ENST00000374379
154	tAa/tGa	SYNONYMOUS STOP	exon_4_41621205_41621953	SILENT	LIMCH1	LOW	ENST00000509638
		INTERGENIC	NONE			MODIFIER	
		UPSTREAM	NONE		AL669831.1	MODIFIER	ENST00000358533
		UTR_5_PRIME	exon_1_948803_948956	NONE	ISG15	MODIFIER	ENST00000379389
		SPLICE SITE ACCEPTOR		NONE	RP11-34P13.2	HIGH	ENST00000538476
		SPLICE SITE DONOR		NONE	SAMD11	HIGH	ENST00000342066

Table 21: Definition of genomic annotations as produced by snpEff.

Name	Meaning
AMINO ACID CHANGE	The exact position and the change of the amino acid.
CODON CHANGE	The change of the nucleotide within the context of the codon.
EFFECT	The predicted effect the change implies.
EXON ID	The Exon Id the variant belongs to.
FUNCTIONAL CLASS	Functional class of the SNP: silent (synonymous), missense (non-synonymous), nonsense (stop-gaining), readthrough (stop-loss), NA (unclassified), none.
GENE NAME	The gene entry associated with the location of the variant call. If present, gene name will be displayed. If not, ”NA” will be displayed.
IMPACT	Effect impact. Can be one of High, Moderate, Low and Modifier.
TRANSCRIPT ID	The transcript Id.

Table 22: Impact, Description and Examples of Effects as reported by snpEff.

Impact	Effects	Description	Examples
High	SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).	
	SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).	
	START_LOST	Variant causes start codon to be mutated into a non-start codon.	aTg/aGg, M/R
	EXON_DELETED	A deletion removes the whole exon.	
	FRAME_SHIFT	Insertion or deletion causes a frame shift	An indel size is not multple of 3
	STOP_GAINED	Variant causes a STOP codon	Cag/Tag, Q/*
	STOP_LOST	Variant causes stop codon to be mutated into a non-stop codon	Tga/Cga, */R
Moderate	NON_SYNONYMOUS _CODING	Variant causes a codon that produces a different amino acid	Tgg/Cgg, W/R
	CODON_CHANGE	One or many codons are changed	An MNP of size multiple of 3
	CODON_INSERTION	One or many codons are inserted	An insert multiple of three in a codon boundary
	CODON_CHANGE_PLUS _CODON_INSERTION	One codon is changed and one or many codons are inserted	An insert of size multiple of three, not at codon boundary
	CODON_DELETION	One or many codons are deleted	A deletion multiple of three at codon boundary
	CODON_CHANGE_PLUS _CODON_DELETION	One codon is changed and one or more codons are deleted	A deletion of size multiple of three, not at codon boundary
	UTR_5_DELETED	The variant deletes and exon which is in the 5'UTR of the transcript	
	UTR_3_DELETED	The variant deletes and exon which is in the 3'UTR of the transcript	
Low	SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon.	Ttg/Ctg, L/L (TTG and CTG can be START codons)
	NON_SYNONYMOUS_START		
	START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon.	
	SYNONYMOUS_CODING	Variant causes a codon that produces the same amino acid	Ttg/Ctg, L/L
	SYNONYMOUS_STOP	Variant causes stop codon to be mutated into another stop codon.	taA/taG, */*
	NON_SYNONYMOUS_STOP		
Modifier	UTR_5_PRIME	Variant hits 5'UTR region	
	UTR_3_PRIME	Variant hits 3'UTR region	
	REGULATION		
	UPSTREAM	Upstream of a gene (default length: 5K bases)	
	DOWNSTREAM	Downstream of a gene (default length: 5K bases)	
	GENE	The variant hits a gene.	
	TRANSCRIPT	The variant hits a transcript.	
	EXON	The vairant hits an exon.	
	INTRON_CONSERVED	The variant is in a highly conserved intronic region	
	INTRON	Variant hist and intron. Technically, hits no exon in the transcript.	
	INTRAGENIC	The variant hits a gene, but no transcripts within the gene	
	INTERGENIC	The variant is in an intergenic region	
	INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region	
	NONE		
	CHROMOSOME		
	CUSTOM		
	CDS	The variant hits a CDS.	

Table 23: Filters applied for single nucleotide variant sites.

Name	Expression	Description
LowCovFilter	$\leq$ 20	Depth of Coverage.
QDFilter	<2.0	Quality by read depth.
MQFilter	<-12.5	Root Mean Square of the Mapping quality of the reads across all samples.
FSFilter	>60.0	Phred-scaled p-value using Fisher's Exact Test to detect strand bias.
HaplotypeFilter	>13.0	Consistency of the site with two (and only two) segregating haplotypes.
MQFilter	<-12.5	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for mapping qualities.
ReadPosFilter	<-8.0	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele.

Table 24: Filter applied for small Insertion / Deletion variant sites.

Name	Expression	Description
QDFilter	<2.0	Quality by read depth.
ReadPosFilter	<-20.0	The phred-scaled p-value (u-based z-approximation) from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele.
FSFilter	>200.0	Phred-scaled p-value using Fisher's Exact Test to detect strand bias.

Eurofins Genomics' products, services and applications reach the best quality and safety levels. They are carried out under strict QM and QA systems and comply with the following standards:

ISO 9001	Globally recognised as the standard quality management certification	GLP	The gold standard to conduct non-clinical safety studies
ISO 17025	Accredited analytical excellence	GCP	Pharmacogenomic services for clinical studies
ISO 13485	Oligonucleotides according to medical devices standard	cGMP	Products and testing according to pharma and biotech requirements

Eurofins Genomics • Anzinger Str. 7a • 85560 Ebersberg • Germany