Running workflow "gatk-variant-call-paired-datasets-collection"

Step 1: Input dataset

reads in paired dataset collection

type to filter

Step 2: Input dataset collection

reference genome

type to filter

Step 3: FASTQ Groomer(version 1.0.4)

File to groom
Output dataset 'output' from step 2

Input FASTQ quality scores type
Sanger & Illumina 1.8+

Advanced Options
Hide Advanced Options

Step 4: Bowtie2(version 2.2.6.2)

Is this single or paired library
Paired-end Dataset Collection

FASTQ Paired Dataset
Output dataset 'output_file' from step 3

Write unaligned reads (in fastq format) to separate file(s)
Not available.

Write aligned reads (in fastq format) to separate file(s)
Not available.

Do you want to set paired-end options?
Yes

Set the minimum fragment length for valid paired-end alignments
Not available.

Set the maximum fragment length for valid paired-end alignments
500

Select the upstream/downstream mate orientations for a valid paired-end alignment against the forward reference strand
--fr

Disable no-mixed behavior
Not available.

Disable no-discordant behavior
Not available.

Allow mate dovetailing
Not available.

Disallow one mate alignment to contain another

Not available.

Disallow mate alignments to overlap
Not available.

Will you select a reference genome from your history or use a built-in index?
Use a genome from the history and build index

Select reference genome
Output dataset 'output' from step 1

Set read groups information?
Set read groups (SAM/BAM specification)

Auto-assign
Not available.

Read group identifier (ID)
group_id

Auto-assign
Not available.

Read group sample name (SM)
group_sample_name

Platform/technology used to produce the reads (PL)
ILLUMINA

Auto-assign
Not available.

Library name (LB)
Not available.

Sequencing center that produced the read (CN)
Not available.

Description (DS)
Not available.

Date that run was produced (DT)
Not available.

Flow order (FO)
Not available.

The array of nucleotide bases that correspond to the key sequence of each read (KS)
Not available.

Programs used for processing the read group (PG)
Not available.

Predicted median insert size (PI)
Not available.

Platform unit (PU)
Not available.

Select analysis mode
1: Default setting only

Do you want to use presets?
No, just use defaults

Save the bowtie2 mapping statistics to the history
Not available.

Step 5: Realigner Target Creator(version 2.8.0)

Choose the source for the reference list
History

BAM file
Output dataset 'output' from step 4

Using reference file
Output dataset 'output' from step 1

Known Variants

Basic or Advanced GATK options
Advanced

Pedigree files

Pedigree strings

How strict should we be in validating the pedigree information
STRICT

Read Filters

Operate on Genomic intervals

Exclude Genomic intervals

Interval set rule
UNION

Amount of padding (in bp) to add to each interval
Not available.

Type of reads downsampling to employ at a given locus
NONE

Type of BAQ calculation to apply in the engine
OFF

BAQ gap open penalty (Phred Scaled)
40.0

Use the original base quality scores from the OQ tag
Not available.

Value to be used for all base quality scores, when some are missing
-1

How strict should we be with validation
STRICT

Interval merging rule
ALL

Read group black lists

Disable experimental low-memory sharding functionality.
Not available.

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run
Not available.

Fix mis-encoded base quality scores. Q0 == ASCII 33 according to the SAM specification, whereas Illumina encoding starts at Q64. The idea here is simple: we just iterate over all reads and subtract 31 from every quality score.
Not available.

Basic or Advanced Analysis options
Advanced

Window size for calculating entropy or SNP clusters (windowSize)
10

Fraction of base qualities needing to mismatch for a position to have high entropy (mismatchFraction)
0.15

Minimum reads at a locus to enable using the entropy calculation (minReadsAtLocus)
4

Maximum interval size
500

Step 6: Indel Realigner(version 2.8.0)

Choose the source for the reference list
History

BAM file
Output dataset 'output' from step 4

Using reference file
Output dataset 'output' from step 1

Restrict realignment to provided intervals
Output dataset 'output_interval' from step 5

Known Variants

LOD threshold above which the realigner will proceed to realign
5.0

Use only known indels provided as RODs
Not available.

Basic or Advanced GATK options
Advanced

Pedigree files

Pedigree strings

How strict should we be in validating the pedigree information
STRICT

Read Filters

Operate on Genomic intervals

Exclude Genomic intervals

Interval set rule
UNION

Amount of padding (in bp) to add to each interval
Not available.

Type of reads downsampling to employ at a given locus
NONE

Type of BAQ calculation to apply in the engine
OFF

BAQ gap open penalty (Phred Scaled)
40.0

Use the original base quality scores from the OQ tag

Not available.

Value to be used for all base quality scores, when some are missing
-1

How strict should we be with validation
STRICT

Interval merging rule
ALL

Read group black lists

Disable experimental low-memory sharding functionality.
Not available.

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run
Not available.

Fix mis-encoded base quality scores. Q0 == ASCII 33 according to the SAM specification, whereas Illumina encoding starts at Q64. The idea here is simple: we just iterate over all reads and subtract 31 from every quality score.
Not available.

Basic or Advanced Analysis options
Advanced

percentage of mismatching base quality scores at a position to be considered having high entropy
0.15

Simplify BAM
Not available.

Consensus Determination Model
USE_READS

Maximum insert size of read pairs that we attempt to realign
3000

Maximum positional move in basepairs that a read can be adjusted during realignment
200

Max alternate consensuses to try
30

Max reads (chosen randomly) used for finding the potential alternate consensuses
120

Max reads allowed at an interval for realignment
20000

Don't output the original cigar or alignment start tags for each realigned read in the output bam
Not available.

Step 7: Unified Genotyper(version 2.8.0)

Choose the source for the reference list
History

BAM files

> **BAM file 1**
>
> **BAM file**
> Output dataset 'output_bam' from step 6

Using reference file

Output dataset 'output' from step 1

Provide a dbSNP Reference-Ordered Data (ROD) file
Don't set dbSNP

Genotype likelihoods calculation model to employ
BOTH

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be called
30.0

The minimum phred-scaled confidence threshold at which variants not at 'trigger' track sites should be emitted (and filtered if less than the calling threshold)
30.0

Basic or Advanced GATK options
Advanced

Pedigree files

Pedigree strings

How strict should we be in validating the pedigree information
STRICT

Read Filters

Operate on Genomic intervals

Exclude Genomic intervals

Interval set rule
UNION

Amount of padding (in bp) to add to each interval
Not available.

Type of reads downsampling to employ at a given locus
NONE

Type of BAQ calculation to apply in the engine
OFF

BAQ gap open penalty (Phred Scaled)
40.0

Use the original base quality scores from the OQ tag
Not available.

Value to be used for all base quality scores, when some are missing
-1

How strict should we be with validation
STRICT

Interval merging rule
ALL

Read group black lists

Disable experimental low-memory sharding functionality.
Not available.

Makes the GATK behave non deterministically, that is, the random numbers generated will be different in every run
Not available.

Fix mis-encoded base quality scores. Q0 == ASCII 33 according to the SAM specification, whereas Illumina encoding starts at Q64. The idea here is simple: we just iterate over all reads and subtract 31 from every quality score.

Not available.

Basic or Advanced Analysis options
Advanced

Heterozygosity value used to compute prior likelihoods for any locus
0.001

The PCR error rate to be used for computing fragment-based likelihoods
0.0001

How to determine the alternate allele to use for genotyping
DISCOVERY

Should we output confident genotypes (i.e. including ref calls) or just the variants?
EMIT_VARIANTS_ONLY

Compute the SLOD
Not available.

Minimum base quality required to consider a base for calling
17

Maximum fraction of reads with deletions spanning this locus for it to be callable
0.05

Maximum number of alternate alleles to genotype
6

Minimum number of consensus indels required to trigger genotyping run
5

Heterozygosity for indel calling
0.000125

Indel gap continuation penalty
10

Indel gap open penalty
45

Indel haplotype size
80

Vary gap penalties by context
Not available.

Annotation Types
None

Additional annotations

Annotation Interfaces/Groups
Nothing selected.

Annotations to exclude
None

Ploidy (number of chromosomes) per sample. For pooled data, set to (Number of samples in each pool * Sample Ploidy)
2

Step 8: SnpEff(version 4.0.0)

Sequence changes (SNPs, MNPs, InDels)
Output dataset 'output_vcf' from step 7

Input format

VCF

Output format
VCF (only if input is VCF)

Genome source
Named on demand

Snpff Genome Version Name (e.g. GRCh38.76)
hg19

Upstream / Downstream length
5000 bases

Set size for splice sites (donor and acceptor) in bases
2 bases

Annotation options
Nothing selected.

Use custom interval file for annotation
Selection is Optional

Only use the transcripts in this file.
Selection is Optional

Filter output
Nothing selected.

Filter out specific Effects
No

Chromosomal position
Use default (based on input type)

Text to prepend to chromosome name
Not available.

Produce Summary Stats
True

Do not report usage statistics to server
True

Send results to a new history

tools used in this workflow:

"tool_id": null,
"tool_id": null,
"tool_id": "toolshed.g2.bx.psu.edu/repos/devteam/fastq_groomer/fastq_groomer/1.0.4",
"tool_id": "toolshed.g2.bx.psu.edu/repos/devteam/bowtie2/bowtie2/2.2.6.2",
"tool_id": "toolshed.g2.bx.psu.edu/repos/iuc/gatk2/gatk2_realigner_target_creator/2.8.0",
"tool_id": "toolshed.g2.bx.psu.edu/repos/iuc/gatk2/gatk2_indel_realigner/2.8.0",
"tool_id": "toolshed.g2.bx.psu.edu/repos/iuc/gatk2/gatk2_unified_genotyper/2.8.0",
"tool_id": "toolshed.g2.bx.psu.edu/repos/iuc/snpeff/snpEff/4.0.0",