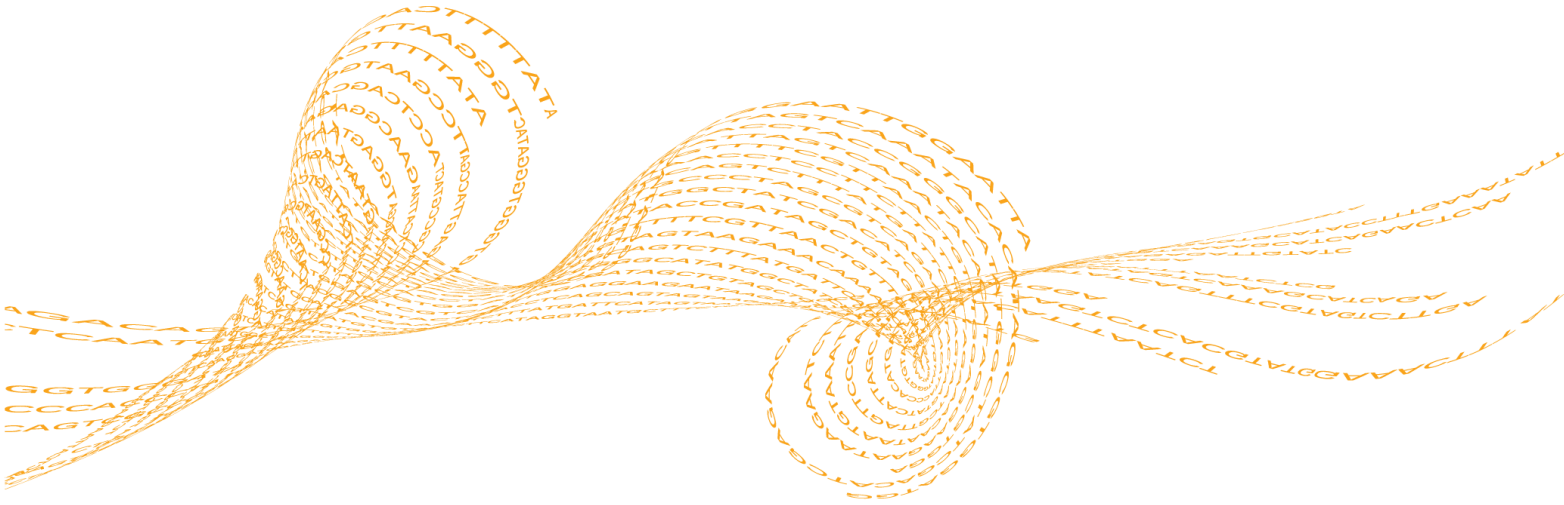
BWA Whole Genome Sequencing v1.0

BaseSpace App Guide

For Research Use Only. Not for use in diagnostic procedures.

Introduction	3
Workflow Diagram	5
Set Analysis Parameters	6
Analysis Methods	7
Analysis Output	10
Revision History	19
Technical Assistance	



ILLUMINA PROPRIETARY
15050952 v01
January 2016

illumina®

Introduction

After BaseSpace® generates the FASTQ files containing the base calls and quality scores of the run, you can use the Burrows-Wheeler Aligner (BWA) Whole Genome Sequencing v1.0 App to analyze the sequencing data. The app analyzes the data in two parts; first, it aligns to the reference genome, and then it assembles and performs variant calls.

Compatible Libraries

See the BaseSpace support page for a list of library types that are compatible with the BWA Whole Genome Sequencing v1.0 App.

Workflow Requirements

- This app does not support mate-pair or other non-forward and -reverse styles of paired-end sequencing.
- This app does not support annotation of non-human genomes.
- A minimum read length of 21 bp and a maximum read length of 150 bp.
- Minimum recommended data set size is enough data to yield 10 × coverage after alignment of the genome being sequenced. See Table 1 for details.
- Maximum data set size is fewer than 200 gigabases, which equates to the following:
  - Approximately 1 billion reads assuming 2 × 100.
  - Approximately 665 million reads assuming 2 × 150.
- Completedjobinfo.xml may not print all statistics.
- Sample name length has a maximum of 32 characters.
- GQX can be entered as any value, although the maximum recommended value is 99.
- GATK indel realignment on chrM (at very high coverage) displays this warning:
  - `Reads will be written to bam out of order`
  - One of the Picard tools, ValidateSamFile, displays this error:
  - `MAPQ should be 0 for unmapped read` and `Mate unmapped flag does not match read unmapped flag of mate`.
- Isaac App is recommended if PhiX is being sequenced at a coverage > 5,000,000 ×.

For recommended minimum number of reads for 10 × coverage for different species, see Table 1. The number of reads listed yields 10 × coverage with an additional 5% to account for unaligned reads.

Table 1 Recommended Minimums for 10 × Coverage

	Genome Size	Data Size	Reads for 2 × 100 (million)	Reads for 2 × 150 (million)
Arabidopsis thaliana	63.4 Mb	666 Mb	3.33	2.22
Bos taurus	2.65 Gb	28 Gb	140.00	93.33
Escherichia coli K-12 DH10B	4.7 Mb	50 Mb	0.25	0.17
Escherichia coli K-12 MG1655	4.6 Mb	49 Mb	0.25	0.16
Drosophila melanogaster	139.5 Mb	1.5 Gb	7.33	4.88

	Genome Size	Data Size	Reads for 2 × 100 (million)	Reads for 2 × 150 (million)
Human	3.3 Gb	35 Gb	175.00	116.67
Mus musculus	2.6 Gb	28 Gb	140.00	93.33
PhiX (Illumina)	5386 b	57 Kb	282.77	188.51
Rattus norvegicus	2.9 Gb	31 Gb	155.00	103.33
Rhodobacter sphaeroides 2.4.1	4.6 Mb	49 Mb	0.25	0.16
Saccharomyces cerevisiae	12.2 Mb	129 Mb	0.65	0.43
Staphylococcus aureus NCTC 8325	12.8 Mb	135 Mb	0.68	0.45

Versions

These components are used in the BWA WGS App.

Software	Version
BWA	0.6.1-r104-tpx-isis
GATK	1.6-22-g3ec78bd
CNV Variant Caller (CNVseg)	2.2.4
Structural Variant Caller (Grouper)	1.4.2
SAMtools	0.1.18
Tabix	0.2.5 (r1005)

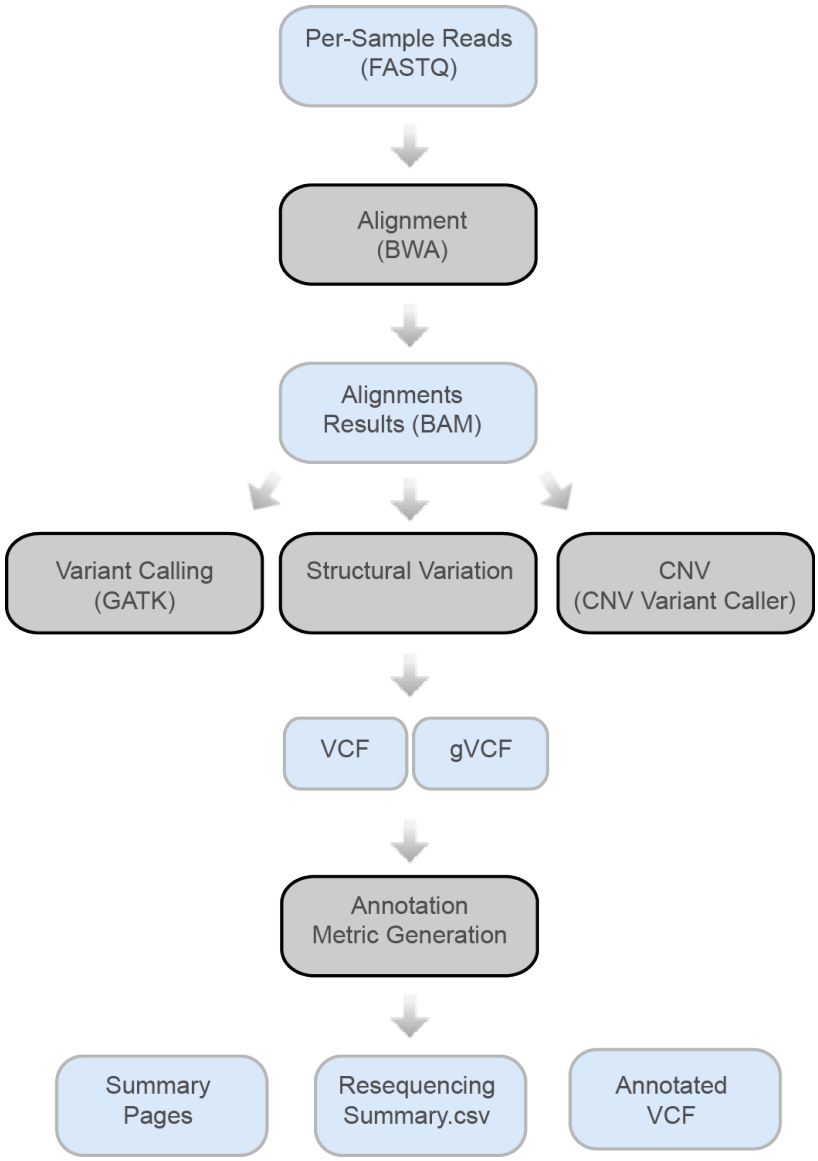
Reference Genomes

These genomes are available for alignment:

- Human, UCSC hg19
  The human reference genome is PAR-Masked, which means that the Y chromosome sequence has the Pseudo Autosomal Regions (PAR) masked (set to N) to avoid mismapping of reads in the duplicate regions of sex chromosomes.
- Arabidopsis thaliana (NCBI build9.1)
- Bos taurus (Ensembl UMD3.1)
- Escherichia coli K-12 DH10B (NCBI 2008-03-17)
- Escherichia coli K-12 MG1655 (NCBI 2001-10-15)
- Drosophila melanogaster (UCSC dm3)
- Mus musculus (UCSC MM9)
- Phi X (Illumina)
- Rattus norvegicus (UCSC RN5)
- Rhodobacter sphaeroides 2.4.1 (NCBI 2005-10-07)
- Saccharomyces cerevisiae (UCSC sacCer2)
- Staphylococcus aureus NCTC 8325 (NCBI 2006-02-13)

Workflow Diagram

Figure 1 BWA WGS App Workflow



Set Analysis Parameters

1. Navigate to BaseSpace, and then click the **Apps** tab.
2. Click **BWA Whole Genome Sequencing v1.0.**
3. From the drop-down list, select **version 1.0.0**, and then click **Launch** to open the app.
4. In the **Analysis Name** field, enter the analysis name.
   By default, the analysis name includes the app name, followed by the date and time that the analysis session starts.
5. From the **Save Results To** field, select the project that stores the app results.
6. From the **Sample(s)** field, browse to the sample you want to analyze, and select the checkbox. You can select multiple samples.
7. From the **Reference Genome** field, select the reference genome you want to align.
8. From the **Enable SV/CNV calling** field, select the checkbox to perform Isaac Variant Caller calls structural variation (SV) and copy number variation (CNV) calling.
   This option applies to paired-end data.
   For more information, see *Large Indel and Structural Variant Calls* on page 7 and *CNV Variant Caller* on page 8).
9. From the **Annotation** field, select a preferred gene and transcript annotation reference database.
10. Optional, click the **Advance** drop-down list for additional parameter fields.
    a. From the **Min GQX for Variants** field, enter the quality GQX for variants. GQX is the minimum of the GQ (genotype quality) and QUAL (low quality filter), which makes it a conservative filter. The default value is 30. The maximum recommended is 99.
    b. From the **Max Strand Bias for Variants** field, enter the maximum allowed strand bias for variant calling. This option filters for reads when the differences in allele frequencies for forward- and reverse-strand reads are too high. The default value is 10. Variants with a high strand bias are flagged as filtered. Strand bias is higher when the evidence for a variant is confined to reads from just one strand.
    c. From the **FlagPCRDuplicates** field, select the checkbox to have PCR duplicates flagged in the BAM files and not be used for variant calling. PCR duplicates are defined as paired-end reads generated from two clusters that have the exact same alignment positions for each read. Optical duplicates are already filtered out during RTA processing.
11. Click **Continue**.
    The BWA Whole Genome Sequencing v1.0 App begins analysis of the sample. When analysis is complete, the status of the app session is updated automatically and an email is sent to notify you.

Analysis Methods

The BWA Whole Genome Sequencing v1.0 App uses these methods to analyze the sequencing data.

BWA

The BWA Whole Genome Sequencing workflow uses the Burrows-Wheeler Aligner (BWA, which adjusts parameters based on read lengths and error rates, and then estimates the insert size distribution.

For more information, see github.com/lh3/bwa.

After BWA alignment, GATK performs variant calling.

GATK

The Genome Analysis Toolkit (GATK) is the standard variant caller after BWA alignment.

Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) first calls raw variants for each sample read. Then GATK analyzes the variants against known variants, and applies a calibration procedure to compute a false discovery rate for each variant. Variants are flagged as homozygous (1/1) or heterozygous (0/1) in the VCF file sample column.

The GATK best practices were guidelines for the app; they are described here: www.broadinstitute.org/gatk/guide/topic?name=best-practices.

Fo

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5): 491-8.

Large Indel and Structural Variant Calls

The large indel and structural variant caller uses the series of modules described here, and then generates output files in VCF 4.1 format.

Before ReadBroker

- **StatsGenerator**—Computes summary statistics on insert sizes, read orientation, and alignment scores for each input BAM file.
- **AnomalousReadFinder**—Grouper processes chromosomes in chunks. This method enables parallel execution and, therefore, faster performance. AnomalousReadFinder examines all alignments in a block and classifies reads and read pairs as follows:
  - Classifies reads as either shadow (unaligned) or semialigned partial or clipped alignment).
  - Classifies read pairs as either InsertionPair, DeletionPair, InversionPair, TandemDuplicationPair, or ChimericPair, according to which type of structural variant an anomalously mapped read pair is associated.
- **ClusterFinder**—Clusters reads based on their type and the position of their alignment. Only reads of the same type are clustered together at this stage, except shadow and semialigned reads, which can be clustered together.

- **ClusterMerger**—Associates clusters of various anomalous read types with shadow/semi-aligned read clusters, which breakpoints can cause. A breakpoint is a pair of bases that are adjacent in the sample genome but not in the reference. Two clusters are merged if they share the read or if they agree on the position and length of the structural variant. This information is inferred from read alignment orientation and distance.

ReadBroker

- Interchromosomal translocations yield chimeric read pairs where 1 read aligns to one chromosome and its partner aligns to another. Because Grouper examines each chromosome individually, the ReadBroker step is performed to join the information from chimeric read pairs across chromosomes.

After ReadBroker

- **SmallAssembler**—Assembles reads in clusters into contigs using a *de Bruijn* method and iteratively assembles reads into contigs until all reads in the cluster are assembled. It also produces a file containing the reads that were used to assemble the contig, with a realignment to the contig sequence.
- **SpanContigs**—Uses the presence of nearby anomalous read pairs to determine whether to extend the search range used by the subsequent AlignContig step from its default.
- **AlignContig**—Computes a dynamic programming alignment of a contig to a region of the reference genome; merges full or partial duplicate calls of the same event into a single call.
- **VariantFilter**—Removes all structural variants that overlap with gaps identified in UCSC gaps. The UCSC gaps file defines regions of the genome that have not been sequenced.
- **DeletionGenotyper**—Assigns a genotype to all deletions.
- **SomaticGenotyper**—Assigns a quality score (Q-score) to all structural variants. Higher Q-scores indicate a higher probability that this structural variant is somatic.
- **DeletionGenotyper**—Assigns a genotype to all deletions.

CNV Variant Caller

The CNV variant caller is designed to identify copy number variants (CNVs) in diploid genomes using Hidden Markov Models (HMM) or unbalanced Haar wavelets. The method adopts a count-based approach for CNV calling, which comprises the following steps:

1. Pre-processing step, during which read depth is computed at each position and then filtered based on CpG islands, assembly gaps, telomeric/centromeric regions. Either alignability tracks or coverage tracks obtained from a pool of reference sample are used to normalize the data. Counts or count ratios are produced as an output.
2. Segmentation of read counts/ratios using fixed or variable bin size and a copy number assignment.

Normalization

A single sample or a pool of reference samples is used for normalization, by deriving a ratio between a test and the reference. Window size is fixed (by default to 100 bp). The HMM model with Gaussian emission distribution is used for segmentation. A bin exclusion criterion (less than 10% of build coverage in both samples) is applied.

The reference for CNV normalization is an alignability measure that is meant to gauge the probability of a position aligning to a single unique region of the genome. In detail, the notion of alignability for reads of length k is as follows: given a map M that, for a fixed read length k and any position P in a genome G, stores at M(P) the number of occurrences in G of the k-mer that starts at P for a given position P in G, define the overlap set of P as the k-mers that overlap P. The alignability of P is the proportion of this overlap set that is unique.

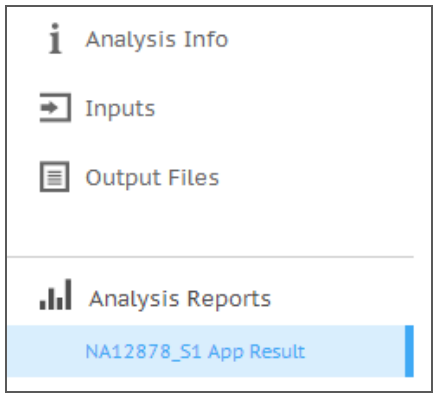
Variant Scoring

After copy number assignment, each CNV call is assigned a quality score based on a two-sample t-test. Each counts/ratio in a 1 kb window on each size of a breakpoint (or half the length of a variant call, whichever is smaller) is compared using t-test. This test is based on the null hypothesis that there is no difference in coverage on each size of the breakpoint. Obtained p-values are then reported as Q-scores on a Phred scale as -10 log10.

Analysis Output

To view the results, click the **Projects** tab, then the project name, and then the analysis.

Figure 2 Output Navigation Bar



After analysis is complete, access the output through the left navigation bar.

- **Analysis Info**—Information about the analysis session, including log files.
- **Inputs**—Overview of input settings.
- **Output Files**—Output files for the sample.
- **Sample Analysis Reports**—Analysis reports for each sample.

Analysis Info

The Analysis Info page displays the analysis settings and execution details.

Row Heading	Definition
Name	Name of the analysis session.
Application	App that generated this analysis.
Date Started	Date and time the analysis session started.
Date Completed	Date and time the analysis session completed.
Duration	Duration of the analysis.
Session Type	Number of nodes used.
Status	Status of the analysis session. The status shows either Running or Complete.

Log Files

Click the **Log Files** link to access the app log files. You can locate some log files in **Output Files**.

File Name	Description
CompletedJobInfo.xml	Contains information about the completed analysis session.

File Name	Description
Logging.zip	Contains all detailed log files for each step of the workflow.
WorkflowError.txt	Workflow standard error output (contains error messages created when running the workflow).
WorkflowLog.txt	Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).
monoOut.txt	Wrapper mono call standard output that contains the command calling the workflow and anything that WorkflowLog.txt does not catch. Available when the mono component does not work as expected.

NOTE
For explanation about mono, see www.mono-project.com.

Output Files

The Output Files page provides access to the output files for each sample analysis.

- BAM Files
- VCF Files
- Genome VCF Files
- Resequencing File
- Summary File

BAM File Format

A BAM file (*.bam) is the compressed binary version of a SAM file that is used to represent aligned sequences up to 128 Mb. SAM and BAM formats are described in detail at https://samtools.github.io/hts-specs/SAMv1.pdf.

BAM files use the file naming format of SampleName_S#.bam, where # is the sample number determined by the order that samples are listed for the run.

BAM files contain a header section and an alignments section:

- **Header**—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.
- **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and the match descriptor string.

The alignments section includes the following information for each or read pair:

- **RG:** Read group, which indicates the number of reads for a specific sample.
- **BC:** Barcode tag, which indicates the demultiplexed sample ID associated with the read.
- **SM:** Single-end alignment quality.
- **AS:** Paired-end alignment quality.
- **NM:** Edit distance tag, which records the Levenshtein distance between the read and the reference.
- **XN:** Amplicon name tag, which records the amplicon tile ID associated with the read.

BAM index files (*.bam.bai) provide an index of the corresponding BAM file.

VCF File Format

VCF is a widely used file format developed by the genomics scientific community that contains information about variants found at specific positions in a reference genome.

VCF files use the file naming format SampleName_S#.vcf, where # is the sample number determined by the order that samples are listed for the run.

VCF File Header—Includes the VCF file format version and the variant caller version. The header lists the annotations used in the remainder of the file. If MARS is listed, the Illumina internal annotation algorithm annotated the VCF file. The VCF header includes the reference genome file and BAM file. The last line in the header contains the column headings for the data lines.

VCF File Data Lines—Each data line contains information about a single variant.

VCF File Headings

Heading	Description
CHROM	The chromosome of the reference genome. Chromosomes appear in the same order as the reference FASTA file.
POS	The single-base position of the variant in the reference chromosome. For SNPs, this position is the reference base with the variant; for indels or deletions, this position is the reference base immediately before the variant.
ID	The rs number for the SNP obtained from dbSNP.txt, if applicable. If there are multiple rs numbers at this location, the list is semicolon delimited. If no dbSNP entry exists at this position, a missing value marker ('.') is used.
REF	The reference genotype. For example, a deletion of a single T is represented as reference TT and alternate T. An A to T single nucleotide variant is represented as reference A and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T is represented as reference A and alternate AT. An A to T single nucleotide variant is represented as reference A and alternate T.
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is $10^{-(Q/10)}$. For example, the set of Q30 calls has a 0.1% error rate. Many variant callers assign quality scores based on their statistical models, which are high in relation to the error rate observed.

VCF File Annotations

Heading	Description
FILTER	If all filters are passed, **PASS** is written in the filter column. • **LowDP**—Applied to sites with depth of coverage below a cutoff. • **LowGQ**—The genotyping quality (GQ) is below a cutoff. • **LowQual**—The variant quality (QUAL) is below a cutoff. • **LowVariantFreq**—The variant frequency is less than the given threshold. • **R8**—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. • **SB**—The strand bias is more than the given threshold. Used with the Somatic Variant Caller and GATK.
INFO	Possible entries in the INFO column include: • **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**—The total number of alleles in called genotypes. • **CD**—A flag indicating that the SNP occurs within the coding region of at least 1 RefGene entry. • **DP**—The depth (number of base calls aligned to a position and used in variant calling). • **Exon**—A comma-separated list of exon regions read from RefGene. • **FC**—Functional Consequence. • **GI**—A comma-separated list of gene IDs read from RefGene. • **QD**—Variant Confidence/Quality by Depth. • **TI**—A comma-separated list of transcript IDs read from RefGene.
FORMAT	The format column lists fields separated by colons. For example, GT:GQ. The list of fields provided depends on the variant caller used. Available fields include: • **AD**—Entry of the form X,Y, where X is the number of reference calls, and Y is the number of alternate calls. • **DP**—Approximate read depth; reads with MQ=255 or with bad mates are filtered. • **GQ**—Genotype quality. • **GQX**—Genotype quality. GQX is the minimum of the GQ value and the QUAL column. In general, these values are similar; taking the minimum makes GQX the more conservative measure of genotype quality. • **GT**—Genotype. 0 corresponds to the reference base, 1 corresponds to the first entry in the ALT column, and so on. The forward slash (/) indicates that no phasing information is available. • **NL**—Noise level; an estimate of base calling noise at this position. • **PL**—Normalized, Phred-scaled likelihoods for genotypes. • **SB**—Strand bias at this position. Larger negative values indicate less bias; values near 0 indicate more bias. Used with the Somatic Variant Caller and GATK. • **VF**—Variant frequency; the percentage of reads supporting the alternate allele.
SAMPLE	The sample column gives the values specified in the FORMAT column.

Genome VCF Files

Genome VCF (gVCF) files are VCF v4.1 files that follow a set of conventions for representing all sites within the genome in a reasonably compact format. The gVCF files include all sites within the region of interest in a single file for each sample.

The gVCF file shows no-calls at positions with low coverage, or where a low-frequency variant (< 3%) occurs often enough (> 1%) that the position cannot be called to the reference. A genotype (GT) tag of **./.** indicates a no-call.

For more information, see sites.google.com/site/gvcftools/home/about-gvcf.

Summary File

The BWA WGS App produces an overview of statistics for each sample and the aggregate results in a comma-separated values (CSV) format. The *resquencing_ summary.csv file contains the same data as the Sample Summary Report but is formatted for easier analysis. These files are located in the results folder for each sample and the aggregate results.

Statistic	Definition
Sample ID	IDs of samples reported on in the file.
Run Folder	Run folders for samples reported on in the file.
Fragment length median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Fragment length min	Minimum length of the sequenced fragment.
Fragment length max	Maximum length of the sequenced fragment.
Fragment length SD	Standard deviation of the sequenced fragment length.
Number of Reads	Total number of reads passing filter for this sample.
Percent Aligned (per read)	Percentage of reads passing filter that aligned.
Percent Q30 (per read)	The percentage of bases with a quality score of 30 or higher.
MismatchRate (per read)	The average percentage of mismatches across both reads 1 and 2 over all cycles.
SNVs All	Total number of Single Nucleotide Variants present in the data set passing the quality filters.
SNVs Passing Filters	SNVs passing variants filter.
SNVs (Percent found in dbSNP)	100*(Number of SNVs in dbSNP/Number of SNVs).

Statistic	Definition
SNV Ts/Tv ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).
SNV Het/Hom ratio	Number of heterozygous SNVs/Number of homozygous SNVs.
Indels	Total number of indels present in the data set passing the quality filters.
Insertions Passing Variants	Insertions passing variant filters.
Deletions Passing Variants	Deletions passing variant filters.
Indels (Percent found in dbSNP)	100*(Number of Indels in dbSNP/Number of Indels).
Insertions (Percent found in dbSNP)	100*(Number of insertions in dbSNP/ Number of insertions)
Deletions(Percent found in dbSNP)	100*(Number of deletions in dbSNP/ Number of deletions)
Indel Het/Hom ratio	Number of heterozygous indels/Number of homozygous indels.
Insertion Het/Hom ratio	Ratio of the number of heterozygous to homozygous insertions.
Deletion Het/Hom ratio	Ratio of the number of heterozygous to homozygous deletions.
SmallVariantStatisticsFlag	Flags whether SmallVariantStatistics was run (1 means that it was run)
SVStatisticsFlag	Flags whether SVStatistics was run (1 means that it was run)
CNVStatisticsFlag	Flags whether CNVStatistics was run (1 means that it was run)

Sample Analysis Reports

The BWA WGS App provides an overview of statistics per sample in the Analysis Reports sample pages. To download the statistics, click **PDF Summary Report**.

- Alignment Summary
- Small Variants
- Structural Variants
- Coverage Histogram Summary

Alignment Summary

Table 2 Alignment Summary

Statistic	Definition
Number of Reads	Total number of reads passing filter for this sample.
Coverage	Total number of aligned bases divided by the genome size.
Percent Duplicate Paired Reads	Percentage of paired reads that have duplicates.
Fragment Length Median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Fragment Length Standard Deviation	Standard deviation of the sequenced fragment length.

Table 3 Read Level Statistics

Statistic	Definition
Percent Aligned	The percentage of reads passing filter that aligned to the reference genome.
Percent Q30	The percentage of bases with a quality score of 30 or higher.
Mismatch Rate	The average percentage of mismatches across both reads 1 and 2 over all cycles.

Small Variants Summary

Table 4 Small Variants Summary

Statistic	Definition
Total Passing	The total number of variants present in the data set that passed the variant quality filters.
Percent Found in dbSNP	100*(Number of variants in dbSNP/Number of variants).
Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.
Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T).

Table 5 Variants by Sequence Context

Statistic	Definition
Number in Genes	The number of variants that fall into a gene.
Number in Exons	The number of variants that fall into an exon.
Number in Coding Regions	The number of variants that fall into a coding region.
Number in UTR Regions	The number of variants that fall into an untranslated region (UTR).
Number in Mature microRNA	The number of variants that fall into a mature microRNA.
Number in Splice Site Regions	The number of variants that fall into a splice site region.

Table 6 Variants by Consequence

Statistic	Definition
Frameshifts	The number of variants that cause a frameshift.
Non-synonymous	The number of variants that cause an amino acid change in a coding region.
Synonymous	The number of variants that are within a coding region, but do not cause an amino acid change.
Stop Gained	The number of variants that cause an additional stop codon.
Stop Lost	The number of variants that cause the loss of a stop codon.

Structural Variants Summary

This table breaks structural variant output into the classes of variants called, and reports the total number and their overlap with annotated genes. All counts are based on PASS filter variants.

Variant Class	Definition
CNV	A copy-number variation (CNV) is a large category of structural variation, which includes insertions, deletions and duplications. CNVs are generally greater than 10 kb. CNVs below 10 kb are filtered but still present in the VCF file.
Insertion	In an insertion, nucleotides are added between 2 adjacent nucleotides in the reference sequence. The insertions in the structural variants category are 51 bp or greater.
Tandem Duplication	In a tandem duplication, a segment of a chromosome is duplicated front to end, with both segments in the same orientation. The segments are 51 bp or greater.

Variant Class	Definition
Deletion	In a deletion, contiguous nucleotides are absent compared to the reference sequence. The deletions in the structural variants category are 51 bp or greater.
Inversion	An inversion is a chromosome rearrangement in which a segment of a chromosome is reversed end to end. An inversion occurs when a single chromosome undergoes breakage and rearrangement within itself. The segments are 51 bp or greater.

Coverage Histogram

The coverage histogram shows the number of reference bases plotted against the depth of coverage (read depth). The features include the following:

- The drop-down menu lets you look at the overall picture, or highlight a particular chromosome.
- The **Fix Y Scale** checkbox lets you keep the Y Scale the same when comparing multiple chromosomes.
- The **Export TSV** link lets you to export the coverage data in a tab-separated txt. file.

Revision History

Document	Date	Description of Change
Document # 15050952 v01	January 2016	Reorganized topics, updated writing style.

Notes

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 7 Illumina General Contact Information

Website	www.illumina.com
Email	techsupport@illumina.com

Table 8 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com