eurofins | Genomics

Data Analysis Report: Variant Analysis v2.4

Project / Study: EF-DEMO

Date: February 19, 2019

Table of Contents

1 Analysis workflow

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

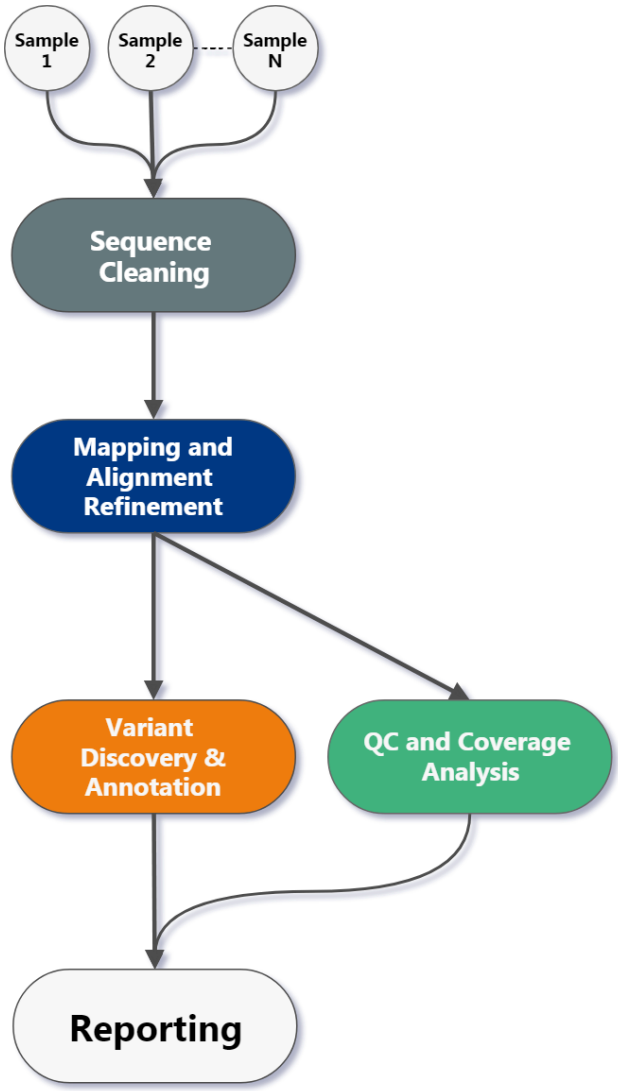

Figure 1: Variant Analysis v2.4 Workflow

2 Reference Database

Table 1: Information about the reference database.

Tag	Description
Name	Escherichia coli
Version	ecok12.30
Source	Ensembl
Size (bp)	4.641 MB
Sequences	1

Table 2: Information about additional reference data used.

Type	Version	Source
Annotation	30	Ensembl

3 Results

In total, 1 sample has been analysed. Please see table 11 in the appendix for details.

3.1 Sequence Cleaning

To improve subsequent analyses, a sequence cleaning was performed. First, sequencing adapter sequence that may be contained in reads due to read-through of short fragments is removed. Then, 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, clipped reads were discarded if they were shorter than 36 bp. Only high quality mate pairs (i. e. both forward and reverse read passed cleaning) were used for the next analysis step.

Detailed cleaning metrics for each sample can be found in file `*.cleaning_metrics.tsv` (see Deliverables, chapter 4)

Table 3: Sequence cleaning metrics.

No.	Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
1	DH1	9,261,414	550,402 (5.9%)	524,882 (5.7%)	8,186,130 (88.4%)

Total Reads: Total number of sequence reads analysed for each sample.
LQ Reads: Number (percentage) of discarded low quality reads.
Single Reads: Number (percentage) of high quality reads that lost their (low quality) mate during cleaning.
HQ Reads: Number (percentage) of high quality reads used for further analysis (always mate pairs).

3.2 Mapping and Alignment Processing

Mapping to the reference sequence / database is done using BWA[1] with default parameters. Please note that the mapping efficiency depends on the accuracy of the reference and the quality of sequence reads. Reads are then classified according to the following 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 with unmapped mates.
- Cross-Contig: Mapped reads with mates mapped to a different contig / chromosome.

For targeted sequencing (e. g. exome sequencing, amplicon panels), the targeted regions are subregions of the reference sequence. For whole genome sequencing, the target region is the full reference sequence. Unmapped reads, non-unique reads, singletons, cross-contig reads, and off-target reads are discarded. Only uniquely mapped reads are processed further.

Remaining reads are deduplicated using sambamba[2] in order to remove the artificial coverage caused by the PCR amplification step during the library preparation and / or sequencing. If a read maps to the same genomic location and has the same orientation as another already mapped read, the reads are considered as duplicates. For paired-end data, all mates of compared pairs have to fulfill the criteria in order to be designated as PCR duplicates. One copy of the duplicated reads is kept for futher analyses, the others are discarded.

As a next step, a base quality recalibration is performed to improve the base quality scores of reads. 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 a mismatch. This process is achieved by analysing the covariation among several different features of a base. The reported quality score, sequencing cycle, and sequencing context are considered for this step. Base quality recalibration is done using GATK[3, 4] modules.

Detailed alignment metrics for each sample can be found in file `*.alignment_metrics.tsv`. (see Deliverables, chapter 4).

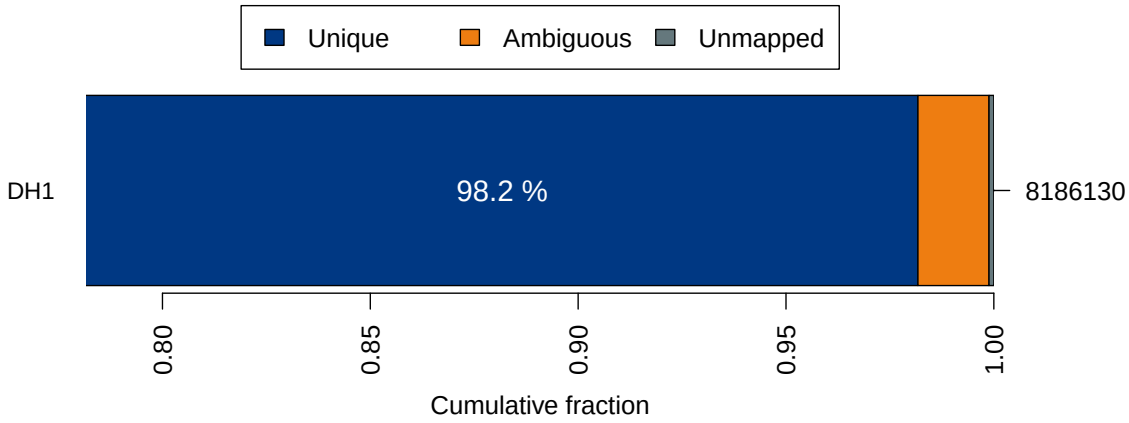

Figure 2: Summary of alignment results. For each sample, the fraction of uniquely mapped, non-uniquely mapped (ambiguous) and unmapped reads relative to the total number of reads per sample (right y-axis) is shown.

Table 4: Mapped read metrics observed per sample. Percentage of reads in category **Unique** is calculated based on the number of reads mapping to entire reference. Percentage of reads in category **Deduplicated** (reads without duplicates) is calculated based on the number of reads maped uniquely.

No.	Sample	Mapped HQ Reads	Unique	Deduplicated
1	DH1	8,176,655 (99.88%)	8,036,658 (98.29%)	7,978,599 (99.52%)

3.3 Coverage Report

The coverage plot shows the base coverage distribution of the aligned data. Depth of coverage is plotted on X-axis and the percentage of the respective reference covered is plotted on Y-axis. The shape of the curve indicates the uniformity of the reference coverage in the samples analysed.

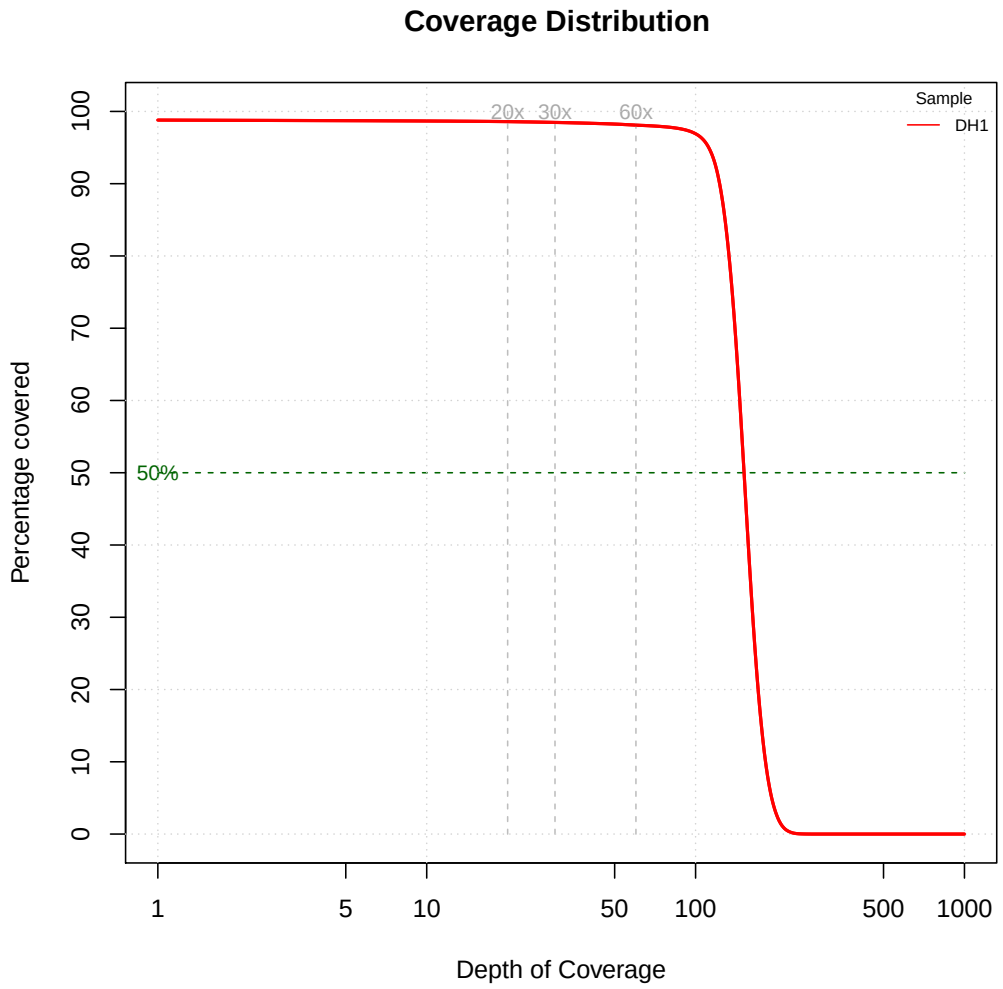

Figure 3: Coverage plot (excluding duplicated fragments).

Table 5: 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
DH1	694.66 MB	149.66	98.8	98.7	98.7	98.6	98.5	98.1	97.5	91.8

3.4 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 are detailed in tables 16 and 17. More information about annotations produced by snpEff can be found here. Several metrics that are used to evalutate the quality of a variant, are annotated using GATK's Variant Annotator module.

Customised filters are applied to the variants to filter false positive variants using GATK's Variant Filtration module. Filters used are described in tables 18 and 19. In this step, variants are classified as PASSED or FILTER_NAME (failed).

Please note that the variants are NOT VALIDATED and are provided as they are 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 is contained in the delivery package (see Deliverables, chapter 4) in the corresponding VCF and CSV files. The file formats are described in tables 14 and 15. The detected variants (SNP and InDels) are summarised in the following table(s).

Table 6: Variant metrics for all samples.

No.	Sample	Total	Passed	SNP	InDel	Known	Unknown
1	DH1	308	265	247	18	0	265

Table 7: Variant annotations for all samples.

No.	Sample	Passed	Missense	Nonsense	Silent
1	DH1	265	139	10	65

4 Deliverables

Table 8: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
PROJECT.Variant_Analysis_Report.pdf	PDF	PDF reader
PROJECT.alignment_metrics.tsv	TSV	Spreadsheet Editor
PROJECT.cleaning_metrics.tsv	TSV	Spreadsheet Editor
SAMPLE.alignment.bam	BAM	IGV, Tablet
SAMPLE.alignment.bam.bai	BAI	None
SAMPLE.snpEff_genes.txt	TXT	Text Editor
SAMPLE.snpEff_summary.csv	CSV	Spreadsheet Editor
SAMPLE.snpEff_summary.html	HTML	Web Browser
SAMPLE.variants.csv	CSV	Spreadsheet Editor
SAMPLE.variants.vcf	VCF	Text Editor

Table 9: Short descriptions of file contents.

File	Description
PROJECT.Variant_Analysis_Report.pdf	This report.
PROJECT.alignment_metrics.tsv	This file contains various alignment metrics.
PROJECT.cleaning_metrics.tsv	This file contains various sequence cleaning metrics.
SAMPLE.alignment.bam	Contains quality cleaned, mapped, filtered, and deduplicated reads in BAM format. This file is used for variant calling.
SAMPLE.alignment.bam.bai	The index file associated with SAMPLE.alignment.bam.
SAMPLE.snpEff_genes.txt	A per-gene summary of variant types.
SAMPLE.snpEff_summary.csv	A comma separated representation of the SAMPLE.snpEff_summary.html file for further processing.
SAMPLE.snpEff_summary.html	This file contains overall statistics of the snpEff run. These include quality and coverage histograms, distributions of variants accross chromosomes, classifications of variants to various types, etc. You can find a detailed description here.
SAMPLE.variants.csv	This file contains all identified variants of an individual sample in CSV format.
SAMPLE.variants.vcf	This file contains all identified variants of an individual sample in VCF format.

5 Formats

Table 10: References and descriptions of file format.

Format	Description
BAM[6]	Compressed binary version of the Sequence Alignment / Mapping (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
CSV	Comma separated table style text file. It can be imported into spreadsheet editors like MS OFFICE Excel.
HTML	Standard markup language for creating web pages and web applications
TSV	Tab separated table style text file. This can be imported into spreadsheet processing software like MS OFFICE Excel.
TXT	Text file of arbitrary style. It can be opened by any text editor. We recommend to use Notepad++
VCF[7]	Variant Call Format (VCF) is a format to describe and report the variants.

6 FAQ

Q: How can I open a CSV, TSV, or VCF file in Excel?
A: You can open CSV, TSV, VCF, or any other text file using Excel. Please follow this procedure:

1. Start Excel
2. Click on the "File" menu button in the top left corner
3. Click on the "Open" menu button in the left menu pane
4. Click on the dropdown-menu in the bottom right corner of the small window that opens. Initially, it should show "All Excel files (*.xls; *.xlsx)".
5. Select the topmost entry "All files (*.*)"
6. Navigate to the directory with the text files. They should be visible now.
7. Open the files and click through the appearing "Text Import Wizard" dialog (Next, Next, Done).

Depending on the content of the text file you want to import, you might want to change some settings in the "Text Import Wizard" dialog. Most often, you want to change the decimal separator. The provided text files use the dot as decimal separator and comma as thousands separator. Make sure that you set both correctly. To do this, click on the "Advanced" button in pane 3 of the "Text Import Wizard" dialog. You can find additional information in this article at the Microsoft Office support site.

Q: How can I view alignments and variants?
A: A convenient tool to view alignments and variant data is the *Integrative Genomics Viewer* (IGV) for Unix, MS Windows, and MacOS X. It can be downloaded and installed locally, or can be run as web-application.

- Before loading alignments or variant data into IGV, the reference genome FASTA file has to be loaded via the *Genomes -> Load Genome from File* menu. Make sure that you load the same reference genome FASTA file that was used during mapping.
- To load alignments into IGV select the BAM files via the *File -> Load from File* menu. Please note that you need to zoom-in to about 30kb to see alignments. You can set this visibility range threshold and other displaying and filtering options via the *View -> Preferences -> Alignments* menu, or the right-click context menu.
- To load variant data into IGV select the VCF files via the *File -> Load from File* menu. IGV can color mismatch bases and InDel positions. Use the right-click context menu to configure this and other displaying and filtering options. Not all mismatch positions in alignments might have been considered significant by the variant analysis tool and therefore might not be contained in the variant tracks.
- Please visit the IGV online manual to get more information about loading genomes, viewing alignments, and viewing variants.

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

[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] 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.

[7] 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.

[8] 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.

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

[11] 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.

[12] 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.

[13] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120, August 2014.

A Sequence Data Used

Naming convention for FASTQ files:
<project-id>_<sample-id>_<lib-id>_<run-id>_<lane-no>_<read-no>.fastq.gz

<project-id> the unique identifier of this project.

<sample-id> the sample name as provided by the customer.

<lib-id> a unique identifier of the sequencing library created in the lab. Multiple sequencing libraries may have been created from the same sample material, depending e.g. on project setup.

<run-id> a unique identifier of the sequencing run that created this file.

<lane-no> a number specifying the lane of the sequencing device used for sequencing.

<read-no> either _1 or _2. For paired-end runs, these numbers identify the associated forward and reverse read files (mate pairs).

Table 11: Analysed samples.

No.	Sample	File Name
1	DH1	EF-DEMO_DH1_lib12345_1.fastq.gz EF-DEMO_DH1_lib12345_2.fastq.gz

B Relevant Programs

Table 12: Name, version and description of relevant programs.

Program	Version	Description
bamtools[8]	2.3.0	BamTools provides a small, but powerful suite of command-line utility programs for manipulating and querying BAM files for data.
BamUtil[9]	1.0.10	BamUtil is a repository that contains several programs that perform operations on SAM/BAM files
BWA[1]	0.7.15	BWA is a software package for mapping low-divergent sequences against a large reference genome
GATK[3, 4]	3.7	GATK is a java-based command-line toolkit that process SAM / BAM / VCF files.
Picard[10]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[11]	3.2.4	R is a programming language and environment for statistical computing.
sambamba[2]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
SAMTools[12]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
snpEff[5]	4.3	SnpEff is a genetic variant annotation and effect prediction toolbox.
Trimmomatic[13]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.

C Tables

Table 13: Defintion of the VCF INFO and FORMAT fields in **.variants.vcf* files.

Field	Description
AB	Allele balance for each het genotype
AC	Allele count in genotypes, for each ALT allele, in the same order as listed
AD	Allelic depths for the ref and alt alleles in the order listed
AF	Allele Frequency, for each ALT allele, in the same order as listed
AN	Total number of alleles in called genotypes
ANN	Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO'
BaseCounts	Counts of each base
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
DP	Approximate read depth (reads with MQ=255 or with bad mates are filtered)
DP	Approximate read depth; some reads may have been filtered
DS	Were any of the samples downsampled?
ExcessHet	Phred-scaled p-value for exact test of excess heterozygosity
FS	Phred-scaled p-value using Fisher's exact test to detect strand bias
GQ	Genotype Quality
GT	Genotype
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
LOF	Predicted loss of function effects for this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected'
LikelihoodRankSum	Z-score from Wilcoxon rank sum test of Alt Vs. Ref haplotype likelihoods
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
MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
NMD	Predicted nonsense mediated decay effects for this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected'
PL	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification
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

Table 14: Examples of fixed fields of the comma separated variant report table in **.variants.csv* files.

CHROMOSOME	POSITION	DBSNP ID	REFERENCE BASE	OBSERVED BASE	FILTER	MUTATION FREQUENCY	COVERAGE
chr1	8064578	rs72634228	A	G	LowCovFilter	0.50	6
chr1	165376227	rs111545739	C	CCG	PASS	1.00	2
chr1	165376231	rs764778331	CTG	C	PASS	1.00	2
chr2	141571329	rs3749010	T	C	LowCovFilter	1.00	5
chr3	69813146	.	GT	G	PASS	0.50	3
chr4	1739816	rs28557273	C	T	QDFilter	0.50	29

Table 15: Definition of fixed fields of the comma separated variant report table in **.variants.csv* files.

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].
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".
MUTATION FREQUENCY	The mutation frequency with which a particular variant occurs in a population.
COVERAGE	Sequencing depth or coverage at the variant position.

Table 16: Example of fixed fields of the comma separated variant report table in **.variants.csv* files.

EFFECT	IMPACT	CODON CHANGE	AMINO ACID CHANGE	GENE NAME	BIOTYPE	TRANSCRIPT ID
synonymous_variant	LOW	c.4731G>A	p.Ala1577Ala	MTOR	protein_coding	ENST00000361445.8_2
missense_variant	MODERATE	c.7078A>G	p.Asn2360Asp	SPEN	protein_coding	ENST00000375759.7_1
upstream_gene_variant	MODIFIER	n.-3584T>C	.	PTPRF	processed_transcript	ENST00000477970.1_1
intron_variant	MODIFIER	c.298+29T>C	.	NOTCH2NL	protein_coding	ENST00000362074.7_1

Table 17: Definition of fixed fields of the comma separated variant report table in **.variants.csv* files.

Name	Meaning
EFFECT	The predicted effect the change implies.
IMPACT	Effect impact. Can be one of High, Moderate, Low and Modifier.
CODON CHANGE	The exact position and the change of the nucleotide within the context of the codon.
AMINO ACID CHANGE	The exact position and the change of the amino acid.
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.
BIOTYPE	The bare minimum is at least a description on whether the transcript is Coding or Noncoding.
TRANSCRIPT ID	The transcript Id.

Table 18: Filters applied for single nucleotide variant sites.

Name	Expression	Description
LowCovFilter	$\leq$ 20	Depth of Coverage.
QDFilter	<2.0	Quality by read depth.
MQFilter	<40.0	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.
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.
MQRankSumLow	<-12.5	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities

Table 19: 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 17025	Accredited analytical excellence	GLP	The gold standard to conduct non-clinical safety studies
ISO 13485	Oligonucleotides according to medical devices standard	GCP	Pharmacogenomic services for clinical studies
cGMP	Products and testing according to pharma and biotech requirements		

Eurofins Genomics Europe Sequencing GmbH • Jakob-Stadler-Platz 7 • 78467 Constance • Germany