Quick-StartGuide

GeneRead DNAseq variant analysis

This guide describes how to analyze sequencing reads generated from a GeneRead DNAseq Gene Panel kit for targeted exon enrichment. The first section provides instructions for uploading reads, setting analysis parameters, and downloading the variant report. Next, technical details are provided for how the sequence reads are mapped to the reference genome. Thirdly, the method for variant identification and annotation from the aligned reads is described. Finally, definitions for the terms used in the variant report are provided.

Minimum computer requirements

GeneRead Targeted Exon Enrichment Panel Data Analysis is web-based and compatible with the following web browsers:

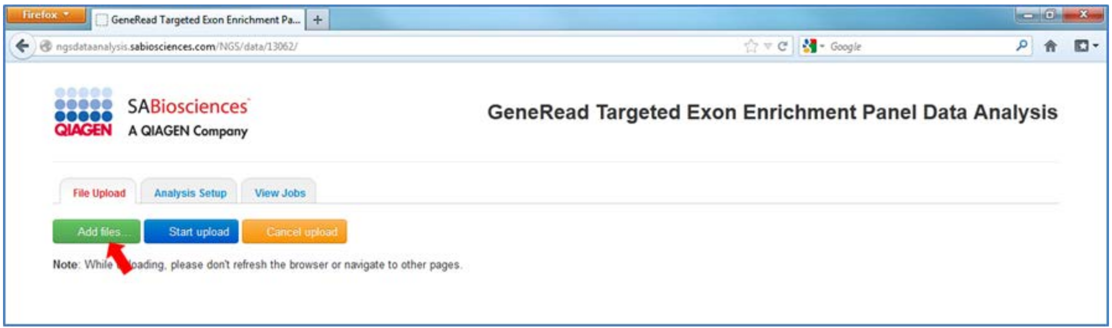
- Mozilla® Firefox®
- Google® Chrome®

Note: We apologize for the inconvenience, but due to technical issues our software is not compatible with Microsoft® Internet Explorer®.

Getting started with GeneRead Targeted Exon Enrichment Panel Data Analysis

1. **Go to http://ngsdataanalysis.sabiosciences.com.**
2. **Click on "Add files" and select either the .SFF file (generated from Ion Torrent™ PGM) or the .FASTQ file (generated from Illumina® sequencers such as MiSeq).**

   Note: Read files should contain reads from only one sample. For example, if your sequencing library contains barcoded samples, the reads must already be split by sample.

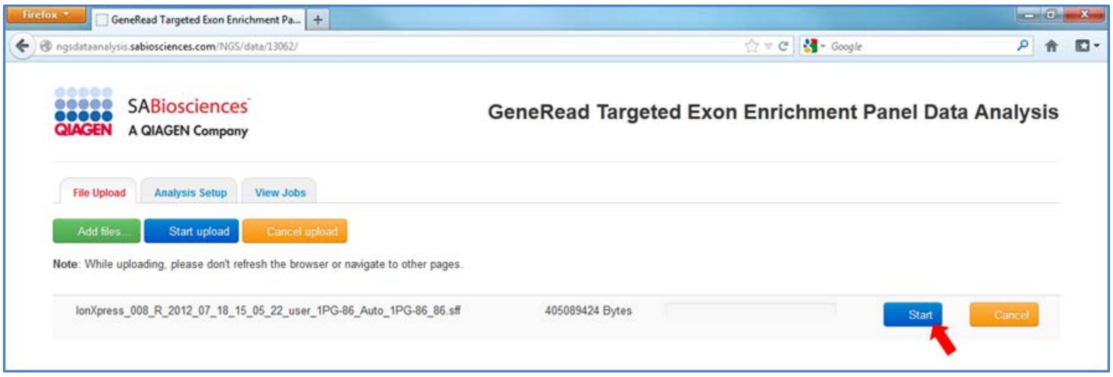


GeneRead Targeted Exon Enrichment Panel Data Analysis

Note: While uploading, please don't refresh the browser or navigate to other pages.

Sample & Assay Technologies

3. **Click “Start” to start uploading the selected file.**



Firefox
GeneRead Targeted Exon Enrichment Pa...
ngsdataanalysis.sabiosciences.com/NGS/data/13062/
Google
SABiosciences
QIAGEN
A QIAGEN Company

GeneRead Targeted Exon Enrichment Panel Data Analysis

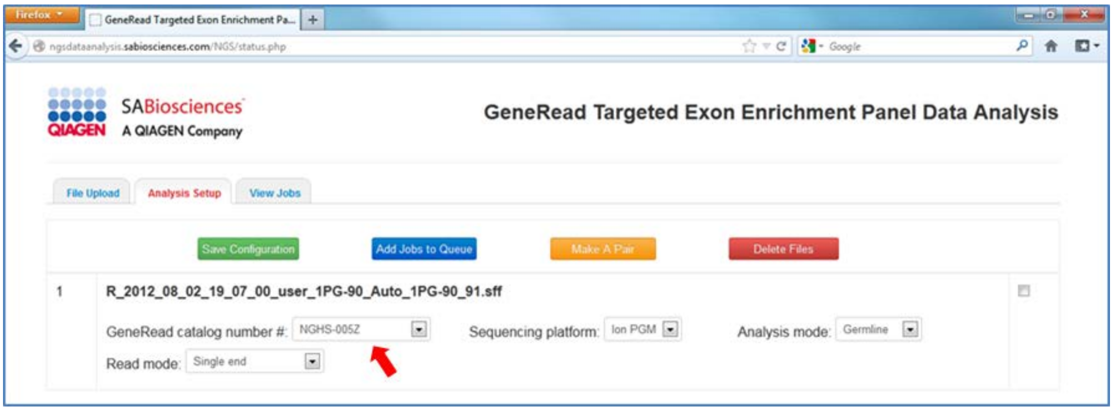
Note: While uploading, please don't refresh the browser or navigate to other pages.

IonXpress_008_R_2012_07_18_15_05_22_user_1PG-86_Auto_1PG-86_86.sff 405089424 Bytes

4. **The upload process may take a few minutes or longer to complete.**

5. **Select several pieces of information about each sequencing file in the “Analysis Setup” tab:**

   - **GeneRead catalog number**
   - **Sequencing platform: Ion Torrent PGM / Ion Proton™ or Illumina**
   - **Analysis Mode: Germline or Somatic**
   - **Read Mode: Single-end or Paired-end**



Firefox
GeneRead Targeted Exon Enrichment Pa...
ngsdataanalysis.sabiosciences.com/NGS/status.php
Google
SABiosciences
QIAGEN
A QIAGEN Company

GeneRead Targeted Exon Enrichment Panel Data Analysis

1 R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.sff ☐

GeneRead catalog number #: NGHS-005Z

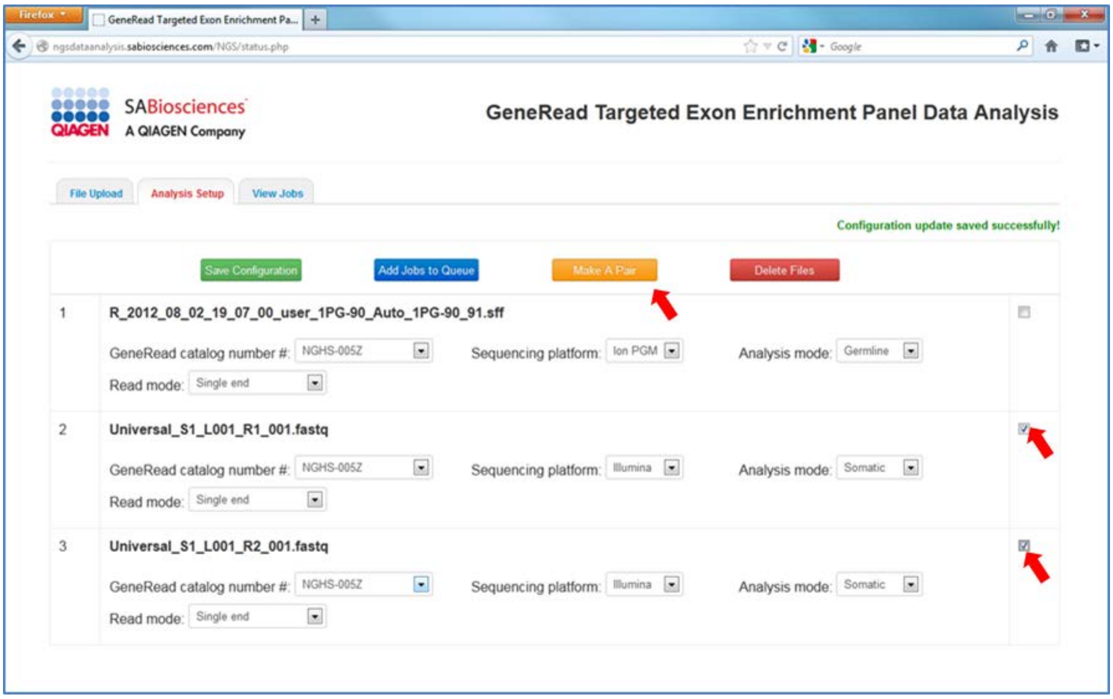
Sequencing platform: Ion PGM

Analysis mode: Germline

Read mode: Single end

6. **For paired-end analysis, specify which file pairs correspond to paired-end reads (“Make A Pair”). Select the first and second file in a pair, and click on “Make A Pair”.**

   Note: The files must be paired two at a time, but the order in which the files are paired is not important. The “Read Mode” for both files will be automatically updated to “Paired-end.”



GeneRead Targeted Exon Enrichment Panel Data Analysis

Configuration update saved successfully!

1 R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.sff □

GeneRead catalog number #: NGHS-005Z Sequencing platform: Ion PGM Analysis mode: Germline

Read mode: Single end

2 Universal_S1_L001_R1_001.fastq ☑

GeneRead catalog number #: NGHS-005Z Sequencing platform: Illumina Analysis mode: Somatic

Read mode: Single end

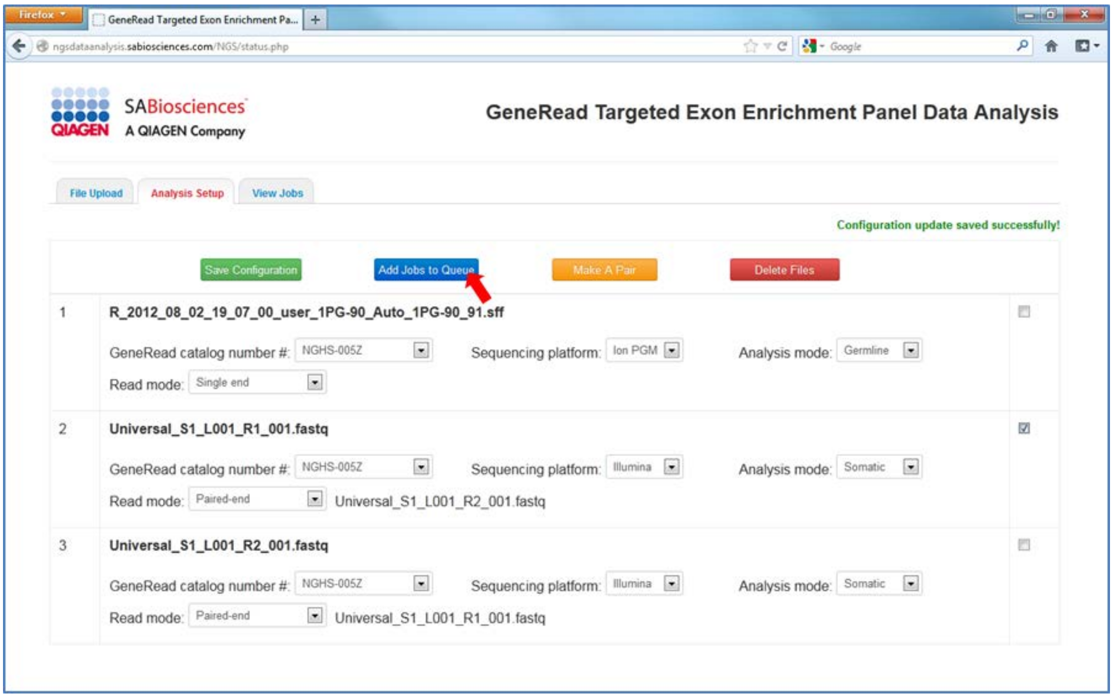
3 Universal_S1_L001_R2_001.fastq ☑

GeneRead catalog number #: NGHS-005Z Sequencing platform: Illumina Analysis mode: Somatic

Read mode: Single end

7. **Select the file to analyze and click on "Add Jobs to Queue." For a paired-end run, select only the first file in the pair. Analysis will start.**

   Note: Analysis will typically take 30 minutes to a few hours. Run time depends on the file size, the number of files, and how many processes are running concurrently on the servers.



GeneRead Targeted Exon Enrichment Panel Data Analysis

Configuration update saved successfully!

1 R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.sff □

GeneRead catalog number #: NGHS-005Z Sequencing platform: Ion PGM Analysis mode: Germline

Read mode: Single end

2 Universal_S1_L001_R1_001.fastq ☑

GeneRead catalog number #: NGHS-005Z Sequencing platform: Illumina Analysis mode: Somatic

Read mode: Paired-end Universal_S1_L001_R2_001.fastq

3 Universal_S1_L001_R2_001.fastq □

GeneRead catalog number #: NGHS-005Z Sequencing platform: Illumina Analysis mode: Somatic

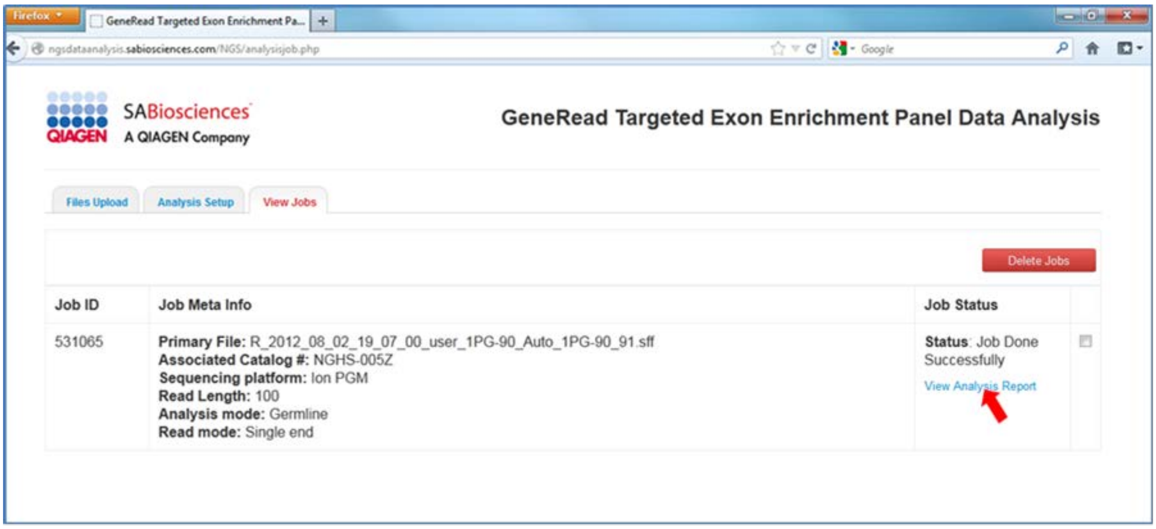
Read mode: Paired-end Universal_S1_L001_R1_001.fastq

8. **The status and results of the analysis can be checked by clicking the "View Jobs" tab.**

9. **While analysis is running, the status will display, “Job in progress.”**

10. **The status will change to “Job Done Successfully” when analysis of the reads is complete.**

11. **Click on “View Analysis Report” to see the results summary and links to download the results.**



GeneRead Targeted Exon Enrichment Panel Data Analysis

Job ID	Job Meta Info	Job Status	
531065	**Primary File:** R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.sff **Associated Catalog #:** NGHS-005Z **Sequencing platform:** Ion PGM **Read Length:** 100 **Analysis mode:** Germline **Read mode:** Single end	**Status**: Job Done Successfully View Analysis Report	□

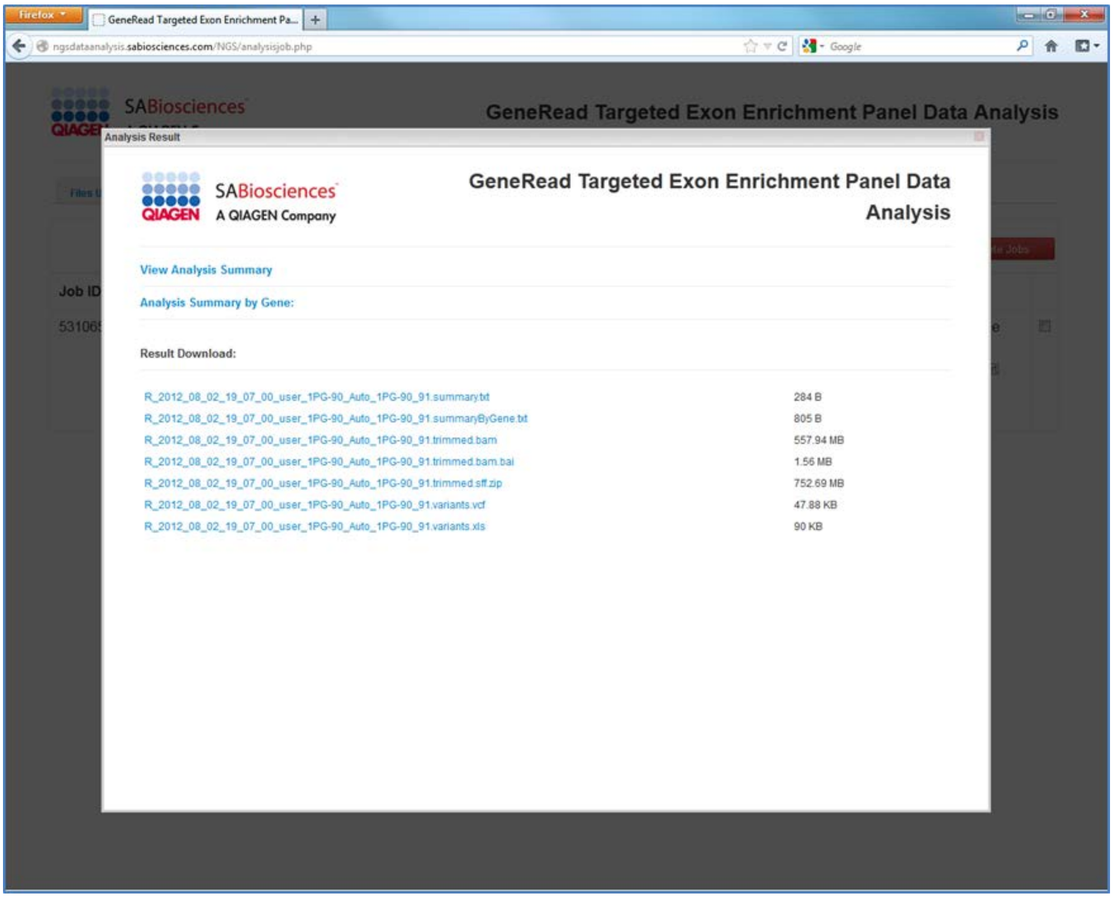
12. **Files available for download are:**

   - **summary.txt: summary statistics for read mapping and variant calling**
   - **summaryByGene.txt: summary statistics for read mapping and variant calling by gene**
   - **reads.trimmed.sff (IonTorrent) or reads.trimmed.fastq (Illumina): reads in the same format as they were uploaded, after quality filtering and primer trimming**
   - **reads.trimmed.bam: reads aligned to human genome reference**
   - **reads.trimmed.bam.bai: index for BAM file (which is necessary to view the alignments in IGV)**
   - **variants.vcf: variants in variant call format (VCF)**
   - **variants.xls: variants in Excel® format**

Note: Finished jobs and the associated data file will be stored at the analysis site for a maximum of 30 days. After that, all files will be automatically deleted.

Note: See the following sections for technical background on how the software maps sequence reads and identifies variants.

Note: Analysis files available for download will appear in a pop-up box.



Firefox
GeneRead Targeted Exon Enrichment Pa...
ngsdataanalysis.sabiosciences.com/NGS/analysisjob.php
Google
SABiosciences
GeneRead Targeted Exon Enrichment Panel Data Analysis
Job ID
Analysis Result

SABiosciences A QIAGEN Company

GeneRead Targeted Exon Enrichment Panel Data Analysis

View Analysis Summary

Analysis Summary by Gene:

Result Download:

R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.summary.txt	284 B
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.summaryByGene.txt	805 B
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.trimmed.bam	557.94 MB
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.trimmed.bam.bai	1.56 MB
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.trimmed.sff.zip	752.69 MB
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.variants.vcf	47.88 KB
R_2012_08_02_19_07_00_user_1PG-90_Auto_1PG-90_91.variants.xls	90 KB

Read alignment workflow

The uploaded sequence reads are aligned to the reference genome in several steps.

Preliminary alignment to reference genome

A preliminary alignment is performed using the full read set. Illumina reads (such as from MiSeq) are aligned using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Ion Torrent PGM and Ion Proton reads are aligned using TMAP (https://github.com/nh13/TMAP/tree/master/doc). Reads are aligned with hard-clipping on the 5′ end of the read, and soft-clipping on the 3′ end of the read.

Trimming primer region from reads

Priming sites are identified, and primer regions are trimmed at primer binding sites. The trimmed read file is available for download for users who wish to perform their own alignment and variant calling.

Quality filter

Reads with an untrimmed length of less than 45 bp are discarded. These short reads can result from either undesired PCR products or poor sequencing quality.

Final read alignment

Next, final alignment of trimmed reads to the reference genome is performed. Alignment parameters are identical to those used in the preliminary alignment. This final read alignment is used for variant calling.

Alignment summary

Finally, basic read and coverage statistics are generated, including the total number of reads mapped, the median depth of coverage, and the number of targeted bases covered at a depth of at least 10, 30, and 100. An alignment summary is provided for the entire gene panel, and also for each gene included in the gene panel.

Variant calling workflow

This section describes the steps involved in the GeneRead variant calling pipeline. The pipeline incorporates tools and programs from multiple sources, including many that are part of the Genome Analysis Toolkit (GATK: http://www.broadinstitute.org/gatk/) from the Broad Institute.

Inputs

The following files are inputs for the variant calling pipeline:

- reads.trimmed.bam: This file contains the mapped reads after primer trimming.
- target.bed: This file contains the coordinates of the target amplicons. This file is automatically generated from the catalog number of the GeneRead DNAseq Gene Panel used.
- params.txt: This file contains various parameters for variant calling. Some user-supplied parameters are included, including sequencing platform (such as Ion Torrent PGM or MiSeq) and the desired workflow (germline or somatic).

Variant calling

Software used for variant calling varies depending on the sequencing platform. For Ion Torrent / Ion Proton data, the Torrent Variant Caller (TVC) version 2.2 is used for calling the variants. For Illumina (MiSeq) data, the GATK Unified Genotyper program (GATKLite version 2.1–8) is used. Both programs generate a variant call format (VCF) file containing both SNPs and indels identified from the input reads. The GeneRead data analysis workflow runs the GATK Variant Annotator program using the output from TVC in order to populate the INFO field in the VCF file with parameters needed for downstream filtering. The INFO field in the GATK Unified Genotyper output is already populated with all necessary parameters, so the Variant Annotator step is not necessary for these data (Figure 8).

Variant filtering

Variant filtering is performed in two stages. Stage 1 of the filtering marks variants that fail some of the thresholds for variant calling. However, note that the variants that fail the filters are not removed from the VCF file. Instead, the FILTER field in the VCF file is marked with the names of the filters failed by the variants. We use two filters derived from the recommendations in the GATK best practices document.

GatkSNPFilter includes the following thresholds:

- Fisher Strand (FS: Fisher's exact test for strand bias). Variants with FS>60 are marked.
- RMS Mapping Quality (MQ). Variants with MQ<40 are marked.
- ReadPosRankSum. Variants with ReadPosRankSum<–8.0 are marked.

GatkIndelFilter includes the following thresholds:

- Variants with FS>200 are marked.
- Variants with ReadPosRankSum<–20.0 are marked.

Stage 2 of filtering removes SNPs and indels that do not pass certain variant frequency and coverage thresholds. Unlike in Stage 1, the variants that fail Stage 2 are removed from the VCF file. The following filters are used:

- SNPs with <4% variant allele frequency are removed in the somatic workflow, and SNPs with <20% variant allele frequency are removed in the germline workflow.
- Indels with <20% variant allele frequency are removed in the somatic workflow, and indels with <25% variant allele frequency are removed in the germline workflow.

Functional annotation

After filtering, the variants undergo various annotation steps to add useful information. The following annotation steps are part of the variant calling workflow:

- snpEff annotations: The snpEff program (http;//snpeff.sourceforge.net/) predicts the effects of variants on genes. Information such as gene name, transcript ID, codon change, animo acid change, and others are added.
- dbSNP binding: If the variant is present in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), the corresponding dbSNP ID(s) is added to the variant.
- Cosmic binding: If the variant is present in Cosmic (http://www.sanger.ac.uk/perl/genetics/CGP/cosmic), the corresponding Cosmic ID is added to the variant.
- dbNSFP annotations: SNPSift (http://snpeff.sourceforge.net/SnpSift.html) is used to extract available annotations from dbNSFP (database of nonsynonymous SNPs' functional predictions). Annotations derived from dbNSFP include SIFT score, Polyphen score, Interpro domain, and Uniport ID.

The final output from the variant calling pipeline is a VCF file named "variants.vcf", which contains all the added annotations. This VCF file is also converted to Excel format ("variants.xls") for convenience.

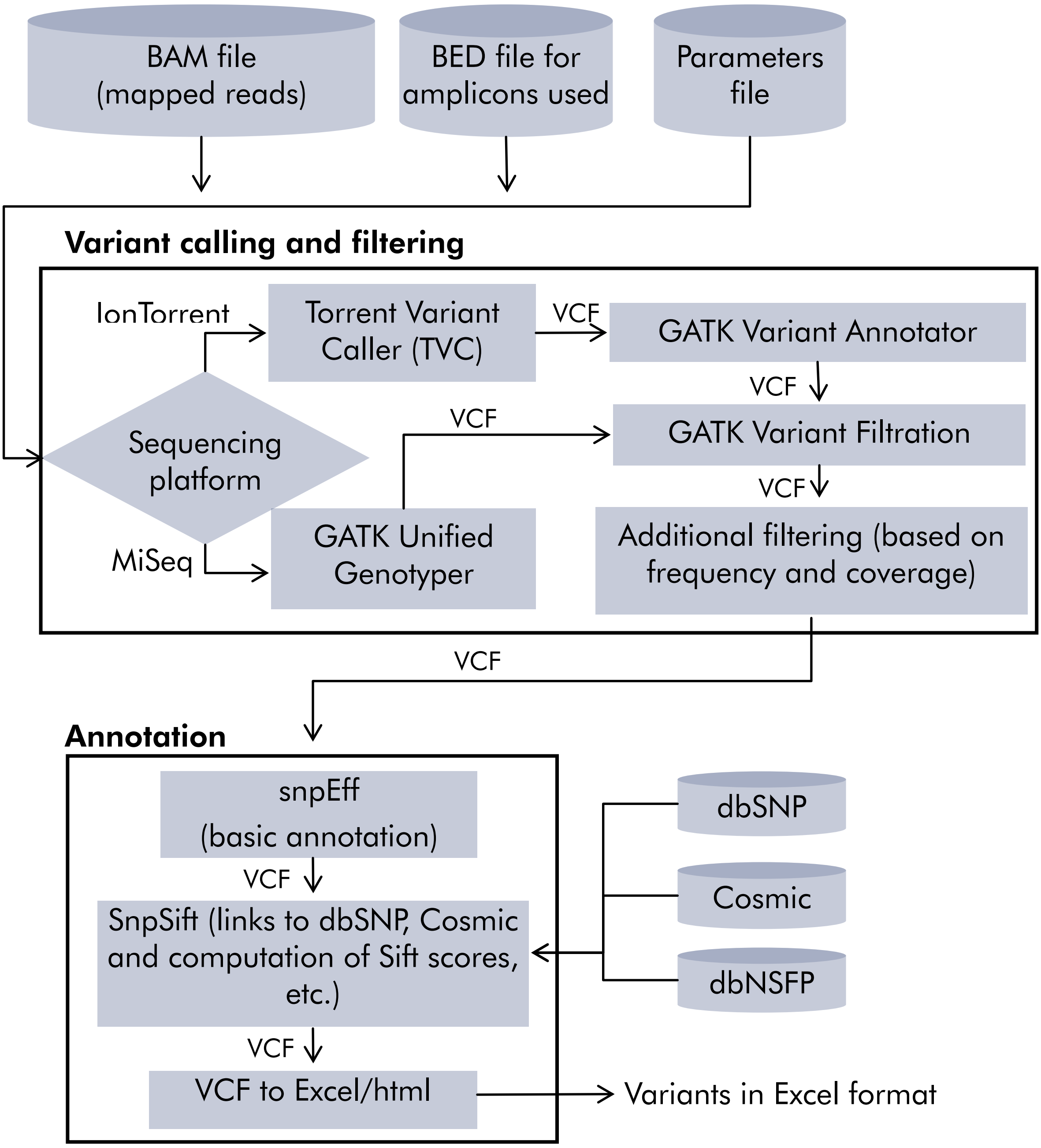


Figure 1. GeneRead variant calling workflow.

Functional annotation

After filtering, the variants undergo various annotation steps to add useful information. The following annotation steps are part of the variant calling workflow:

- snpEff annotations: The snpEff program (http;//snpeff.sourceforge.net/) predicts the effects of variants on genes. Information such as gene name, transcript ID, codon change, animo acid change, and others are added.
- dbSNP binding: If the variant is present in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), the corresponding dbSNP ID(s) is added to the variant.
- Cosmic binding: If the variant is present in Cosmic (http://www.sanger.ac.uk/perl/genetics/CGP/cosmic), the corresponding Cosmic ID is added to the variant.
- dbNSFP annotations: SNPSift (http://snpeff.sourceforge.net/SnpSift.html) is used to extract available annotations from dbNSFP (database of nonsynonymous SNPs' functional predictions). Annotations derived from dbNSFP include SIFT score, Polyphen score, Interpro domain, and Uniport ID.

The final output from the variant calling pipeline is a VCF file named "variants.vcf", which contains all the added annotations. This VCF file is also converted to Excel format ("variants.xls") for convenience.

Terms and definitions in the variant report

Position: Position within the chromosome.

ID: dbSNP or Cosmic ID. Currently, dbSNP version 135 and cosmic version 60 are used. This field can have multiple IDs separated by a space.

Ref: Reference allele at the position. UCSC hg19 is currently used as the reference. For point mutations, the reference allele will be a single nucleotide. For InDels or MNPs, the reference allele can be longer than one nucleotide.

Alt: A list of alternate alleles separated by commas.

Gene Name: Gene within which the mutation is contained.

Mutations Type: Type of mutation or variant. The value will be either SNP, MNP, INS (insertion), or DEL (deletion).

Codon Change: Change in the nucleotide space. Codon change is presented in the format c.999A>B, where "999" is the chromosomal position, "A" is the reference allele, and "B" is the alternate allele.

AA Change: Change in the amino acid space. AA change is presented in the format p.A999B, where "A" is a string with the reference amino acid(s), "999" is the first position within the residue that is different, and "B" is a string with the alternative amino acid(s) present in the variant.

Filtered Coverage: Coverage depth which is taken into consideration in variant calling. This does not include reads with low base quality at the current position, and hence is generally smaller than the actual read depth at the position.

Allele Coverage: Coverage depth of each allele at the position. The REF allele is listed first, followed by each of the alleles in ALT field, in the same order as they appear in the ALT field.

Allele Frequency: Frequency of the REF and ALT alleles, listed in the same order as in the "Allele Coverage" field.

Variant Frequency: Frequency of the most common ALT allele (within the sample) at the current position.

SNPEFF Effect: Most significant effect of the variant, as predicted by snpEff. For the list of all possible effects, please refer to http://snpeff.sourceforge.net/faq.html#What_effects_are_predicted?.

SNPEFF Impact: Impact of the SNP, as predicted by snpEff. The value will be listed as "HIGH," "MODERATE," "LOW," or "MODIFIER," in decreasing order of importance. For detailed information about the possible values, please refer to http://snpeff.sourceforge.net/faq.html#How_is_impact_categorized?_(VCF_output).

Filter: "PASS" indicates that variant passes all the filters. Otherwise, this field contains a comma separated list of filters that the variant files. "BroadSNPFilter" indicates that the variant fails some of the thresholds recommended by the Broad Institute for SNPs, as listed below:

- MQ<40 (MQ: RMS Mapping Quality) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_RMSMappingQuality.html
- FS>60 (FS: Fisher Strand) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_FisherStrand.html
- ReadPosRankSum<–8.0 http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_ReadPosRankSumTest.html

"BroadIndelFilter" indicates that the variant fails some of the thresholds recommended by the Broad Institute for indels:

- FS>200
- ReadPosRankSum<–20.0

P-Value: The Phred quality score of the SNP converted to a p-value. The value indicates the likelihood that the variant call is a false positive.

SIFT Score: A score assigned by the SIFT program that indicates if the amino acid substitution affects protein function. Please refer to http://sift.jcvi.org/.

Interpro Domain: Domain or conserved site in which the variant is located.

Uniprot Acc: Uniprot accession number.

Polyphen Score: Score computed by the Polyphen program that predicts the impact of amino acid substitution of the structure and function of the protein. Please refer to http://genetics.bwh.harvard.edu/pph/pph_help_text.html.

For up-to-date licensing and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN handbooks can be requested from QIAGEN Technical Service or your local QIAGEN distributor. Selected handbooks can be downloaded from www.qiagen.com/literature. Safety data sheets (SDS) for any QIAGEN product can be downloaded from www.qiagen.com/safety.

GeneRead DNAseq Gene Panels are intended for molecular biology applications. These products are not intended for the diagnosis, prevention, or treatment of a disease.

Trademarks: QIAGEN® (QIAGEN Group); Chrome®, Google® (Google, Inc.); Illumina® (Illumina, Inc.); Ion Proton™, Ion Torrent™ (Life Technologies Corporation); Excel®, Internet Explorer®, Microsoft® (Microsoft Corporation); Firefox®, Mozilla® (Mozilla Foundation).

www.qiagen.com
Australia ■ 1-800-243-800
Austria ■ 0800/281010
Belgium ■ 0800-79612
Canada ■ 800-572-9613
China ■ 021-51345678
Denmark ■ 80-885945
Finland ■ 0800-914416
France ■ 01-60-920-930
Germany ■ 02103-29-12000
Hong Kong ■ 800 933 965
Ireland ■ 1800 555 049
Italy ■ 800 787980
Japan ■ 03-6890-7300
Korea (South) ■ 1544 7145
Luxembourg ■ 8002 2076
The Netherlands ■ 0800 0229592
Norway ■ 800-18859
Singapore ■ 65-67775366
Spain ■ 91-630-7050
Sweden ■ 020-790282
Switzerland ■ 055-254-22-11
UK ■ 0808-2343665
USA ■ 800-426-8157

QIAGEN

Sample & Assay Technologies