Source BioScience
Illumina Sequencing

Variant Calling Report

A:AK

Contents

Samples and references ***3***

Advanced Bioinformatics Workflow ***4***

1. Adapter and Quality Trimming *5*

2. Mapping, Deduplicating, Sorting and Indexing Trimmed Reads *6*

3. Variant Calling and Annotation *7*

Definitions ***9***

FASTQ format *9*

Quality scores *9*

Note *10*

SAM/BAM format *10*

VCF File Format *11*

Samples and references

The following samples were included in the analysis:

Sample name	Sequenced at
S1	Source BioScience
S2	Source BioScience
S3	Source BioScience
S4	Source BioScience
S5	Source BioScience
S6	Source BioScience
S7	Source BioScience
S8	Source BioScience
S9	Source BioScience
S10	Source BioScience
S11	Source BioScience
S12	Source BioScience
S13	Source BioScience
S14	Source BioScience

The following reference sequence and regions file were used for mapping:

Reference species	Reference source	Reference size (bp)	Regions
Homo Sapiens	UCSC (hg19)	3.23 Gb	Agilent SureSelect Exome V7

Advanced Bioinformatics Workflow

We have setup a personalised pipeline for your sequenced samples. The pipeline consisted of three parts:

1. Adapter and Quality Trimming
2. Mapping, Deduplicating, Sorting and Indexing Trimmed Reads
3. Variant Calling and Annotation

1. Adapter and Quality Trimming

Raw FASTQ files were adapter and quality (Q30) trimmed using bbduk (bbmap Version 38.51).

Output files:

Trimmed FASTQ files for each direction (R1 = forward, R2 = reverse) for each sample:

Sample_name/Sample_name_ R1_001_trimmed.fastq.gz
Sample_name/Sample_name_R2_001_trimmed.fastq.gz

For more information about the FASTQ format, please refer to the Definitions section below.

2. Mapping, Deduplicating, Sorting and Indexing Trimmed Reads

Data was aligned (or mapped) to the reference genome using BWA-mem (version 0.7.17-r1188). Duplicate reads were then marked and removed (deduplicated) to enable accurate quantification of variants. The resulting BAM files were sorted by coordinate and indexed for visualisation with common genome browser software.

Alignment statistics for all samples are summarised in table 1. For more detailed metrics please refer to the metrics files provided (see output files below).

Sample name	Total reads	HQ aligned reads	Mean read length
S1	27,430,726	25,663,301	125.0
S2	28,177,540	26,382,081	125.1
S3	42,406,782	39,793,450	125.4
S4	31,363,592	29,379,812	122.9
S5	78,985,474	73,902,285	122.9
S6	27,118,486	25,364,708	125.4
S7	63,137,382	59,060,120	123.4
S8	37,194,072	34,758,449	124.1
S9	70,835,960	66,211,670	123.5
S10	36,886,374	34,552,517	121.2
S11	31,635,626	29,609,519	122.9
S12	30,006,248	27,969,462	123.2
S13	34,631,042	32,381,539	123.0
S14	30,377,174	28,429,925	122.6
S15	43,008,728	41,666,267	93.1
S16	37,729,724	36,571,548	98.0

Table 1: Alignment statistics following alignment to the reference sequence. Total reads = number of reads in trimmed FASTQ files (R1 and R2); HQ aligned reads = number of aligned reads that map with a high quality (mapQ >= 20); Mean read length = mean length of reads that could be aligned.

Output files:

Final BAM files for each sample as well as the corresponding index files (BAI):

Sample_name/Sample_name_final.bam
Sample_name/Sample_name_final.bai

Metrics files:

Sample_name/Sample_name _metrics.txt
Sample_name/Sample_name_insert_metrics.txt
Sample_name/Sample_name_insert_size_histogram.pdf
Sample_name/Sample_name_alignment_metrics.txt

For more information about the BAM file format please refer to the Definitions section below.

3. Variant Calling and Annotation

Variant calling was performed using GATK (GATK v4.1.0) according to the GATK Best Practices recommendations. Firstly, base quality scores were recalibrated using databases of known polymorphic sites (dbSNP and 1000G) to improve the accuracy of base quality scores. Variant calling was then performed with *GATK4 HaplotypeCaller (GVCF mode)* for the genomic regions specified in the regions BED file. During this step a local reassembly and realignment of reads occurs in regions surrounding potential variant sites. The optional setting "--dont-use-soft-clipped-bases=true" was applied during variant calling to exclude any poor quality bases from the analysis, as well as the optional setting "bamout", to generate a BAM file of the realigned reads, for visualisation in genome browser software. Raw variants from all samples were then combined and joint genotyping performed. Next SNPs and INDELs were extracted and soft-filtered according to the settings shown in table 2. SNPs or INDELS that did not meet the filter criteria were retained in final VCF files but labelled with the respective filter flag. Finally, variants were annotated with matches by chromosome position to the publicly available databases UCSC, dbSNP, ClinVar and dbnsfp. VCF files were indexed to allow visualision in genome browser software and converted to a table for visualisation in spreadsheet software.

Filter name	Filter description	SNP setting	INDEL setting
QD	QualByDepth: variant confidence (QUAL) divided by unfiltered depth of non-hom-ref samples	< 2.0	< 2.0
MQ	RMSMappingQuality: root mean square mapping quality over all reads for that variant/site	< 40.0	N/A
FS	FisherStrand: Phred-scalded probability of strand bias for that variant/site	> 60.0	> 200.0
SOR	StrandOddsRatio: strand bias test that only assesses ratios of reads that cover both alleles (vs FS which penalizes for variants with only one allele)	> 4.0	> 10.0

Table 2: Variant calling filters: Filter thresholds used for soft-filtering SNPs and INDELs. Any variants that pass these criteria carry the FILTER tag "PASS"; any variants that failed these criteria carry the respective FILTER tag, e.g. "QD".

Variant calling statistics for all samples are summarised in table 3. For more detailed metrics please refer to the metrics files provided (see output files below).

Sample name	Total variants	Total SNPs	Total INDELs	Filtered SNPs	Filtered INDELS	Het/Hom ratio
S1	28,013	27,028	985	26,104	965	1.76
S2	27,061	26,113	948	25,250	933	1.68
S3	27,629	26,625	1,004	25,652	989	1.84
S4	27,509	26,522	987	25,604	976	1.80
S5	27,639	26,626	1,013	25,634	996	1.74
S6	27,702	26,691	1,011	25,835	995	1.71
S7	28,035	27,034	1,001	26,014	980	1.72
S8	27,373	26,410	963	25,584	950	1.73
S9	27,527	26,550	977	25,530	955	1.69
S10	27,205	26,258	947	25,356	931	1.72
S11	27,031	26,048	983	25,305	973	1.70
S12	26,990	26,029	961	25,244	946	1.78
S13	27,231	26,285	946	25,369	933	1.67
S14	27,356	26,427	929	25,605	916	1.78
S15	25,438	24,647	791	24,088	782	1.68
S16	25,037	24,302	735	23,780	728	1.63

Table 3: Variant calling statistics. Total variants = total number of unfiltered SNPs and INDELs, also shown separately. Filtered SNPs and INDELs = number of SNPs and INDELs remaining after filtering, using criteria shown in table 2. Ratio of heterozygous to homozygous-reference variants (SNPs & INDELs). Values around 1.6 are typically seen in European exome samples (Wang J et al, Bioinformatics, 2015, 31(3):318-323)

Output files:

Final VCF files for each sample as well as the corresponding index files (IDX):

Sample_name/Sample_name_ SNPs_indels.vcf

Sample_name/Sample_name_ SNPs_indels.vcf.idx

Annotated VCF files for each sample as well as the corresponding index files (IDX):

Sample_name/Sample_name_ SNPs_indels_annotated.vcf

Sample_name/Sample_name_ SNPs_indels_annotated.vcf.idx

Table of annotated variants:

Sample_name/Sample_name_ SNPs_indels_annotated.tsv

Metrics files:

Sample_name/Sample_name _variant_calling_summary_metrics.txt

Sample_name/Sample_name_variant_calling_detail_metrics.txt

For more information about the VCF file format please refer to the Definitions section below.

Definitions

FASTQ format

A sample entry is provided and explained below:

@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@

The first line is prefixed by the "@" symbol and contains the read name. These names are parsed until the first encountered whitespace. Due to this behavior, adding additional tags to the header line is not problematic for extant FASTQ parsers.

The second line contains the sequence bases.

The third line is prefixed by a + symbol and sometimes repeats the read name. The read name is omitted in the minimal FASTQ case.

The fourth line contains the base qualities where BQ + 33 = ASCII value shown in the base quality string

The header line is interpreted as follows:

@ <instrument-name>:<run ID>:<flowcell ID>:<lane-number>:<tile-number>:
<x-pos>: <y-pos><read number>:<is filtered>:<control number>:<barcode sequence>
<is filtered> is Y if the read is filtered, N otherwise.

<control number> is 0 when none of the control bits are on, otherwise it is an even number.

<barcode sequence> represents the USE_BASES masked barcode sequence, empty otherwise.

Quality scores

A quality score (or Q-score) expresses an error probability. In particular, it serves as a convenient and compact way to communicate very small error probabilities. Given an assertion, A, the probability that A is not true, $P(\neg A)$, is expressed by a quality score, $Q(A)$, according to the relationship:

$$Q(A) = -10[\log_{10} P(\neg A)],$$

where $P(\neg A)$ is the estimated probability of an assertion A being wrong.
The relationship between the quality score and error probability is demonstrated with the following table:

Quality score $\mathrm{Q(A)}$	Error probability $\mathrm{P(\neg A)}$
10	0.1
20	0.01
30	0.001
40	0.0001

Note

- *Clusters passing filter are flagged as PF, if no more than one base call in the first 25 cycles has a chastity of < 0.6, which is defined as the ratio of the intensity of the most intense signal for a cluster divided by the sum of the most intense plus the second most intense signal.*

SAM/BAM format

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with `@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information. BAM is a compressed SAM file in BGZF format. BAM files can be indexed (BAI) for fast random access. This is required e.g. for loading it into a genome browser where BAI files have to be located in the same directory as corresponding BAM files.
Every row of the output of a SAM/BAM file contains the following information:

Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Where flag can be one of the following:

Char	Flag	Description
P	0x0001	the read is paired in sequencing
P	0x0002	the read is mapped in a proper pair
U	0x0004	the query sequence itself is unmapped
U	0x0008	the mate is unmapped
R	0x0010	strand of the query (1 for reverse)
R	0x0020	strand of the mate
1	0x0040	the read is the first read in a pair
2	0x0080	the read is the second read in a pair
S	0x0100	the alignment is not primary
F	0x0200	QC failure

VCF File Format

Header:

File meta-information is included after the ## string, often as key=value pairs.

- INFO fields are described as follows:
  ##INFO=<ID=ID,Number=number,Type=type,Description=”description”>
- FILTERs that have been applied to the data are described as follows:
  ##FILTER=<ID=ID,Description=”description”>
- Likewise, Genotype fields specified in the FORMAT field are as follows:
  ##FORMAT=<ID=ID,Number=number,Type=type,Description=”description”>

Header line:

The header line names the 10 columns. These columns are as follows:

1. #CHROM
2. POS
3. ID
4. lREF
5. ALT
6. QUAL
7. FILTER
8. INFO
9. FORMAT
10. Sample ID

Data lines:

There are 8 fixed fields per record. All data lines are tab-delimited. In all cases, missing values are specified with a dot (”.”). Fixed fields are:

CHROM Chromosome: an identifier from the reference genome.

POS Position: The reference position, with the 1st base having position 1. Positions are sorted numerically, in increasing order, within each reference sequence CHROM.

ID Semi-colon separated list of unique identifiers where available. If this is a dbSNP variant it is encouraged to use the rs number(s).

REF Reference base(s): Each base must be one of A,C,G,T,N. Bases should be in uppercase. Multiple bases are permitted. The value in the POS field refers to the position of the first base in the String. For InDels,

the reference String must include the base before the event (which must be reflected in the POS field).

ALT Comma separated list of alternate non-reference alleles called on at least one of the samples. Options are base Strings made up of the bases A,C,G,T,N, or an angle-bracketed ID String (”<ID>”).

QUAL Phred-scaled quality score for the assertion made in ALT. i.e. give -10log_10 prob(call in ALT is wrong). If ALT is ”.” (no variant) then this is -10log_10 p(variant), and if ALT is not ”.” this is -10log_10 p(no variant). High QUAL scores indicate high confidence calls. Although traditionally people use integer phred scores, this field is permitted to be a floating point to enable higher resolution for low confidence calls if desired.

FILTER Filter: 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 list of codes for filters that fail. e.g. “QD” indicates that at this site the QualByDepth is below the filter parameter used (refer to the filter table above for filter parameters used here).

INFO Additional information: (Alphanumeric String) INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data].

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.
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**.**
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.
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.

FORMAT Data types and order. This is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the format.

GT genotype, encoded as alleles values separated by ”/” or “|”. The allele values are 0 for the reference allele (what is in the reference sequence), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on. For diploid calls examples could be 0/1 or 1|0 etc. For haploid calls, e.g. on Y, male X, mitochondrion, only one allele value should be given. All samples must have GT call information; if a call cannot be made for a sample at a given locus, ”.” must be

specified for each missing allele in the GT field (for example ./. for a diploid). The meanings of the separators are:

/ : genotype unphased

| : genotype phased

AD Allelic depths for the ref and alt alleles in the order listed.

DP Approximate read depth (reads with MQ=255 or with bad mates are filtered).

GQ Genotype Quality, encoded as a Phred quality -10log_10p(genotype call is wrong).

PL Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification.

If any of the fields is missing, it is replaced with the missing value (“.”)