**Clinical Genome Analytics**

**Workflow for Whole Exome Analysis**

**Prepared by**

**Anuj Srivastava and Grace A Stafford**

**Note** most of these applications will name the output based on the input. In this document, the example file names will be based on this convention. Specific input for each usage example are bracketed by < > ; parameters will be set and remain unchanged.

**Assumption**: job initiated in same directory as data files. If not, include path to data. **NOTE** that subsequent commands assume output path of prior commands will not change.

**Step1: NGS QC**

**Module Requirement:**

module load NGSQCToolkit/2.3

* 1. **Quality Statistics generation and bad read filtering**

**Illumina sequencing; CASAVA 1.8+; paired end reads; human data**

**Usage:**

IlluQC\_PRLL.pl -pe <read1.fastq> <read2.fastq> <adaptor\_file> <A> –s 30 -c 12 -o filter

**Input:**

* Forward paired end file [required]
* Reverse paired end file [required]
* Adaptor\_database file or application library [required]
* FASTQ encoding [required]

**Output:**

Folder named “filter” created here which has the following files

* read1.fastq\_filtered (here)
* read2.fastq\_filtered (here)
* \*\_unPaired\_HQReads
* \*\_stat
* \*.html,
* 6 plots for each fastq file\*\_avgQual.png, \*\_baseCompostion.png, \*\_QualRangePerBase.png, \*\_ gcDistribution.png, \*\_ qualDistribution.png, \*\_ QualRangePerBase.png,
* \*\_summary.png (summary plot)

**Parameters Description:**

* -pe: Forward reverse paired end file
* Adaptor\_file: Should be replaced by known adaptor sequence used in particular type of sequencing; if not available, use the option dependending on type of sequence from below:

Primer/Adaptor libraries:

1 = Genomic DNA/Chip-seq Library

2 = Paired End DNA Library

3 = DpnII gene expression Library

4 = NlaIII gene expression Library

5 = Small RNA Library

6 = Multiplexing DNA Library

N = Do not filter for Primer/Adaptor

* FASTQ quality value variants: 5 stands for = Illumina (1.8+) (Phred+33, 33 to 7); if not known then please use A [automatic detection]; various options are given below

FASTQ variants:

1 = Sanger (Phred+33, 33 to 73)

2 = Solexa (Phred+64, 59 to 104)

3 = Illumina (1.3+) (Phred+64, 64 to 104)

4 = Illumina (1.5+) (Phred+64, 66 to 104)

5 = Illumina (1.8+) (Phred+33, 33 to 74)

A = Automatic detection of FASTQ variant

* -s: -cutOffQualScore (The cut-off value for PHRED quality score for high-quality filtering)
* -c: Number of CPUs to be used
* -o: Output will be stored in the given folder; ***provide name*** (filter here)

Note: Check here: Inside \*\_stat file that *Percentage of HQ filtered reads*” should be greater than 50 percent; if not then exit with a message that “greater than 50% of reads filtered out (data is of bad quality)

**Snippet of \*\_stat file:**

Library type Paired-end

Input files data\_1.fastq data\_2.fastq

Primer/Adaptor library Paired End DNA Library

Cut-off read length for HQ 70%

Cut-off quality score 30

Only statistics Off

Number of CPUs 12

QC statistics

File name data\_1.fastq data\_2.fastq

Total number of reads 37205349 37205349

Total number of HQ reads 33912757 33912757

Percentage of HQ reads 91.15% 91.15%

Total number of bases 2827606524 2827606524

Total number of bases in HQ reads 2577369532 2577369532

Total number of HQ bases in HQ reads 2518231490 2518391414

Percentage of HQ bases in HQ reads 97.71% 97.71%

Number of Primer/Adaptor contaminated HQ reads 3023 22

Total number of HQ filtered reads 33909713 33909713

**Percentage of HQ filtered reads 91.14% 91.14%**

Detailed QC statistics

File name data\_1.fastq data\_2.fastq data\_1.fastq\_filtered data\_2.fastq\_filtered

Minimum read length 76 76 76 76

Maximum read length 76 76 76 76

Average read length 76.00 76.00 76.00 76.00

Total number of reads 37205349 37205349 33909713 33909713

Total number of reads with non-ATGC bases 870598 309108 713977 247870

Percentage of reads with non-ATGC bases 2.34% 0.83% 2.11% 0.73%

Total number of bases 2827606524 2827606524 2577138188 2577138188

Total number of HQ bases 2669093673 2665709183 2518010286 2518167765

Percentage of HQ bases 94.39% 94.27% 97.71% 97.71%

Total number of non-ATGC bases 901356 694930 715483 321321

Percentage of non-ATGC bases 0.03% 0.02% 0.03% 0.01%

**1.2 Quality based trimming of filtered files**

**Usage:**

TrimmingReads.pl –i <filter/read1.fastq\_filtered> -irev <filter/read2.fastq\_filtered> –q 30

**Input:**

* Forward paired end file with paired end intact after filtering [required]
* Reverse paired end file with paired end intact after filtering [required]

**Output:**

* read1.fastq\_filtered\_trimmed (here)
* read2.fastq\_filtered\_trimmed (here)

**Parameters Description:**

* -i <Forward read/sequence file>
* -irev <Reverse read/sequence file of paired-end data>
* -q | -qualCutOff <Integer> (Only for FASTQ files) Cut-off PHRED quality score for trimming reads from right end (3' end)

**Reference:**

Patel et al. (2012) “NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PloS ONE*

**Step2: Alignment**

**Module Requirements:**

* module load samtools/0.1.18
* module load bwa/0.5.10

**2.1 Running BWA “aln” step**

**Usage:**

bwa aln -t 12 /pathTo/hg19.fasta <read1.fastq\_filtered\_trimmed> > aln1.sai

bwa aln -t 12 /pathTo/hg19.fasta <read2.fastq\_filtered\_trimmed> > aln2.sai

**Input:**

* Reference Genome Indexed with BWA [required]
* Read file forward; passed QC [required]
* Read file reverse; passed QC [required]

Note: Need to run bwa “aln” separately for forward and reverse file

**Output:**

* Suffix array index file; **provide file names** (aln1.sai and aln2.sai here)

**Parameters Description:**

|  |  |
| --- | --- |
| * **-t** *INT* | Number of threads (multi-threading mode) |

**2.2 Running BWA “sampe” step**

**Usage:**

bwa sampe –r @RG"\t"ID:<SAMPLE1\_RG1>"\t"LB:<SAMPLE1\_LIB1>"\t"SM:<SAMPLE1>"\t"PL:<ILLUMINA>/pathTo/reference.fasta <aln1.sai> <aln2.sai> <read1.fastq\_filtered\_trimmed> <read2.fastq\_filtered\_trimmed> > aln.sam

**Input:**

* Suffix array index file (aln1.sai and aln2.sai here) [required]
* Input fasta file (read1.fastq\_filtered\_trimmed read2.fastq\_filtered\_trimmed here)[required]
* -r: read group information (required for downstream analysis)

**Output:**

* \*.sam Alignment file in SAM format; **provide file names** (aln.sam here)

**Parameters Description:**

|  |  |
| --- | --- |
| * **-r STR** | Specify the read group in a format like ‘@RG\tID:foo\tSM:bar’. [null] |

Note: If we continue to use the same conventions as GES uses for Genologics and the HiSeq at JAX BHB, the read group info can be derived from the fastq file.

Given a pair of fastq files, the @RG info can be pulled from the name (of either of pair) and the first line of the file. Highlighted colors match read group info needed.

0916-6180\_GES12\_36721\_TTAGGC\_L007\_R2\_ALL.fastq

@HISEQ2000:158:D0R99ACXX:7:1101:1387:1965 2:N:0:TTAGGC

GTGTCTGTGGGGCCTTCCGCTCTGGCTCCTGTGGA

+

BB@FFFFFHHHHGJJFIJJIGIJJJIGJJJEBGII

@RG

ID HISEQ2000:158:D0R99ACXX:7 (the machine name, run#, flowcell ID, lane #)

LB GES12\_36721\_TTAGGC (GES sample ID with barcode)

PL ILLUMINA (platform)

SM 0916-6180 (customer sample ID)

if the given fastq file not in above format [need a script to check it]; then add timestamp to

these information:

@RG

ID ID\_timestamp

LB LIB\_timestamp

PL ILLUMINA

SM SAMPLE\_timestamp

**References:**

Li et al. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*

Li et al. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*

**Step3: Variant Pre-processing**

**Module requirements (or path):**

* module load python
* module load java
* module load R
* /opt/compsci/picard/1.8.4/
* /opt/compsci/GATK/2.2-16/GenomeAnalysisTK.jar

**Step3.1: Picard SortSam.jar (generating sorted alignment bam file)**

**Usage:**

java -Xmx2g -jar picard/SortSam.jar \

SO=coordinate \

INPUT=<aln.sam> \

OUTPUT=aln.bam \

VALIDATION\_STRINGENCY=SILENT \

CREATE\_INDEX=true

**Input:**

* The BWA generated sam file (aln.sam here) [required]

**Output:**

* The coordinate sorted alignment file in bam format; **provide file name** (aln.bam here)
* Index of coordinate sorted alignment file (aln.bai here)

**Parameters Description:**

|  |  |
| --- | --- |
| * SO: SORT\_ORDER. If not supplied OUTPUT is in the same order as INPUT. Default value: null. Possible values: {unsorted, queryname, coordinate} |  |

* VALIDATION\_STRINGENCY: Validation stringency for all SAM files read by this program. Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded. Default value: STRICT. This option can be set to 'null' to clear the default value. Possible values: {STRICT, LENIENT, SILENT}
* CREATE\_INDEX: Whether to create a BAM index when writing a coordinate-sorted BAM file. Default value: false. This option can be set to 'null' to clear the default value. Possible values: {true, false}

**Step3.2: Picard Mark Duplicates (Removing duplicates in BAM file)**

**Usage:**

java -Xmx2g -jar picard/MarkDuplicates.jar \

INPUT=<aln.bam>  \

OUTPUT=aln\_dedup.bam \

METRICS\_FILE=metrics \

REMOVE\_DUPLICATES=true \

CREATE\_INDEX=true \

VALIDATION\_STRINGENCY=LENIENT

**Input:**

* Coordinated sorted BAM file (aln.bam here) [required]
* File to write duplication metrics to (metrics here) [Required]

**Output:**

* Duplicate removed bam file; **provide file name** (aln\_dedup.bam here)
* Index of duplicate removed bam file (aln\_dedup.bai here)

**Parameters Description:**

* REMOVE\_DUPLICATES: If true do not write duplicates to the output file instead of writing them with appropriate flags set. Default value: false. This option can be set to 'null' to clear the default value. Possible values: {true, false}
* METRICS\_FILE: File to write duplication metrics

**Step3.3: Realignment around indels Part I (Target Interval Creation)**

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-I <aln\_dedup.bam> \

-R hg19.fasta \

--num\_threads 8 \

-T RealignerTargetCreator \

-o forIndelRealigner.intervals \

--known /pathTo/< Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf\

--known /pathTo/< 1000G\_phase1.indels.hg19.vcf

**Input**

* Duplicated removed bam file (aln\_dedup.bam here) [required]
* Reference genome file (hg19.fasta here) [required]

**Output**

* Interval file output to be used by IndelRealigner; ***provide name*** (forIndelRealigner.intervals here)

**Parameters Description:**

* --known: Input VCF file with known indels
* --num\_threads: How many data threads should be allocated to running this analysis

**Databases could be used with this command (-known):**

* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

Note: index files (\*.idx) must be available in same directory as the \*.vcf files

Note: we will use “known” files provided by the Broad GATK bundle. We will use the files that best correspond to the reference genome used.

**Step3.4: Realignment around indels Part II (Performing the local re-alignment)**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-I <aln\_dedup.bam> \

-R hg19.fasta \

-T IndelRealigner \

-targetIntervals <forIndelRealigner.intervals> \

-o realignedBam.bam \

-known /pathTo/< Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf \

-known /pathTo/< 1000G\_phase1.indels.hg19.vcf

**Input:**

* Duplicated removed bam file (aln\_dedup.bam here) [required]
* Reference genome file (hg19.fasta here) [required]
* Interval file created by RealignerTargetCreator (forIndelRealigner.intervals here) [required]

**Output:**

* Bam file with realigned indels; ***provide name*** (realignedBam.bam here)
* Index file of realigned bam file (realignedBam.bai here)

**Parameters Description:**

* -known: Input VCF file with known indels

**Databases could be used with this command (-known):**

* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

**Step3.5: Base Quality Recalibration Part I**

Notes: Requires R packages (see appendix) to create graphs.

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T BaseRecalibrator \

-I <realignedBam.bam> \

- nct 8 \

-R resources/hg19.fasta \

-knownSites /pathTo/<dbsnp\_137.hg19.vcf\

-knownSites /pathTo/<Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf\

-knownSites /pathTo/<1000G\_phase1.indels.hg19.vcf\

-o <recal\_data.grp>

--plot\_pdf\_file Pre\_recal\_data.pdf

**Input:**

* Realigned file realignedBam.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]
* -knownSites (database of known SNP and Indel sites) [not required]
* File name to store pre recalibration plots; ***provide name*** (Pre\_recal\_data.pdf here) [not required]

**Output:**

* output recalibration table file Recal\_data.grp
* recalibration plots pdf file

**Parameters Description:**

* -knownSites: A database of known polymorphic sites to skip over in the recalibration algorithm
* --plot\_pdf\_file: The output recalibration pdf file to create

**Databases could be used with this command (-knownSites):**

* The most recent dbSNP release (build ID > 132) [dbsnp\_137.hg19.vcf]
* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

**Step 3.5: Base Quality Recalibration Part II (**Just creating plots after model generation)

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T BaseRecalibrator \

-I <realignedBam.bam> \

-nct 8 \

-R resources/hg19.fasta \

-BQSR recal\_data.grp \

-knownSites /pathTo/<dbsnp\_137.hg19.vcf\

-knownSites /pathTo/<Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf\

-knownSites /pathTo/<1000G\_phase1.indels.hg19.vcf\

-o Post\_recal\_data.grp \

--plot\_pdf\_file Post\_recal\_data.pdf

**Input:**

* Realigned file realignedBam.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]
* output recalibration table file (recal\_data.grp here) [required]
* -knownSites (database of known SNP and Indel sites) [not required unless supplied in previous step]
* File name to store pre recalibration plots ; ***provide name***(Post\_recal\_data.pdf here) [not required]

**Output:**

* post recalibration table file
* Post recalibration plots pdf file

**Parameters Description:**

* -knownSites: A database of known polymorphic sites to skip over in the recalibration algorithm
* --plot\_pdf\_file: The output recalibration pdf file to create
* -BQSR: The input covariates table file which enables on-the-fly base quality score recalibration

**Databases could be used with this command (-knownSites):**

* The most recent dbSNP release (build ID > 132) [dbsnp\_137.hg19.vcf]
* Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf
* 1000G\_phase1.indels.hg19.vcf (currently from the 1000 Genomes Phase I indel calls)

**Step3.5: Base Quality Recalibration Part III (**printing the reads**)**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T PrintReads \

-R hg19.fasta \

-nct 8 \

-I <realignedBam.bam> \

-BQSR <recal\_data.grp> \

-o realigned\_BQSR.bam

**Input:**

* Realigned file realignedBam.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]
* output recalibration table file (recal\_data.grp here) [required]
* Outfile name with fixed base qualities [required]

**Output:**

* Bamfile with fixed base qualities; *provide name* (realigned\_BQSR.bam here)
* Index files of Bam alignment (realigned\_BQSR.bai here)

**Parameters Description:**

* PrintReads (commands to print read from fixed BQSR file)

Note: The file size becomes almost double after BQSR step as additional information been added to sam fields.

**Step 3.6: Picard CalculateHsMetrics (**generating target enrichment information**)**

**Usage:**

java -Xmx2g -jar CalculateHsMetrics.jar \

BAIT\_INTERVALS=<bait.txt> \

TARGET\_INTERVALS=<target.txt> \

INPUT=<realigned\_BQSR.bam> \

OUTPUT=Metricsfile.txt \

VALIDATION\_STRINGENCY=LENIENT

**Input**

* Realigned base quality fixed file realigned\_BQSR.bam (here) [required]
* bait interval file (bait here) [required]
* target interval file (target here) [required]

**Output:**

* Output file with target enrichment information; provide name (Metricsfile.txt here) [required]

**Parameters Description:**

* BAIT\_INTERVAL: An interval list file that contains the locations of the baits used. [Required]
* TARGET\_INTERVALS: An interval list file that contains the locations of the targets [Required]

**References:**

* **McKenna A et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data Genome Res**
* **DePristo et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics**
* <http://picard.sourceforge.net>

**Note: The \_INTERVALS files require specific formatting, including header information that corresponds to the “@” lines from BWA generated SAM file in step 2.2. The file should be in five columns with chromosome names matching those in genome file. A snippet of target file is given below; bait file should be in same format:**

@HD VN:1.0 GO:none SO:coordinate

@SQ SN:chrM LN:16571

@SQ SN:chr1 LN:249250621

@SQ SN:chr2 LN:243199373

@SQ SN:chr3 LN:198022430

@SQ SN:chr4 LN:191154276

@SQ SN:chr5 LN:180915260

@SQ SN:chr6 LN:171115067

@SQ SN:chr7 LN:159138663

@SQ SN:chr8 LN:146364022

@SQ SN:chr9 LN:141213431

@SQ SN:chr10 LN:135534747

@SQ SN:chr11 LN:135006516

@SQ SN:chr12 LN:133851895

@SQ SN:chr13 LN:115169878

@SQ SN:chr14 LN:107349540

@SQ SN:chr15 LN:102531392

@SQ SN:chr16 LN:90354753

@SQ SN:chr17 LN:81195210

@SQ SN:chr18 LN:78077248

@SQ SN:chr19 LN:59128983

@SQ SN:chr20 LN:63025520

@SQ SN:chr21 LN:48129895

@SQ SN:chr22 LN:51304566

@SQ SN:chrX LN:155270560

@SQ SN:chrY LN:59373566

@RG ID:Leukemia\_test PL:ILLUMINA LB:SAMPLE1\_LIB1 SM:SAMPLE1

@PG ID:bwa PN:bwa VN:0.5.9-r26-dev

chr1 69114 69332 + --

chr1 69383 69577 + --

chr1 69644 69940 + --

chr1 69951 70028 + --

chr1 566170 566275 + --

**NOTE: the values of the 4th and 5th columns are not important in this step; merely their presence.**

**Step 4: Variant Calling**

**Module requirements:**

* module load python
* module load java
* module load R
* /opt/compsci/GATK/2.2-16/GenomeAnalysisTK.jar

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-R resources/hg19.fasta \

-T UnifiedGenotyper \

-nt 8 \

-I <realigned\_BQSR.bam> \

-glm BOTH \

--dbsnp /pathTo/dbsnp\_137.hg19.vcf \

-o variants.raw.vcf \

-stand\_call\_conf 50.0 \

-stand\_emit\_conf 10.0 \

-dcov 200 \

**-L <targets.bed>**

**Input:**

* Realigned base quality fixed file realigned.fixed\_BQSR.bam (here) [required]
* Reference genome file (hg19.fasta here) [required]

**Output:**

* SNP and indel calls in variant call format; provide name (variants.raw.vcf here)
* VCF index file (variants.raw.vcf.idx here)

**Parameter description:**

-stand\_call\_conf: The minimum phred-scaled confidence threshold at which variants should be called.

-stand\_emit\_conf: The minimum phred-scaled confidence threshold at which variants should be emitted (and filtered with LowQual if less than the calling threshold). This argument allows one to emit low quality calls as filtered records.

**Databases could be used with this command:**

These tools do NOT require known sites, but if SNPs are provided with the -dbsnp argument they will use them for variant annotation. We use this file:

* The most recent dbSNP release (build ID > 132) [dbsnp\_137.hg19.vcf]
* Bed file for target intervals -L target\_intervals.bed (agilent target list)

Note: If present, bed file must match reference genome used for alignment (eg, it should not contain chr\_un or chr\_random if using UCSC hg19). It should end with any of the following extension (.bed, .list, .picard, .interval\_list, or .intervals) and should be coordinate sorted. I prepared a whole exome file from UCSC using refseq annotation; it’s on Data for AL dir on rockhopper [whole\_genome\_refseq\_hg19\_24\_chr\_only.bed]; for test dataset use Target.bed in same directory.

Snippet of target\_intervals.bed (no header is should like the file in step 3.6):

chr1 69090 70008

chr1 566176 566276

chr1 801942 802433

chr1 861307 861407

chr1 865534 865716

chr1 866394 866494

**Step5: Variant Quality Recalibration**

Note: The UnifiedGenotyper uses a fundamentally different likelihood model when calling different classes of variation and so therefore the VQSR must be run separately for SNPs and INDELs to build separate adaptive error models.

**Module requirements:**

* module load python
* module load java
* module load R
* /opt/compsci/GATK/2.2-16/GenomeAnalysisTK.jar

**Step 5.1: Select SNP**

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-T SelectVariants \

-R reference/hg19.fasta \

--variant <variants.raw.vcf> -o output\_SNP.vcf -selectType SNP

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP/indels calls in variant call format (variants.raw.vcf here) [required]

**Output:**

* VCF file with SNP only; **provide name** (output\_SNP.vcf here)
* VCF file with SNP only (output\_SNP.vcf.idx here)

**Parameter description:**

* -selectType: Selecting the SNPs only

**Step 5.2: Recalibrate SNP**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T VariantRecalibrator \

-R reference/hg19.fasta \

--num\_threads 8 \

-input <output\_SNP.vcf> \

-resource:<hapmap,known=false,training=true,truth=true,prior=15.0 hapmap\_3.3.hg19.vcf> \

-resource:<omni,known=false,training=true,truth=false,prior=12.0 1000G\_omni2.5.hg19.vcf> \

-resource:<dbsnp,known=true,training=false,truth=false,prior=6.0 dbsnp\_137.hg19.vcf> \

-an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ \

-mode SNP \

-recalFile pathTo/output.recal \

-tranchesFile pathTo/output.tranches \

-rscriptFile pathTo/output.plots.R \

--maxGaussians 4 \

-percentBad 0.05

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP/indels calls in variant call format (output\_SNP.vcf here) [required]

**Output:**

* Recal file to be used by ApplyRecalibration; **provide name** (output.recal)
* Tranches file to be used by ApplyRecalibration; **provide name** (output.tranches)
* The output rscript file generated by the VQSR to aid in visualization of the input data and learned model ; ***provide name*** (output.plots.R here)
* Plots of recalibration; ***provide name*** (Output.plots.R.pdf and output.tranches.pdf here)

**Parameter description:**

* -mode: Recalibration mode to employ: 1.) SNP for recalibrating only snps (emitting indels untouched in the output VCF); 2.) INDEL for indels
* -rscriptFile: The output rscript file generated by the VQSR to aid in visualization of the input data and learned model
* **-**resource: A list of sites for which to apply a prior probability of being correct but which aren't used by the algorithm
* -an: The names of the annotations which should used for calculations. Description given below:

1. **QD: QualByDepth**: Variant confidence (from the QUAL field) / unfiltered depth Low scores are indicative of false positive calls and artifacts. Note that QualByDepth requires sequencing reads associated with the samples with polymorphic genotypes.
2. **HaplotypeScore:** Consistency of the site with two (and only two) segregating haplotypes. Higher scores are indicative of regions with bad alignments, often leading to artifactual SNP and indel calls. Note that the Haplotype Score is only calculated for sites with read coverage.
3. **MQRankSum:** The u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with ref bases vs. those with the alternate allele) Note that the mapping quality rank sum test cannot be calculated for sites without a mixture of reads showing both the reference and alternate alleles. )
4. **ReadPosRankSum:** The 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; if the alternate allele is only seen near the ends of reads this is indicative of error). Note that the read position rank sum test cannot be calculated for sites without a mixture of reads showing both the reference and alternate alleles.
5. **FS:** Phred-scaled p-value using Fisher's Exact Test to detect strand bias (the variation being seen on only the forward or only the reverse strand) in the reads? More bias is indicative of false positive calls. Note that the fisher strand test may not be calculated for certain complex indel cases or for multi-allelic site
6. **MQ: Mapping Quality**
7. **InbreedingCoeff:** Likelihood-based (using PL field) test for the inbreeding among samples. A continuous generalization of the Hardy-Weinberg test for disequilibrium that works well with limited coverage per sample. See the 1000 Genomes Phase I release for more information. Note that the Inbreeding Coefficient will not be calculated for files with fewer than a minimum (generally 10) number of samples.

**Databases with paramteres could be used with this command:**

* -resource:hapmap,VCF,known=false,training=true,truth=true,prior=15.0 hapmap\_3.3. hg19.vcf
* -resource:omni,VCF,known=false,training=true,truth=false,prior=12.0 1000G\_omni2.5. hg19.vcf
* -resource:dbsnp,VCF,known=true,training=false,truth=false,prior=6.0 dbsnp\_137. hg19.vcf

**Note: If model building fails to due to low data; then user should get the raw snp output in step 5.1 with the message that not able to run VQSR due to low depth**

**Step 5.3: Apply SNP recalibration**

**Usage:**

java -Xmx3g -jar GenomeAnalysisTK.jar \

-T ApplyRecalibration \

-R reference/hg19.fasta \

--num\_threads 8 \

-input <output\_SNP.vcf> \

--ts\_filter\_level 99.0 \

-tranchesFile pathTo/output.tranches \

-recalFile pathTo/output.recal \

-mode SNP \

-o SNP.recalibrated.filtered.vcf

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP/indels calls in variant call format (output\_SNP.vcf here) [required]
* Recal file generated by VariantRecalibrator (output.recal) [required]
* tranches file generated by VariaintRecalibrator (output.tranches) [required]

**Output:**

* Output file with analysis ready SNPs; **provide name** (SNP.recalibrated.filtered.vcf here)
* Index vcf file (SNP.recalibrated.filtered.vcf.idx here)

**Parameter description:**

* --ts\_filter\_level: The truth sensitivity level at which to start filtering
* -mode: Recalibration mode to employ: 1.) SNP for recalibrating only SNPs (emitting indels untouched in the output VCF); 2.) INDEL for indels

**Step 5.4: Select Indel**

**Usage:**

java -Xmx2g -jar GenomeAnalysisTK.jar \

-T SelectVariants \

-R reference/hg19.fasta \

--variant <variants.raw.vcf> -o output\_INDEL.vcf -selectType INDEL

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP/indels calls in variant call format (variants.raw.vcf here) [required]

**Output:**

* VCF file with INDEL only; **provide name** (output\_INDEL.vcf here)
* Index file (output\_INDEL.vcf.idx here)

**Parameter description:**

* -selectType: Selection the INDELs only

**Step 5.5: Recalibrate Indel**

**Usage:**

java -Xmx4g -jar GenomeAnalysisTK.jar \

-T VariantRecalibrator \

-R reference/hg19.fasta \

--num\_threads 8 \

-input <output\_INDEL.vcf> \

-resource:mills,known=true,training=true,truth=true,prior=12.0 pathTo/Mills\_and\_1000G\_gold\_standard.indels.hg19.vcf \

-an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an FS -an MQ -

-mode INDEL \

--maxGaussians 4 \

-percentBad 0.05 \

-recalFile pathTo/output\_indel.recal \

-tranchesFile pathTo/output\_indel.tranches \

-rscriptFile pathTo/output\_indel.plots.R

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP/indels calls in variant call format (output\_INDEL.vcf here) [required]

**Output:**

* Recal file to be used by ApplyRecalibration; **provide name** (output\_indel.recal)
* Tranches file to be used by ApplyRecalibration; **provide name** (output\_indel.tranches)
* The output rscript file generated by the VQSR to aid in visualization of the input data and learned model ; ***provide name*** (output\_indel.plots.R here)
* Plots of recalibration; ***provide name*** (Output\_indel.plots.R.pdf and output\_indel.tranches.pdf here)

**Parameter description:**

* No new parameter except analysis mode changed to INDEL

**Database with parameters could be used with this command:**

-resource:mills,VCF,known=false,training=true,truth=true,prior=12.0 [Mills\_and\_1000G\_gold\_standard.indels.hg19.]

**Note: If model building fails to due to low data; then user should get the raw indel output in step 5.4 with the message that not able to run VQSR due to low depth**

**Step 5.6: Apply Indel recalibration**

Note: run after SNP recalibration finished

**Usage:**

java -Xmx3g -jar GenomeAnalysisTK.jar \

-T ApplyRecalibration \

-R reference/hg19.fasta \

--num\_threads 8 \

-input <output\_INDEL.vcf> \

--ts\_filter\_level 95.0 \

-tranchesFile pathTo/output\_indel.tranches \

-recalFile pathTo/output\_indel.recal \

-mode INDEL \

-o INDEL.recalibrated.filtered.vcf

**Input:**

* Reference genome file (hg19.fasta here) [required]
* SNP calls in variant call format (output\_INDEL.vcf here) [required]
* Recal file generated by VariantRecalibrator (output\_indel.recal) [required]
* tranches file generated by VariaintRecalibrator (output\_indel.tranches) [required]

**Output:**

* Output file with analysis ready Indels; **provide name** (INDEL.recalibrated.filtered.vcf here)
* Output variant index file (INDEL.recalibrated.filtered.vcf.idx here)

**Parameter description:**

* No new parameter here except mode is INDEL

**Broad Recommended parallelism configurations (Guide Article #1975 2013-01-14 18:02:57)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tool** | **Full name** | **Type of traversal** | **NT** | **NCT** | **SG** |
| RTC | RealignerTargetCreator | RodWalker | + | - | - |
| IR | IndelRealigner | ReadWalker | - | - | + |
| BR | BaseRecalibrator | LocusWalker | - | + | + |
| PR | PrintReads | ReadWalker | - | + | - |
| RR | ReduceReads | ReadWalker | - | - | + |
| UG | UnifiedGenotyper | LocusWalker | + | + | + |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tool** | **RTC** | **IR** | **BR** | **PR** | **RR** | **UG** |
| Available modes | NT | SG | NCT,SG | NCT | SG | NT,NCT,SG |
| Cluster nodes | 1 | 4 | 4 | 1 | 4 | 4 / 4 / 4 |
| CPU threads (-nct) | 1 | 1 | 8 | 4-8 | 1 | 3 / 6 / 24 |
| Data threads (-nt) | 24 | 1 | 1 | 1 | 1 | 8 / 4 / 1 |
| Memory (Gb) | 48 | 4 | 4 | 4 | 4 | 32 / 16 / 4 |

**Broad recommended sets of known sites per tool** (http://gatkforums.broadinstitute.org/discussion/1247/what-should-i-use-as-known-variantssites-for-running-tool-x)

**Summary table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tool | dbSNP 129 | dbSNP >132 | Mills indels | 1KG indels | HapMap | Omni |
| RealignerTargetCreator |  |  | X | X |  |  |
| IndelRealigner |  |  | X | X |  |  |
| BaseRecalibrator |  | X | X | X |  |  |
| (UnifiedGenotyper/ HaplotypeCaller) |  | X |  |  |  |  |
| VariantRecalibrator |  | X | X |  | X | X |
| VariantEval | X |  |  |  |  |  |

**RealignerTargetCreator and IndelRealigner**

These tools require known indels passed with the -known argument to function properly. We use both the following files:

* Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf
* 1000G\_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)

**BaseRecalibrator**

This tool requires known SNPs and indels passed with the -knownSites argument to function properly. We use all the following files:

* The most recent dbSNP release (build ID > 132)
* Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf
* 1000G\_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)

**UnifiedGenotyper / HaplotypeCaller**

These tools do NOT require known sites, but if SNPs are provided with the -dbsnp argument they will use them for variant annotation. We use this file:

* The most recent dbSNP release (build ID > 132)

**VariantRecalibrator**

This tool requires known SNPs and indels passed with the -resource argument to function properly. We use all the following files:

* HapMap genotypes and sites
* OMNI 2.5 genotypes and sites for 1000 Genomes samples
* The most recent dbSNP release (build ID > 132)
* Mills\_and\_1000G\_gold\_standard.indels.b37.sites.vcf

For best results, these resources should be passed with these parameters:

-resource:hapmap,VCF,known=false,training=true,truth=true,prior=15.0 hapmap\_3.3.b37.sites.vcf \

-resource:omni,VCF,known=false,training=true,truth=false,prior=12.0 1000G\_omni2.5.b37.sites.vcf \

-resource:dbsnp,VCF,known=true,training=false,truth=false,prior=6.0 dbsnp\_135.b37.vcf \

-resource:mills,VCF,known=false,training=true,truth=true,prior=12.0 gold.standard.indel.b37.vcf

**Appendix**

**Files required**

Human reference genome

Illumina adapter sequences used

Desired ‘known’ files from Broad GATK bundle

Interval files for exome capture bait and target (create in house from vendor’s files)

Bedfile for Unified genotyper (bait/target from vendor)

**Accessory applications required**

R packages:

library("ggplot2")

library(gplots)

library("reshape")

library("grid")

library("tools") #For compactPDF in R 2.13+

library(gsalib)

(check R version compatibilities)