variant calling Pipeline on the Test data

We are running this pipeline for the H3Abionet accrediation. I take note of the following:

- The previous iteration was using an older version of the tools
- I am now running a current version of the tools, including the known sites data
- The plan is to convert this pipeline into a snakemake, but that depends on how long it is going to take us.
- Nevertheless, this pipeline should be well documented and made available via GitHub (Icould just change the naming conventions, or request for permision to do so after the accreditation process is complete
- I need to make this easy and reproducible but providing all installation instructions

Required installations

We are setting up a conda envoronment that can be se dto easily reproduce what we did in this analysis. We could decide to use tools like snakemake pipelining language.

- Install `conda install -c bioconda bwa` from Anaconda distribution.
- Install `conda install -c bioconda picard` from Bioconda for de-duplication. I need to actually test different tools and be able to defend the one I finaly choose.

Human Genome Version

In our analysis, alignment and the rest, we have been using the human genome, which see,s to be version b37 or h919. We now need to confirm the version that was used...Or rather, we can choose a specific version and stick with it.

Hg38:http://hgdownload.cse.ucsc.edu/goldenpath/hg38/chromosomes/ (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/chromosomes/)

Hg19:http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/ (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/)

Verdict: We are using b37 assembly of the reference genome

Details about the test data provided

We have limitted information about the data provided for the analysis. I am assuming we are dealing with exome sequences. If so, we may need to modify some of the paramters to fit the data we are working on.

H3ABionet Next Gen Accreditation Questions

The following are questions to keep in mind when running the NextGen Workflow during the H3ABioNet accreditation exercise. Use them to plan your work in a way that would allow gathering the necessary information for your final report. The report should not be limited to only providing brief answers to these questions; it is expected to be a well-rounded description of the process of running the workflow, and of the results. Please note that only phases I and II of the variant calling SOP (http://h3abionet.org/node-assessments/human-variant-calling (http://h3abionet.org/node-assessments/human-variant-calling)) need to be performed.

Questions related to the nature of the input dataset

- Was the input dataset of sufficiently good quality to perform the analysis?
  - *Based on the FastQC, yes.*
- How did the reads' quality and GC content affect the way analysis was run?
  - *Since the sequences were of good quality, we proceeded with the standard variant calling pipeline.*

Operational questions

- At each step of the workflow, describe which software was used and why:
  - Was the choice affected by the nature and/or quality of the reads?
  - Was the choice made due to the time and cost of the analysis?
  - What are the accuracy and performance considerations for the chosen piece of software?
    - *If we are running two software, we will have to find a way of comparing the variant calls.*

For each software, describe which input parameters were chosen, and why:

- Was the choice affected by the nature and/or quality of the reads?
- Did the available hardware play a role in the parameter choice?
  - *Yes, especially where we used multithreading to make the analysis faster.*
- How did the purpose of the study affect the parameter choice?
  - *In this case, the purpose of the study is just accreditation. There is little information provided on what we need to be doing. Other that that, we understand that we only need to perform the analysis until variant calls*

For each step of the workflow, how do you know that it completed successfully and that the results are usable for the next step?

- *This was determined by looking at the stdout messages, to confirm if the data ran successfully. We will provide specifc details in the corresponding sections.*
- *The various steps completed without any error messages*
- *Expected files, usable in subsequent steps produced*

Runtime analysis: this is useful information for making predictions for the clients and collaborators

- How much time and disk space did each step of the workflow take?

```
1004M   ./artefact_removal
39M ./Variants/gatk
25M ./Variants/freebayes
63M ./Variants
446M    ./dedup
14M ./VQSR
1.4M    ./QC
4.8G    ./BQSR
2.5G    ./bwa
8.7G    ./
1.2G    ./Data/Derived
18G     ./Data/VcfDatabase/b37
```

- How did the underlying hardware perform? Was it possible to do other things, or run other analyses on the same computer at the same time?
  - *We ran the analysis on the icipe hpc with the following specs:...In addition to variant calling analysis, the system was being used for metagenomic analysis, clustalo alignment, etc. So we can confidently say that the hardware performed well.*

Analyzing the results

- How many variants were called with sufficient confidence to be included in further analyses? Are the results good and trustworthy, and can you estimate the sensitivity and selectivity of the analysis? How do you know the workflow completed successfully and the results are worth analyzing further?
- How many variants were located in intronic, exonic, or in non-genic regions? Put this in context of the nature of the input dataset as described in the README.
- How many variants were found in dbSNP and how many were unique to your sample? What does it mean?

```
  - *Variant Eval tool output will be useful for this step of the
 analysis.*
  - *We need to interpret the variants called with the known and u
nkown and confirm whther the novel variants are true of not*
  -
```

*

- What is the fraction of simple variants (SNPs, small indels) versus complex variants (translocations, inversions, etc.)? How is this influenced by your choice of software and parameters?
  - *If possible, find of comparing the freebayes and gatk output for direct comparison.*
- What would be the next steps for your analysis, given this information?
  - We are confident that the quality of the called variant is at acceptable quality for

In [1]:

```
# Quality check using fastqc

!time fastqc -f fastq -o ../Results/QC/ -t 35 ../Data/Derived/*.fq
```

```
Started analysis of set5_read1.fq
Started analysis of set5_read2.fq
Approx 5% complete for set5_read1.fq
Approx 5% complete for set5_read2.fq
Approx 10% complete for set5_read1.fq
Approx 10% complete for set5_read2.fq
Approx 15% complete for set5_read1.fq
Approx 20% complete for set5_read1.fq
Approx 15% complete for set5_read2.fq
Approx 25% complete for set5_read1.fq
Approx 20% complete for set5_read2.fq
Approx 30% complete for set5_read1.fq
Approx 25% complete for set5_read2.fq
Approx 35% complete for set5_read1.fq
Approx 30% complete for set5_read2.fq
Approx 40% complete for set5_read1.fq
Approx 35% complete for set5_read2.fq
Approx 45% complete for set5_read1.fq
Approx 40% complete for set5_read2.fq
Approx 50% complete for set5_read1.fq
Approx 45% complete for set5_read2.fq
Approx 55% complete for set5_read1.fq
Approx 50% complete for set5_read2.fq
Approx 60% complete for set5_read1.fq
Approx 55% complete for set5_read2.fq
Approx 65% complete for set5_read1.fq
Approx 60% complete for set5_read2.fq
Approx 70% complete for set5_read1.fq
Approx 65% complete for set5_read2.fq
Approx 75% complete for set5_read1.fq
Approx 70% complete for set5_read2.fq
Approx 80% complete for set5_read1.fq
Approx 75% complete for set5_read2.fq
Approx 85% complete for set5_read1.fq
Approx 80% complete for set5_read2.fq
Approx 90% complete for set5_read1.fq
Approx 85% complete for set5_read2.fq
Approx 95% complete for set5_read1.fq
Approx 90% complete for set5_read2.fq
Analysis complete for set5_read1.fq
Approx 95% complete for set5_read2.fq
Analysis complete for set5_read2.fq

real    0m29.317s
user    0m49.460s
sys     0m8.454s
```

Quality check

We used FASTQC for qality analysis, and we found that all the reads passed the quality filtering. Therefore, we did not need to perform any read or addpater trimming.

The only question is, whey did we need to perform deduplication analysis?

In [1]:

```
```

Out[1]:

```
0.0030360934182590235
```

In [33]:

```
!time bwa index ../Data/Genome/chr1.fa
```

```
[bwa_index] Pack FASTA... 3.41 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=497912844, availableWord=47034604
[BWTIncConstructFromPacked] 10 iterations done. 76325452 characters
processed.
[BWTIncConstructFromPacked] 20 iterations done. 142216380 characters
processed.
[BWTIncConstructFromPacked] 30 iterations done. 200776364 characters
processed.
[BWTIncConstructFromPacked] 40 iterations done. 252820604 characters
processed.
[BWTIncConstructFromPacked] 50 iterations done. 299073676 characters
processed.
[BWTIncConstructFromPacked] 60 iterations done. 340179500 characters
processed.
[BWTIncConstructFromPacked] 70 iterations done. 376710428 characters
processed.
[BWTIncConstructFromPacked] 80 iterations done. 409175164 characters
processed.
[BWTIncConstructFromPacked] 90 iterations done. 438025868 characters
processed.
[BWTIncConstructFromPacked] 100 iterations done. 463664428 character
s processed.
[BWTIncConstructFromPacked] 110 iterations done. 486447996 character
s processed.
[bwt_gen] Finished constructing BWT in 116 iterations.
[bwa_index] 342.69 seconds elapse.
[bwa_index] Update BWT... 2.18 sec
[bwa_index] Pack forward-only FASTA... 2.20 sec
[bwa_index] Construct SA from BWT and Occ... 61.80 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index ../Data/Genome/chr/chr1.fa
[main] Real time: 416.256 sec; CPU: 412.291 sec

real    6m56.261s
user    6m50.225s
sys     0m2.069s
```

Alignment

In [43]:

```
!time bwa mem -t 48 ../Data/Genome/chr1.fa ../Data/Derived/set5_read1.fq \
../Data/Derived/set5_read2.fq > ../Results/bwa/set5_out.sam
```

```
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 4752476 sequences (480000076 bp)...
[M::process] read 153134 sequences (15466534 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (0, 2
172894, 0, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation F
R...
[M::mem_pestat] (25, 50, 75) percentile: (279, 299, 319)
[M::mem_pestat] low and high boundaries for computing mean and std.d
ev: (199, 399)
[M::mem_pestat] mean and std.dev: (298.96, 29.86)
[M::mem_pestat] low and high boundaries for proper pairs: (159, 439)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] skip orientation RR as there are not enough pairs
[M::mem_process_seqs] Processed 4752476 reads in 1273.559 CPU sec, 5
5.171 real sec
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (0, 6
7837, 0, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation F
R...
[M::mem_pestat] (25, 50, 75) percentile: (279, 299, 319)
[M::mem_pestat] low and high boundaries for computing mean and std.d
ev: (199, 399)
[M::mem_pestat] mean and std.dev: (298.79, 29.82)
[M::mem_pestat] low and high boundaries for proper pairs: (159, 439)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] skip orientation RR as there are not enough pairs
[M::mem_process_seqs] Processed 153134 reads in 40.134 CPU sec, 0.89
2 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem -t 48 ../Data/Genome/chr1.fa ../Data/Derived/set
5_read1.fq ../Data/Derived/set5_read2.fq
[main] Real time: 85.784 sec; CPU: 1329.110 sec

real    1m26.297s
user    21m33.116s
sys     0m36.006s
```

Converting sam to bam

In the next steps, we conver the sam files to

In [4]:

```
!time samtools view -S -b ../Results/bwa/set5_out.sam  \
>  ../Results/bwa/set5_out.bam
```

```
real    2m7.150s
user    2m2.407s
sys     0m2.698s
```

Getting statistics from the data using samtools

At this point, we need to get some statistics from the data. The he samtools flagstat tool provides counts for each of the read categories based primarily on bit flags in the FLAG field. The results:

```
4905632 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
22 + 0 supplementary
0 + 0 duplicates
4905630 + 0 mapped (100.00% : N/A)
4905610 + 0 paired in sequencing
2452805 + 0 read1
2452805 + 0 read2
4905544 + 0 properly paired (100.00% : N/A)
4905608 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Show that all our reads passed quality filtering and mapping. This confirms our conclusion that the reads were of sufficient quality for use in downstream analysis.

In [5]:

```
!time samtools flagstat ../Results/bwa/set5_out.bam \
>../Results/bwa/set5_out.bam.flagstat
```

```
real    0m10.471s
user    0m10.256s
sys     0m0.212s
```

De-duplication

Now we need to perfom some de-duplication

But first, we need to sort the bam file and get some statistics. Why is this step required?

In [6]:

```
!time samtools sort ../Results/bwa/set5_out.bam \
> ../Results/bwa/set5_out_sorted.bam
```

```
[bam_sort_core] merging from 1 files and 1 in-memory blocks...

real    2m32.094s
user    2m25.403s
sys     0m8.018s
```

In [7]:

```
!time samtools index ../Results/bwa/set5_out_sorted.bam
```

```
real    0m10.272s
user    0m9.964s
sys     0m0.165s
```

In [11]:

```
!time picard MarkDuplicates \
I=../Results/bwa/set5_out_sorted.bam \
O=../Results/dedup/set5_out_sorted.dedup.bam \
M=../Results/dedup/set5_out_sorted.metrics
```

```
10:41:33.243 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/picard-
2.18.11-0/picard.jar!/com/intel/gkl/native/libgkl_compression.so
[Thu Sep 13 10:41:33 EAT 2018] MarkDuplicates INPUT=[../Results/bwa/
set5_out_sorted.bam] OUTPUT=../Results/dedup/set5_out_sorted.dedup.b
am METRICS_FILE=../Results/dedup/set5_out_sorted.metrics    MAX_SEQU
ENCES_FOR_DISK_READ_ENDS_MAP=50000 MAX_FILE_HANDLES_FOR_READ_ENDS_MA
P=8000 SORTING_COLLECTION_SIZE_RATIO=0.25 TAG_DUPLICATE_SET_MEMBERS=
false REMOVE_SEQUENCING_DUPLICATES=false TAGGING_POLICY=DontTag CLEA
R_DT=true ADD_PG_TAG_TO_READS=true REMOVE_DUPLICATES=false ASSUME_SO
RTED=false DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUALITIES PROGRAM_
RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=MarkDuplicates READ_NAME
_REGEX=<optimized capture of last three ':' separated fields as nume
ric values> OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 MAX_OPTICAL_DUPLICA
TE_SET_SIZE=300000 VERBOSITY=INFO QUIET=false VALIDATION_STRINGENCY=
STRICT COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=fa
lse CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json U
SE_JDK_DEFLATER=false USE_JDK_INFLATER=false
[Thu Sep 13 10:41:33 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is not a
vailable; Picard version: 2.18.11-SNAPSHOT
INFO    2018-09-13 10:41:33     MarkDuplicates  Start of doWork free
Memory: 502693528; totalMemory: 514850816; maxMemory: 954728448
INFO    2018-09-13 10:41:33     MarkDuplicates  Reading input file a
nd constructing read end information.
INFO    2018-09-13 10:41:33     MarkDuplicates  Will retain up to 34
59161 data points before spilling to disk.
WARNING 2018-09-13 10:41:33     AbstractOpticalDuplicateFinderComman
dLineProgram    A field field parsed out of a read name was expected
to contain an integer and did not. Read name: H3A_WES_chr1_50X_E0.00
5-j1-chr1-r531. Cause: String 'H3A_WES_chr1_50X_E0.005-j1-chr1-r531'
did not start with a parsable number.
INFO    2018-09-13 10:41:45     MarkDuplicates  Read     1,000,000 r
ecords.  Elapsed time: 00:00:12s.  Time for last 1,000,000:   12s.
Last read position: chr1:38,227,850
INFO    2018-09-13 10:41:45     MarkDuplicates  Tracking 5 as yet un
matched pairs. 5 records in RAM.
INFO    2018-09-13 10:41:56     MarkDuplicates  Read     2,000,000 r
ecords.  Elapsed time: 00:00:23s.  Time for last 1,000,000:   11s.
Last read position: chr1:91,169,474
INFO    2018-09-13 10:41:56     MarkDuplicates  Tracking 2 as yet un
matched pairs. 2 records in RAM.
INFO    2018-09-13 10:42:09     MarkDuplicates  Read     3,000,000 r
ecords.  Elapsed time: 00:00:36s.  Time for last 1,000,000:   12s.
Last read position: chr1:155,746,851
INFO    2018-09-13 10:42:09     MarkDuplicates  Tracking 3 as yet un
matched pairs. 3 records in RAM.
INFO    2018-09-13 10:42:24     MarkDuplicates  Read     4,000,000 r
ecords.  Elapsed time: 00:00:50s.  Time for last 1,000,000:   14s.
Last read position: chr1:203,372,865
INFO    2018-09-13 10:42:24     MarkDuplicates  Tracking 2 as yet un
matched pairs. 2 records in RAM.
INFO    2018-09-13 10:42:34     MarkDuplicates  Read 4905630 record
s. 0 pairs never matched.
INFO    2018-09-13 10:42:38     MarkDuplicates  After buildSortedRea
dEndLists freeMemory: 700326272; totalMemory: 902823936; maxMemory:
954728448
INFO    2018-09-13 10:42:38     MarkDuplicates  Will retain up to 29
835264 duplicate indices before spilling to disk.
INFO    2018-09-13 10:42:38     MarkDuplicates  Traversing read pair
```

```
information and detecting duplicates.
INFO    2018-09-13 10:42:39     MarkDuplicates  Traversing fragment
information and detecting duplicates.
INFO    2018-09-13 10:42:39     SortingCollection       Creating mer
ging iterator from 2 files
INFO    2018-09-13 10:42:41     MarkDuplicates  Sorting list of dupl
icate records.
INFO    2018-09-13 10:42:41     MarkDuplicates  After generateDuplic
ateIndexes freeMemory: 804350176; totalMemory: 1054867456; maxMemor
y: 1054867456
INFO    2018-09-13 10:42:41     MarkDuplicates  Marking 2194 records
as duplicates.
INFO    2018-09-13 10:42:41     MarkDuplicates  Found 0 optical dupl
icate clusters.
INFO    2018-09-13 10:42:41     MarkDuplicates  Reads are assumed to
be ordered by: coordinate
INFO    2018-09-13 10:43:55     MarkDuplicates  Before output close
freeMemory: 1047306544; totalMemory: 1060110336; maxMemory: 10601103
36
INFO    2018-09-13 10:43:55     MarkDuplicates  After output close f
reeMemory: 1051500848; totalMemory: 1064304640; maxMemory: 106430464
0
[Thu Sep 13 10:43:55 EAT 2018] picard.sam.markduplicates.MarkDuplica
tes done. Elapsed time: 2.37 minutes.
Runtime.totalMemory()=1064304640

real    2m22.926s
user    17m42.081s
sys     0m19.789s
```

This next command will count the number of sequences that have been marked as duplicates by picard. The question is, what ratio does this represent on the whole? How are we to use these statistics?

In [17]:

```
!samtools view -c -f 1024 ../Results/dedup/set5_out_sorted.dedup.bam
```

```
2194
```

From the deduplication step, 2194 pairs were marked as duplicates.

```
## METRICS CLASS    picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED SECONDARY_OR_SUPPLEME
NTARY_RDS  UNMAPPED_READS  UNPAIRED_READ_DUPLICATES    READ_PAIR_DUPLICAT
ES    READ_PAIR_OPTICAL_DUPLICATES    PERCENT_DUPLICATION ESTIMATED_LIBRA
RY_SIZE
Unknown Library 0   2452804 22  2   0   1097    0   0.000447    274131882
9
```

There were no optical duplicates identified in the dateset.

Artefact removal using GATK IndelRealigner.

The rationale for this stage is? This step may not be required as outlined in this article (https://software.broadinstitute.org/gatk/blog?id=7847). This mean this step may not really be required. The real question is, why did we have to do this, and is what we have already done going to affect further anlysis?

module load gatk/3.3

Create the target list using GATK RealignerTargetCreator as below:

GATK RealignerTargetCreator requires

1. fasta index file for the reference
2. Dictionar file for the reference
3. Readgroups to be defined for the bam files

1) Create fasta index using samtools

In [35]:

```
!time samtools faidx ../Data/Genome/chr1.fa
```

```
real    0m3.015s
user    0m2.870s
sys     0m0.068s
```

2) Create dictionary using Picard-tools

In [41]:

```
!time picard CreateSequenceDictionary \
R=../Data/Genome/chr1.fa \
O=../Data/Genome/chr1.dict
```

```
14:26:53.208 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/picard-
2.18.11-0/picard.jar!/com/intel/gkl/native/libgkl_compression.so
[Thu Sep 20 14:26:53 EAT 2018] CreateSequenceDictionary OUTPUT=../Da
ta/Genome/chr1.dict REFERENCE=../Data/Genome/chr1.fa    TRUNCATE_NAM
ES_AT_WHITESPACE=true NUM_SEQUENCES=2147483647 VERBOSITY=INFO QUIET=
false VALIDATION_STRINGENCY=STRICT COMPRESSION_LEVEL=5 MAX_RECORDS_I
N_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_S
ECRETS=client_secrets.json USE_JDK_DEFLATER=false USE_JDK_INFLATER=f
alse
[Thu Sep 20 14:26:53 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is not a
vailable; Picard version: 2.18.11-SNAPSHOT
[Thu Sep 20 14:26:54 EAT 2018] picard.sam.CreateSequenceDictionary d
one. Elapsed time: 0.03 minutes.
Runtime.totalMemory()=514850816

real    0m2.435s
user    0m3.398s
sys     0m0.473s
```

3) Add readgroup information to bam using Picard-tools AddOrReplaceReadGroups

As readgroup details are not available arbitrary ones were used

Adding readgroup information to bam using Picard-tools AddOrReplaceReadGroups

We may need to find more information here so that we can be able to set this up corectly.

In [44]:

```
!time picard AddOrReplaceReadGroups \
I=../Results/dedup/set5_out_sorted.dedup.bam \
O=../Results/artefact_removal/set5_out_sorted_aln_dedup.bam \
LB=00001 PL=illumina PU=001 SM=out
```

```
14:52:19.292 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/picard-
2.18.11-0/picard.jar!/com/intel/gkl/native/libgkl_compression.so
[Thu Sep 20 14:52:19 EAT 2018] AddOrReplaceReadGroups INPUT=../Resul
ts/dedup/set5_out_sorted.dedup.bam OUTPUT=../Results/artefact_remova
l/set5_out_sorted_aln_dedup.bam RGLB=00001 RGPL=illumina RGPU=001 RG
SM=out    RGID=1 VERBOSITY=INFO QUIET=false VALIDATION_STRINGENCY=ST
RICT COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=fals
e CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json USE
_JDK_DEFLATER=false USE_JDK_INFLATER=false
[Thu Sep 20 14:52:19 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is not a
vailable; Picard version: 2.18.11-SNAPSHOT
INFO    2018-09-20 14:52:19     AddOrReplaceReadGroups  Created read
-group ID=1 PL=illumina LB=00001 SM=out

INFO    2018-09-20 14:52:34     AddOrReplaceReadGroups  Processed
1,000,000 records.  Elapsed time: 00:00:14s.  Time for last 1,000,00
0:   14s.  Last read position: chr1:38,227,850
INFO    2018-09-20 14:52:48     AddOrReplaceReadGroups  Processed
2,000,000 records.  Elapsed time: 00:00:29s.  Time for last 1,000,00
0:   14s.  Last read position: chr1:91,169,474
INFO    2018-09-20 14:53:03     AddOrReplaceReadGroups  Processed
3,000,000 records.  Elapsed time: 00:00:43s.  Time for last 1,000,00
0:   14s.  Last read position: chr1:155,746,851
INFO    2018-09-20 14:53:17     AddOrReplaceReadGroups  Processed
4,000,000 records.  Elapsed time: 00:00:58s.  Time for last 1,000,00
0:   14s.  Last read position: chr1:203,372,865
[Thu Sep 20 14:53:31 EAT 2018] picard.sam.AddOrReplaceReadGroups don
e. Elapsed time: 1.20 minutes.
Runtime.totalMemory()=566231040

real    1m12.780s
user    1m28.597s
sys     0m2.027s
```

In [45]:

```
!time samtools sort ../Results/artefact_removal/set5_out_sorted_aln_dedup.bam \
-o ../Results/artefact_removal/set5_out_sorted_aln_dedup.bam
```

```
[bam_sort_core] merging from 1 files and 1 in-memory blocks...

real    2m33.156s
user    2m25.408s
sys     0m8.367s
```

In [46]:

```
!time samtools index ../Results/artefact_removal/set5_out_sorted_aln_dedup.bam
```

```
real    0m10.498s
user    0m10.157s
sys     0m0.203s
```

Base Score recalibration

module load gatk/3.3

Building the model using known variants

To install gatk, we need to follow the following steps:

1. Use install the Bioconda version of the tool

```
conda install -c conda-forge -c bioconda gatk
```

1. Download the licensed versions using:

`https://github.com/broadinstitute/gatk/releases/download/4.0.8.1/gatk-4.0.8.1.zip` wget "https://software.broadinstitute.org/gatk/download/auth?package=GATK (https://software.broadinstitute.org/gatk/download/auth?package=GATK)" -O GenomeAnalysisTK-3.8-0.tar.bz2

1. Activate the license by running:

```
gatk3-activate GenomeAnalysisTK-3.8-0.tar.bz2
```

To sort out the initial error, I used the guide for this page (https://github.com/bioconda/bioconda-recipes/issues/6038)

GATK4

We may actually beinterested in the latest version of GATK4

https://anaconda.org/bioconda/gatk4 (https://anaconda.org/bioconda/gatk4)

First we had to index the feature file

In [13]:

```
!time gatk IndexFeatureFile -F ../Data/VcfDatabase/common_all_20180418.vcf
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar IndexFeature
File -F ../Data/VcfDatabase/common_all_20180418.vcf
15:41:47.253 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
15:41:52.572 INFO  IndexFeatureFile - ------------------------------
------------------------------
15:41:52.572 INFO  IndexFeatureFile - The Genome Analysis Toolkit (G
ATK) v4.0.8.1
15:41:52.572 INFO  IndexFeatureFile - For support and documentation
go to https://software.broadinstitute.org/gatk/
15:41:52.574 INFO  IndexFeatureFile - Executing as caleb@hpc01.icip
e.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
15:41:52.574 INFO  IndexFeatureFile - Java runtime: OpenJDK 64-Bit S
erver VM v1.8.0_121-b15
15:41:52.574 INFO  IndexFeatureFile - Start Date/Time: September 17,
2018 3:41:47 PM EAT
15:41:52.575 INFO  IndexFeatureFile - ------------------------------
------------------------------
15:41:52.575 INFO  IndexFeatureFile - ------------------------------
------------------------------
15:41:52.575 INFO  IndexFeatureFile - HTSJDK Version: 2.16.0
15:41:52.575 INFO  IndexFeatureFile - Picard Version: 2.18.7
15:41:52.576 INFO  IndexFeatureFile - HTSJDK Defaults.COMPRESSION_LE
VEL : 2
15:41:52.576 INFO  IndexFeatureFile - HTSJDK Defaults.USE_ASYNC_IO_R
EAD_FOR_SAMTOOLS : false
15:41:52.576 INFO  IndexFeatureFile - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_SAMTOOLS : true
15:41:52.576 INFO  IndexFeatureFile - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_TRIBBLE : false
15:41:52.576 INFO  IndexFeatureFile - Deflater: IntelDeflater
15:41:52.576 INFO  IndexFeatureFile - Inflater: IntelInflater
15:41:52.576 INFO  IndexFeatureFile - GCS max retries/reopens: 20
15:41:52.576 INFO  IndexFeatureFile - Using google-cloud-java fork h
ttps://github.com/broadinstitute/google-cloud-java/releases/tag/0.2
0.5-alpha-GCS-RETRY-FIX
15:41:52.577 INFO  IndexFeatureFile - Initializing engine
15:41:52.577 INFO  IndexFeatureFile - Done initializing engine
15:41:53.151 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Data/VcfDatabase/common_all_20180418.vcf
15:41:53.169 INFO  ProgressMeter - Starting traversal
15:41:53.169 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes     Records Processed   Records/Minute
15:42:03.187 INFO  ProgressMeter -        chr1:77774779
0.2                984000        5893980.2
15:42:13.245 INFO  ProgressMeter -       chr1:154975851
0.3               1605000        4796772.3
15:42:24.718 INFO  ProgressMeter -       chr1:159232219
0.5               1658000        3153190.3
15:42:35.370 INFO  ProgressMeter -       chr1:215801025
0.7               2374000        3375355.5
15:42:45.370 INFO  ProgressMeter -        chr2:31327093
```

```
0.9                3250000        3735560.6
15:42:55.466 INFO  ProgressMeter -         chr2:52235746
1.0                3564000        3432589.0
15:43:05.542 INFO  ProgressMeter -         chr2:61148570
1.2                3689000        3058365.1
15:43:15.544 INFO  ProgressMeter -        chr2:162482305
1.4                4881000        3555248.0
15:43:25.677 INFO  ProgressMeter -        chr2:185729946
1.5                5168000        3351926.3
15:43:36.039 INFO  ProgressMeter -        chr2:188679580
1.7                5202000        3034179.7
15:43:46.174 INFO  ProgressMeter -        chr2:190759577
1.9                5226000        2774769.0
15:43:56.499 INFO  ProgressMeter -        chr2:191927589
2.1                5240000        2549278.8
15:44:06.927 INFO  ProgressMeter -        chr2:193974236
2.2                5263000        2360830.8
15:44:17.592 INFO  ProgressMeter -        chr2:195020745
2.4                5276000        2191894.6
15:44:28.536 INFO  ProgressMeter -        chr2:195774917
2.6                5286000        2041360.1
15:44:38.572 INFO  ProgressMeter -        chr2:200675819
2.8                5340000        1937098.7
15:44:49.447 INFO  ProgressMeter -        chr2:201595418
2.9                5350000        1820997.6
15:44:59.582 INFO  ProgressMeter -         chr3:19249448
3.1                6164000        1983981.8
15:45:09.584 INFO  ProgressMeter -         chr3:54721319
3.3                6608000        2018583.1
15:45:20.252 INFO  ProgressMeter -        chr3:107253741
3.5                7259000        2103224.8
15:45:30.634 INFO  ProgressMeter -        chr3:108149753
3.6                7269000        2005573.3
15:45:41.976 INFO  ProgressMeter -        chr3:112915543
3.8                7329000        1921881.8
15:45:53.167 INFO  ProgressMeter -        chr3:113577387
4.0                7338000        1834515.3
15:46:03.481 INFO  ProgressMeter -        chr3:114685484
4.2                7350000        1761801.3
15:46:14.069 INFO  ProgressMeter -        chr3:115948981
4.3                7363000        1693298.9
15:46:24.755 INFO  ProgressMeter -        chr3:116663863
4.5                7372000        1628655.4
15:46:35.399 INFO  ProgressMeter -        chr3:119219272
4.7                7406000        1574466.1
15:46:46.087 INFO  ProgressMeter -        chr3:119815829
4.9                7414000        1518655.5
15:46:56.270 INFO  ProgressMeter -        chr3:120517473
5.1                7422000        1469218.1
15:47:06.567 INFO  ProgressMeter -        chr3:121642303
5.2                7436000        1423625.6
15:47:17.328 INFO  ProgressMeter -        chr3:122494639
5.4                7445000        1378027.4
15:47:27.335 INFO  ProgressMeter -          chr4:1121925
5.6                8441000        1515594.0
15:47:37.338 INFO  ProgressMeter -         chr4:85670373
5.7                9611000        1675514.1
15:47:47.861 INFO  ProgressMeter -         chr4:99166996
5.9                9783000        1654900.6
15:47:58.939 INFO  ProgressMeter -         chr4:99759015
6.1                9790000        1605927.2
```

```
15:48:09.756 INFO  ProgressMeter -       chr4:100585747
6.3                9800000        1561392.2
15:48:22.902 INFO  ProgressMeter -       chr4:101216831
6.5                9808000        1509960.7
15:48:34.033 INFO  ProgressMeter -       chr4:101579964
6.7                9813000        1468777.4
15:48:44.783 INFO  ProgressMeter -       chr4:102043886
6.9                9819000        1431295.9
15:48:55.222 INFO  ProgressMeter -       chr4:102609890
7.0                9826000        1396889.5
15:49:05.395 INFO  ProgressMeter -       chr4:103151764
7.2                9832000        1364841.5
15:49:17.012 INFO  ProgressMeter -       chr4:104705082
7.4                9852000        1331822.3
15:49:29.475 INFO  ProgressMeter -       chr4:111354156
7.6                9933000        1306097.2
15:49:40.233 INFO  ProgressMeter -       chr4:112074475
7.8                9941000        1277044.0
15:49:50.509 INFO  ProgressMeter -       chr4:112860772
8.0                9950000        1250680.9
15:50:01.808 INFO  ProgressMeter -       chr4:113441967
8.1                9957000        1222622.9
15:50:13.013 INFO  ProgressMeter -       chr4:113916832
8.3                9963000        1195935.5
15:50:24.819 INFO  ProgressMeter -       chr4:114429295
8.5                9970000        1169160.9
15:50:36.561 INFO  ProgressMeter -       chr4:114817364
8.7                9975000        1143502.4
15:50:49.748 INFO  ProgressMeter -       chr4:115100805
8.9                9979000        1115849.0
15:51:00.917 INFO  ProgressMeter -       chr4:115401149
9.1                9983000        1093534.1
15:51:13.275 INFO  ProgressMeter -       chr4:115964116
9.3                9991000        1070261.7
15:51:24.856 INFO  ProgressMeter -       chr4:116253316
9.5                9995000        1049000.6
15:51:39.026 INFO  ProgressMeter -       chr4:116406018
9.8                9997000        1023833.5
15:51:49.031 INFO  ProgressMeter -        chr5:11387286
9.9               11157000        1123448.0
15:51:59.037 INFO  ProgressMeter -        chr5:61492972
10.1              11775000         1166095.6
15:52:09.038 INFO  ProgressMeter -       chr5:155869970
10.3              12953000         1261926.3
15:52:19.040 INFO  ProgressMeter -        chr6:61524423
10.4              14132000         1354784.0
15:52:29.041 INFO  ProgressMeter -       chr6:153462682
10.6              15315000         1445102.2
15:52:39.044 INFO  ProgressMeter -        chr7:39524492
10.8              16167000         1501869.6
15:52:53.105 INFO  ProgressMeter -       chr7:106123558
11.0              16965000         1542422.3
15:53:03.113 INFO  ProgressMeter -        chr8:22347269
11.2              18099000         1620941.5
15:53:13.114 INFO  ProgressMeter -        chr8:72939438
11.3              18727000         1652518.4
15:53:28.280 INFO  ProgressMeter -         chr9:4730862
11.6              19741000         1703989.3
15:53:38.284 INFO  ProgressMeter -       chr9:115681277
11.8              20913000         1779539.5
15:53:48.287 INFO  ProgressMeter -       chr10:10897836
```

```
11.9              21385000        1794249.3
15:54:02.310 INFO  ProgressMeter -       chr10:98846829
12.2              22517000        1852895.2
15:54:12.315 INFO  ProgressMeter -       chr11:54904919
12.3              23692000        1923192.4
15:54:22.325 INFO  ProgressMeter -      chr11:122507497
12.5              24580000        1968641.6
15:54:32.320 INFO  ProgressMeter -       chr12:43211777
12.7              25320000        2001182.9
15:54:42.321 INFO  ProgressMeter -       chr13:19289585
12.8              26483000        2065888.2
15:54:52.323 INFO  ProgressMeter -       chr13:81656880
13.0              27296000        2101972.1
15:55:02.330 INFO  ProgressMeter -       chr14:72882395
13.2              28449000        2162983.4
15:55:12.337 INFO  ProgressMeter -       chr15:71309924
13.3              29563000        2219533.3
15:55:22.345 INFO  ProgressMeter -       chr16:68322799
13.5              30744000        2279652.4
15:55:32.345 INFO  ProgressMeter -       chr17:54835578
13.7              31764000        2326533.0
15:55:42.880 INFO  ProgressMeter -        chr18:6089446
13.8              32229000        2330618.7
15:55:52.886 INFO  ProgressMeter -       chr19:15584886
14.0              33384000        2385377.9
15:56:02.891 INFO  ProgressMeter -       chr19:57617032
14.2              33955000        2397610.5
15:56:15.774 INFO  ProgressMeter -       chr21:29587142
14.4              35012000        2435323.7
15:56:25.773 INFO  ProgressMeter -        chrX:40736788
14.5              36187000        2488207.7
15:56:35.781 INFO  ProgressMeter -       chrX:127116281
14.7              37009000        2515876.2
15:56:38.177 INFO  ProgressMeter -        chrY:21636725
14.8              37302978        2528992.6
15:56:38.177 INFO  ProgressMeter - Traversal complete. Processed 373
02978 total records in 14.8 minutes.
15:56:38.389 INFO  IndexFeatureFile - Successfully wrote index to /o
pt/data/accreditation/test/NextGenVariantCalling_set5/Code/../Data/V
cfDatabase/common_all_20180418.vcf.idx
15:56:38.389 INFO  IndexFeatureFile - Shutting down engine
[September 17, 2018 3:56:38 PM EAT] org.broadinstitute.hellbender.to
ols.IndexFeatureFile done. Elapsed time: 14.85 minutes.
Runtime.totalMemory()=12123111424
Tool returned:
/opt/data/accreditation/test/NextGenVariantCalling_set5/Code/../Dat
a/VcfDatabase/common_all_20180418.vcf.idx

real    14m59.493s
user    37m8.449s
sys     29m21.795s
```

Base Calibration

Adjusting the quality scores

There is a change from the previous versions, the information is provided here (https://github.com/broadinstitute/gatk/issues/322). The results from this run are in congruence with the error rate in the raw reads of 0.005.

See here as well: https://gatkforums.broadinstitute.org/gatk/discussion/comment/43986#Comment_43986 (https://gatkforums.broadinstitute.org/gatk/discussion/comment/43986#Comment_43986)

BQSR Quality Assessment Pipeline

The base quality recalibration is run in two steps.

1. Generate the first pass recalibration table file

BaseRecalibrator

In [8]:

```
!time gatk BaseRecalibrator -R ../Data/Genome/chr1.fa \
-I ../Results/artefact_removal/set5_out_sorted_aln_dedup.bam \
--known-sites ../Data/VcfDatabase/common_all_20180418.vcf \
-O ../Results/BQSR/set5_pass1.table
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar BaseRecalibr
ator -R ../Data/Genome/chr1.fa -I ../Results/artefact_removal/set5_o
ut_sorted_aln_dedup.bam --known-sites ../Data/VcfDatabase/common_all
_20180418.vcf -O ../Results/BQSR/set5_pass1.table
12:18:35.844 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:18:35.980 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:18:35.981 INFO  BaseRecalibrator - The Genome Analysis Toolkit (G
ATK) v4.0.8.1
12:18:35.981 INFO  BaseRecalibrator - For support and documentation
go to https://software.broadinstitute.org/gatk/
12:18:35.981 INFO  BaseRecalibrator - Executing as caleb@hpc01.icip
e.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:18:35.981 INFO  BaseRecalibrator - Java runtime: OpenJDK 64-Bit S
erver VM v1.8.0_121-b15
12:18:35.982 INFO  BaseRecalibrator - Start Date/Time: September 28,
2018 12:18:35 PM EAT
12:18:35.982 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:18:35.982 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:18:35.983 INFO  BaseRecalibrator - HTSJDK Version: 2.16.0
12:18:35.983 INFO  BaseRecalibrator - Picard Version: 2.18.7
12:18:35.983 INFO  BaseRecalibrator - HTSJDK Defaults.COMPRESSION_LE
VEL : 2
12:18:35.983 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_R
EAD_FOR_SAMTOOLS : false
12:18:35.983 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_SAMTOOLS : true
12:18:35.983 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_TRIBBLE : false
12:18:35.983 INFO  BaseRecalibrator - Deflater: IntelDeflater
12:18:35.983 INFO  BaseRecalibrator - Inflater: IntelInflater
12:18:35.984 INFO  BaseRecalibrator - GCS max retries/reopens: 20
12:18:35.984 INFO  BaseRecalibrator - Using google-cloud-java fork h
ttps://github.com/broadinstitute/google-cloud-java/releases/tag/0.2
0.5-alpha-GCS-RETRY-FIX
12:18:35.984 INFO  BaseRecalibrator - Initializing engine
12:18:36.600 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Data/VcfDatabase/common_all_20180418.vcf
12:18:36.751 WARN  IndexUtils - Feature file "/opt/data/accreditatio
n/test/NextGenVariantCalling_set5/Code/../Data/VcfDatabase/common_al
l_20180418.vcf" appears to contain no sequence dictionary. Attemptin
g to retrieve a sequence dictionary from the associated index file
12:18:36.859 INFO  BaseRecalibrator - Done initializing engine
12:18:36.864 INFO  BaseRecalibrationEngine - The covariates being us
ed here:
12:18:36.864 INFO  BaseRecalibrationEngine -    ReadGroupCovariate
12:18:36.864 INFO  BaseRecalibrationEngine -    QualityScoreCovariat
e
12:18:36.864 INFO  BaseRecalibrationEngine -    ContextCovariate
```

```
12:18:36.864 INFO  BaseRecalibrationEngine -    CycleCovariate
12:18:36.866 INFO  ProgressMeter - Starting traversal
12:18:36.868 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes       Reads Processed     Reads/Minute
12:18:46.890 INFO  ProgressMeter -         chr1:9324324
0.2                221000        1323353.3
12:18:56.895 INFO  ProgressMeter -        chr1:19166719
0.3                466000        1396185.0
12:19:06.895 INFO  ProgressMeter -        chr1:28148751
0.5                727000        1452741.0
12:19:16.908 INFO  ProgressMeter -        chr1:38511604
0.7                983000        1473063.8
12:19:26.928 INFO  ProgressMeter -        chr1:47765700
0.8               1247000        1494606.5
12:19:36.949 INFO  ProgressMeter -        chr1:60759697
1.0               1481000        1479028.0
12:19:46.970 INFO  ProgressMeter -        chr1:74790110
1.2               1704000        1458467.1
12:19:57.001 INFO  ProgressMeter -        chr1:89019574
1.3               1928000        1443600.0
12:20:07.023 INFO  ProgressMeter -       chr1:101894217
1.5               2167000        1442198.9
12:20:17.039 INFO  ProgressMeter -       chr1:114680040
1.7               2408000        1442348.0
12:20:27.039 INFO  ProgressMeter -       chr1:151131070
1.8               2682000        1460651.7
12:20:37.073 INFO  ProgressMeter -       chr1:158449891
2.0               2970000        1482479.8
12:20:47.086 INFO  ProgressMeter -       chr1:169566282
2.2               3223000        1485059.6
12:20:57.114 INFO  ProgressMeter -       chr1:182421177
2.3               3470000        1484566.1
12:21:07.114 INFO  ProgressMeter -       chr1:196751604
2.5               3701000        1477976.1
12:21:17.135 INFO  ProgressMeter -       chr1:207245610
2.7               3967000        1485146.7
12:21:27.161 INFO  ProgressMeter -       chr1:220632909
2.8               4204000        1481211.8
12:21:37.167 INFO  ProgressMeter -       chr1:232596824
3.0               4453000        1481871.8
12:21:47.215 INFO  ProgressMeter -       chr1:245665671
3.2               4685000        1476784.4
12:21:50.465 INFO  BaseRecalibrator - 140773 read(s) filtered by:
((((((MappingQualityNotZeroReadFilter AND MappingQualityAvailableRea
dFilter) AND MappedReadFilter) AND NotSecondaryAlignmentReadFilter)
AND NotDuplicateReadFilter) AND PassesVendorQualityCheckReadFilter)
AND WellformedReadFilter)
  140773 read(s) filtered by: (((((MappingQualityNotZeroReadFilter A
ND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND NotS
econdaryAlignmentReadFilter) AND NotDuplicateReadFilter) AND PassesV
endorQualityCheckReadFilter)
      140773 read(s) filtered by: ((((MappingQualityNotZeroReadFilte
r AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND N
otSecondaryAlignmentReadFilter) AND NotDuplicateReadFilter)
          138661 read(s) filtered by: (((MappingQualityNotZeroReadFi
lter AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AN
D NotSecondaryAlignmentReadFilter)
              138661 read(s) filtered by: ((MappingQualityNotZeroRea
dFilter AND MappingQualityAvailableReadFilter) AND MappedReadFilter)
                  138661 read(s) filtered by: (MappingQualityNotZero
ReadFilter AND MappingQualityAvailableReadFilter)
```

```
                      138661 read(s) filtered by: MappingQualityNotZ
eroReadFilter
          2112 read(s) filtered by: NotDuplicateReadFilter

12:21:50.466 INFO  ProgressMeter -       chr1:249211294
3.2               4764859        1476727.8
12:21:50.466 INFO  ProgressMeter - Traversal complete. Processed 476
4859 total reads in 3.2 minutes.
12:21:50.741 INFO  BaseRecalibrator - Calculating quantized quality
scores...
12:21:50.768 INFO  BaseRecalibrator - Writing recalibration repor
t...
12:21:52.050 INFO  BaseRecalibrator - ...done!
12:21:52.050 INFO  BaseRecalibrator - Shutting down engine
[September 28, 2018 12:21:52 PM EAT] org.broadinstitute.hellbender.t
ools.walkers.bqsr.BaseRecalibrator done. Elapsed time: 3.27 minutes.
Runtime.totalMemory()=3238526976
Tool returned:
4764859

real    3m19.346s
user    3m57.234s
sys     0m13.003s
```

ApplyBQSR

In [9]:

```
!time gatk ApplyBQSR -R ../Data/Genome/chr1.fa \
-I ../Results/artefact_removal/set5_out_sorted_aln_dedup.bam \
--bqsr ../Results/BQSR/set5_pass1.table \
-O ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar ApplyBQSR -R
../Data/Genome/chr1.fa -I ../Results/artefact_removal/set5_out_sorte
d_aln_dedup.bam --bqsr ../Results/BQSR/set5_pass1.table -O ../Result
s/BQSR/set5_out_aln_dedup_pass1.adjusted.bam
12:25:56.229 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:26:01.367 INFO  ApplyBQSR - ------------------------------------
----------------------
12:26:01.368 INFO  ApplyBQSR - The Genome Analysis Toolkit (GATK) v
4.0.8.1
12:26:01.368 INFO  ApplyBQSR - For support and documentation go to h
ttps://software.broadinstitute.org/gatk/
12:26:01.369 INFO  ApplyBQSR - Executing as caleb@hpc01.icipe.org on
Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:26:01.369 INFO  ApplyBQSR - Java runtime: OpenJDK 64-Bit Server V
M v1.8.0_121-b15
12:26:01.370 INFO  ApplyBQSR - Start Date/Time: September 28, 2018 1
2:25:56 PM EAT
12:26:01.370 INFO  ApplyBQSR - ------------------------------------
----------------------
12:26:01.370 INFO  ApplyBQSR - ------------------------------------
----------------------
12:26:01.370 INFO  ApplyBQSR - HTSJDK Version: 2.16.0
12:26:01.371 INFO  ApplyBQSR - Picard Version: 2.18.7
12:26:01.371 INFO  ApplyBQSR - HTSJDK Defaults.COMPRESSION_LEVEL : 2
12:26:01.371 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR
_SAMTOOLS : false
12:26:01.371 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_SAMTOOLS : true
12:26:01.371 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_TRIBBLE : false
12:26:01.371 INFO  ApplyBQSR - Deflater: IntelDeflater
12:26:01.371 INFO  ApplyBQSR - Inflater: IntelInflater
12:26:01.371 INFO  ApplyBQSR - GCS max retries/reopens: 20
12:26:01.372 INFO  ApplyBQSR - Using google-cloud-java fork https://
github.com/broadinstitute/google-cloud-java/releases/tag/0.20.5-alph
a-GCS-RETRY-FIX
12:26:01.372 INFO  ApplyBQSR - Initializing engine
12:26:01.938 INFO  ApplyBQSR - Done initializing engine
12:26:01.956 INFO  ProgressMeter - Starting traversal
12:26:01.956 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes       Reads Processed     Reads/Minute
12:26:11.975 INFO  ProgressMeter -        chr1:12378211
0.2                325000        1946496.3
12:26:21.974 INFO  ProgressMeter -        chr1:26612325
0.3                705000        2113098.2
12:26:31.977 INFO  ProgressMeter -        chr1:42047175
0.5               1104000        2206455.5
12:26:41.993 INFO  ProgressMeter -        chr1:60559744
0.7               1509000        2261408.2
12:26:52.011 INFO  ProgressMeter -        chr1:86039160
0.8               1909000        2288282.9
12:27:02.033 INFO  ProgressMeter -       chr1:109240812
```

```
1.0                2302000        2299087.8
12:27:12.039 INFO  ProgressMeter -       chr1:147230852
1.2                2700000        2311544.9
12:27:22.058 INFO  ProgressMeter -       chr1:158016432
1.3                3087000        2312330.7
12:27:32.063 INFO  ProgressMeter -       chr1:176050212
1.5                3482000        2318576.8
12:27:42.071 INFO  ProgressMeter -       chr1:198834755
1.7                3883000        2327123.8
12:27:52.087 INFO  ProgressMeter -       chr1:216756556
1.8                4284000        2333968.9
12:28:02.106 INFO  ProgressMeter -       chr1:236978762
2.0                4684000        2339076.2
12:28:08.189 INFO  ApplyBQSR - No reads filtered by: WellformedReadF
ilter
12:28:08.190 INFO  ProgressMeter -       chr1:249211771
2.1                4905632        2331703.4
12:28:08.190 INFO  ProgressMeter - Traversal complete. Processed 490
5632 total reads in 2.1 minutes.
12:28:08.264 INFO  ApplyBQSR - Shutting down engine
[September 28, 2018 12:28:08 PM EAT] org.broadinstitute.hellbender.t
ools.walkers.bqsr.ApplyBQSR done. Elapsed time: 2.20 minutes.
Runtime.totalMemory()=2926575616

real    2m15.076s
user    3m10.256s
sys     0m29.418s
```

2. Generate the second pass recalibration table

In [10]:

```
!time gatk BaseRecalibrator -R ../Data/Genome/chr1.fa \
-I ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam \
--known-sites ../Data/VcfDatabase/common_all_20180418.vcf \
-O ../Results/BQSR/set5_pass2.table
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar BaseRecalibr
ator -R ../Data/Genome/chr1.fa -I ../Results/BQSR/set5_out_aln_dedup
_pass1.adjusted.bam --known-sites ../Data/VcfDatabase/common_all_201
80418.vcf -O ../Results/BQSR/set5_pass2.table
12:30:23.855 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:30:23.988 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:30:23.989 INFO  BaseRecalibrator - The Genome Analysis Toolkit (G
ATK) v4.0.8.1
12:30:23.989 INFO  BaseRecalibrator - For support and documentation
go to https://software.broadinstitute.org/gatk/
12:30:23.989 INFO  BaseRecalibrator - Executing as caleb@hpc01.icip
e.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:30:23.989 INFO  BaseRecalibrator - Java runtime: OpenJDK 64-Bit S
erver VM v1.8.0_121-b15
12:30:23.990 INFO  BaseRecalibrator - Start Date/Time: September 28,
2018 12:30:23 PM EAT
12:30:23.990 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:30:23.990 INFO  BaseRecalibrator - ----------------------------
----------------------------
12:30:23.990 INFO  BaseRecalibrator - HTSJDK Version: 2.16.0
12:30:23.990 INFO  BaseRecalibrator - Picard Version: 2.18.7
12:30:23.991 INFO  BaseRecalibrator - HTSJDK Defaults.COMPRESSION_LE
VEL : 2
12:30:23.991 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_R
EAD_FOR_SAMTOOLS : false
12:30:23.991 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_SAMTOOLS : true
12:30:23.991 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_W
RITE_FOR_TRIBBLE : false
12:30:23.991 INFO  BaseRecalibrator - Deflater: IntelDeflater
12:30:23.991 INFO  BaseRecalibrator - Inflater: IntelInflater
12:30:23.991 INFO  BaseRecalibrator - GCS max retries/reopens: 20
12:30:23.991 INFO  BaseRecalibrator - Using google-cloud-java fork h
ttps://github.com/broadinstitute/google-cloud-java/releases/tag/0.2
0.5-alpha-GCS-RETRY-FIX
12:30:23.991 INFO  BaseRecalibrator - Initializing engine
12:30:24.674 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Data/VcfDatabase/common_all_20180418.vcf
12:30:24.828 WARN  IndexUtils - Feature file "/opt/data/accreditatio
n/test/NextGenVariantCalling_set5/Code/../Data/VcfDatabase/common_al
l_20180418.vcf" appears to contain no sequence dictionary. Attemptin
g to retrieve a sequence dictionary from the associated index file
12:30:24.942 INFO  BaseRecalibrator - Done initializing engine
12:30:24.947 INFO  BaseRecalibrationEngine - The covariates being us
ed here:
12:30:24.947 INFO  BaseRecalibrationEngine -    ReadGroupCovariate
12:30:24.947 INFO  BaseRecalibrationEngine -    QualityScoreCovariat
e
12:30:24.947 INFO  BaseRecalibrationEngine -    ContextCovariate
```

```
12:30:24.947 INFO  BaseRecalibrationEngine -    CycleCovariate
12:30:24.949 INFO  ProgressMeter - Starting traversal
12:30:24.950 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes       Reads Processed     Reads/Minute
12:30:34.957 INFO  ProgressMeter -         chr1:9460620
0.2                223000        1337331.3
12:30:44.979 INFO  ProgressMeter -        chr1:19436673
0.3                475000        1423007.8
12:30:54.983 INFO  ProgressMeter -        chr1:28564258
0.5                739000        1476425.1
12:31:05.012 INFO  ProgressMeter -        chr1:39384629
0.7                995000        1490190.2
12:31:15.019 INFO  ProgressMeter -        chr1:48815133
0.8               1262000        1512313.0
12:31:25.029 INFO  ProgressMeter -        chr1:62222526
1.0               1498000        1496030.2
12:31:35.040 INFO  ProgressMeter -        chr1:76254930
1.2               1732000        1482686.3
12:31:45.048 INFO  ProgressMeter -        chr1:90678393
1.3               1964000        1471197.8
12:31:55.066 INFO  ProgressMeter -       chr1:104581257
1.5               2203000        1466776.2
12:32:05.111 INFO  ProgressMeter -       chr1:116916041
1.7               2451000        1468250.8
12:32:15.121 INFO  ProgressMeter -       chr1:152283977
1.8               2734000        1488958.1
12:32:25.137 INFO  ProgressMeter -       chr1:160011351
2.0               3015000        1505167.0
12:32:35.153 INFO  ProgressMeter -       chr1:171504993
2.2               3265000        1504585.2
12:32:45.171 INFO  ProgressMeter -       chr1:184000903
2.3               3513000        1503198.5
12:32:55.201 INFO  ProgressMeter -       chr1:198501585
2.5               3743000        1494708.8
12:33:05.202 INFO  ProgressMeter -       chr1:209212869
2.7               4005000        1499513.3
12:33:15.204 INFO  ProgressMeter -       chr1:222849443
2.8               4239000        1493885.6
12:33:25.218 INFO  ProgressMeter -       chr1:234606991
3.0               4488000        1493775.9
12:33:35.237 INFO  ProgressMeter -       chr1:247419258
3.2               4719000        1487962.9
12:33:36.994 INFO  BaseRecalibrator - 140773 read(s) filtered by:
((((((MappingQualityNotZeroReadFilter AND MappingQualityAvailableRea
dFilter) AND MappedReadFilter) AND NotSecondaryAlignmentReadFilter)
AND NotDuplicateReadFilter) AND PassesVendorQualityCheckReadFilter)
AND WellformedReadFilter)
  140773 read(s) filtered by: (((((MappingQualityNotZeroReadFilter A
ND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND NotS
econdaryAlignmentReadFilter) AND NotDuplicateReadFilter) AND PassesV
endorQualityCheckReadFilter)
      140773 read(s) filtered by: ((((MappingQualityNotZeroReadFilte
r AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND N
otSecondaryAlignmentReadFilter) AND NotDuplicateReadFilter)
          138661 read(s) filtered by: (((MappingQualityNotZeroReadFi
lter AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AN
D NotSecondaryAlignmentReadFilter)
              138661 read(s) filtered by: ((MappingQualityNotZeroRea
dFilter AND MappingQualityAvailableReadFilter) AND MappedReadFilter)
                  138661 read(s) filtered by: (MappingQualityNotZero
ReadFilter AND MappingQualityAvailableReadFilter)
```

```
                    138661 read(s) filtered by: MappingQualityNotZ
eroReadFilter
          2112 read(s) filtered by: NotDuplicateReadFilter

12:33:36.994 INFO  ProgressMeter -       chr1:249211294
3.2               4764859        1488677.3
12:33:36.994 INFO  ProgressMeter - Traversal complete. Processed 476
4859 total reads in 3.2 minutes.
12:33:37.232 INFO  BaseRecalibrator - Calculating quantized quality
scores...
12:33:37.259 INFO  BaseRecalibrator - Writing recalibration repor
t...
12:33:38.415 INFO  BaseRecalibrator - ...done!
12:33:38.415 INFO  BaseRecalibrator - Shutting down engine
[September 28, 2018 12:33:38 PM EAT] org.broadinstitute.hellbender.t
ools.walkers.bqsr.BaseRecalibrator done. Elapsed time: 3.24 minutes.
Runtime.totalMemory()=3364880384
Tool returned:
4764859

real    3m17.616s
user    3m55.445s
sys     0m15.142s
```

In [13]:

```
!time gatk ApplyBQSR -R ../Data/Genome/chr1.fa \
-I ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam \
--bqsr ../Results/BQSR/set5_pass2.table \
-O ../Results/BQSR/set5_out_aln_dedup_pass2.adjusted.bam
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar ApplyBQSR -R
../Data/Genome/chr1.fa -I ../Results/BQSR/set5_out_aln_dedup_pass1.a
djusted.bam --bqsr ../Results/BQSR/set5_pass2.table -O ../Results/BQ
SR/set5_out_aln_dedup_pass2.adjusted.bam
12:45:01.548 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:45:01.680 INFO  ApplyBQSR - ----------------------------------
----------------------
12:45:01.680 INFO  ApplyBQSR - The Genome Analysis Toolkit (GATK) v
4.0.8.1
12:45:01.680 INFO  ApplyBQSR - For support and documentation go to h
ttps://software.broadinstitute.org/gatk/
12:45:01.680 INFO  ApplyBQSR - Executing as caleb@hpc01.icipe.org on
Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:45:01.681 INFO  ApplyBQSR - Java runtime: OpenJDK 64-Bit Server V
M v1.8.0_121-b15
12:45:01.681 INFO  ApplyBQSR - Start Date/Time: September 28, 2018 1
2:45:01 PM EAT
12:45:01.681 INFO  ApplyBQSR - ----------------------------------
----------------------
12:45:01.681 INFO  ApplyBQSR - ----------------------------------
----------------------
12:45:01.682 INFO  ApplyBQSR - HTSJDK Version: 2.16.0
12:45:01.682 INFO  ApplyBQSR - Picard Version: 2.18.7
12:45:01.682 INFO  ApplyBQSR - HTSJDK Defaults.COMPRESSION_LEVEL : 2
12:45:01.682 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR
_SAMTOOLS : false
12:45:01.682 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_SAMTOOLS : true
12:45:01.683 INFO  ApplyBQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_TRIBBLE : false
12:45:01.683 INFO  ApplyBQSR - Deflater: IntelDeflater
12:45:01.683 INFO  ApplyBQSR - Inflater: IntelInflater
12:45:01.683 INFO  ApplyBQSR - GCS max retries/reopens: 20
12:45:01.683 INFO  ApplyBQSR - Using google-cloud-java fork https://
github.com/broadinstitute/google-cloud-java/releases/tag/0.20.5-alph
a-GCS-RETRY-FIX
12:45:01.683 INFO  ApplyBQSR - Initializing engine
12:45:02.295 INFO  ApplyBQSR - Done initializing engine
12:45:02.316 INFO  ProgressMeter - Starting traversal
12:45:02.316 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes       Reads Processed     Reads/Minute
12:45:12.328 INFO  ProgressMeter -        chr1:13183289
0.2                348000        2085705.7
12:45:22.328 INFO  ProgressMeter -        chr1:27440968
0.3                734000        2200789.6
12:45:32.351 INFO  ProgressMeter -        chr1:43213818
0.5               1126000        2249375.7
12:45:42.366 INFO  ProgressMeter -        chr1:61679850
0.7               1521000        2278651.7
12:45:52.379 INFO  ProgressMeter -        chr1:86362215
0.8               1917000        2297551.0
12:46:02.400 INFO  ProgressMeter -       chr1:109617746
```

```
1.0                2317000        2313760.7
12:46:12.413 INFO  ProgressMeter -       chr1:148018539
1.2                2716000        2324778.5
12:46:22.422 INFO  ProgressMeter -       chr1:158911460
1.3                3117000        2334656.6
12:46:32.428 INFO  ProgressMeter -       chr1:178156999
1.5                3515000        2340420.8
12:46:42.432 INFO  ProgressMeter -       chr1:200826280
1.7                3914000        2345702.4
12:46:52.432 INFO  ProgressMeter -       chr1:218968942
1.8                4311000        2348977.4
12:47:02.450 INFO  ProgressMeter -       chr1:237904335
2.0                4706000        2350375.4
12:47:07.551 INFO  ApplyBQSR - No reads filtered by: WellformedReadF
ilter
12:47:07.551 INFO  ProgressMeter -       chr1:249211771
2.1                4905632        2350284.8
12:47:07.551 INFO  ProgressMeter - Traversal complete. Processed 490
5632 total reads in 2.1 minutes.
12:47:07.628 INFO  ApplyBQSR - Shutting down engine
[September 28, 2018 12:47:07 PM EAT] org.broadinstitute.hellbender.t
ools.walkers.bqsr.ApplyBQSR done. Elapsed time: 2.10 minutes.
Runtime.totalMemory()=2757230592

real    2m9.076s
user    3m14.146s
sys     0m30.394s
```

!time gatk BaseRecalibrator -R ../Data/Genome/chr1.fa \ -I ../Results/BQSR/set5_out_aln_dedup_RG_firstpass.adjusted.bam \ --known-sites ../Data/VcfDatabase/common_all_20180418.vcf \ -O ../Results/BQSR/set5_secondpass.table

3. Finally generate the plots and also keep a copy of the csv (optional)

Currently, we have a problem with this step, as described here (https://github.com/bioconda/bioconda-recipes/issues/5350), but I am sure we can work something out with this. I fixed this error by intslling readline from conda-forge channel (https://anaconda.org/conda-forge/readline):

```
conda install -c conda-forge readline
```

Having confirmed that all is alright, we use the first bam output file for downstream analysis.

In [14]:

```
!time gatk AnalyzeCovariates \
-after ../Results/BQSR/set5_pass2.table \
-before ../Results/BQSR/set5_pass1.table \
-csv ../Results/BQSR/BQSR_Plots.csv \
-plots ../Results/BQSR/BQSR_Plots.pdf
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar AnalyzeCovar
iates -after ../Results/BQSR/set5_pass2.table -before ../Results/BQS
R/set5_pass1.table -csv ../Results/BQSR/BQSR_Plots.csv -plots ../Res
ults/BQSR/BQSR_Plots.pdf
12:47:24.723 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:47:24.854 INFO  AnalyzeCovariates - ----------------------------
------------------------------
12:47:24.854 INFO  AnalyzeCovariates - The Genome Analysis Toolkit
(GATK) v4.0.8.1
12:47:24.854 INFO  AnalyzeCovariates - For support and documentation
go to https://software.broadinstitute.org/gatk/
12:47:24.855 INFO  AnalyzeCovariates - Executing as caleb@hpc01.icip
e.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:47:24.855 INFO  AnalyzeCovariates - Java runtime: OpenJDK 64-Bit
Server VM v1.8.0_121-b15
12:47:24.855 INFO  AnalyzeCovariates - Start Date/Time: September 2
8, 2018 12:47:24 PM EAT
12:47:24.855 INFO  AnalyzeCovariates - ----------------------------
------------------------------
12:47:24.856 INFO  AnalyzeCovariates - ----------------------------
------------------------------
12:47:24.856 INFO  AnalyzeCovariates - HTSJDK Version: 2.16.0
12:47:24.856 INFO  AnalyzeCovariates - Picard Version: 2.18.7
12:47:24.856 INFO  AnalyzeCovariates - HTSJDK Defaults.COMPRESSION_L
EVEL : 2
12:47:24.856 INFO  AnalyzeCovariates - HTSJDK Defaults.USE_ASYNC_IO_
READ_FOR_SAMTOOLS : false
12:47:24.857 INFO  AnalyzeCovariates - HTSJDK Defaults.USE_ASYNC_IO_
WRITE_FOR_SAMTOOLS : true
12:47:24.857 INFO  AnalyzeCovariates - HTSJDK Defaults.USE_ASYNC_IO_
WRITE_FOR_TRIBBLE : false
12:47:24.857 INFO  AnalyzeCovariates - Deflater: IntelDeflater
12:47:24.857 INFO  AnalyzeCovariates - Inflater: IntelInflater
12:47:24.857 INFO  AnalyzeCovariates - GCS max retries/reopens: 20
12:47:24.857 INFO  AnalyzeCovariates - Using google-cloud-java fork
https://github.com/broadinstitute/google-cloud-java/releases/tag/0.2
0.5-alpha-GCS-RETRY-FIX
12:47:24.857 INFO  AnalyzeCovariates - Initializing engine
12:47:24.857 INFO  AnalyzeCovariates - Done initializing engine
12:47:25.606 INFO  AnalyzeCovariates - Generating csv file '../Resul
ts/BQSR/BQSR_Plots.csv'
12:47:25.728 INFO  AnalyzeCovariates - Generating plots file '../Res
ults/BQSR/BQSR_Plots.pdf'
12:47:37.973 INFO  AnalyzeCovariates - Shutting down engine
[September 28, 2018 12:47:37 PM EAT] org.broadinstitute.hellbender.t
ools.walkers.bqsr.AnalyzeCovariates done. Elapsed time: 0.22 minute
s.
Runtime.totalMemory()=2153250816
Tool returned:
Optional.empty

real    0m16.041s
```

```
user    0m27.442s
sys     0m1.146s
```

Variant Calling

There have also been quiet a bit of changes in this tool from the old to the new. We are now moving fully to gatk4. See details here (https://gatkforums.broadinstitute.org/gatk/discussion/8692/version-highlights-for-gatk-version-3-7)

Any annotations have to be applied at this stage since we do not need to use the Variant Annotator in teh pipeine. These are the annotations required in downsytream analysis:

b. Specify which annotations the program should use to evaluate the likelihood of SNPs being real

These annotations are included in the information generated for each variant call by the caller. If an annotation is missing (typically because it was omitted from the calling command) it can be added using the VariantAnnotator tool.

```
Coverage (DP)
```

Total (unfiltered) depth of coverage. Note that this statistic should not be used with exome datasets; see caveat detailed in the VQSR arguments FAQ doc.

```
QualByDepth (QD)
```

Variant confidence (from the QUAL field) / unfiltered depth of non-reference samples.

```
FisherStrand (FS)
```

Measure of strand bias (the variation being seen on only the forward or only the reverse strand). More bias is indicative of false positive calls. This complements the StrandOddsRatio (SOR) annotation.

```
StrandOddsRatio (SOR)
```

Measure of strand bias (the variation being seen on only the forward or only the reverse strand). More bias is indicative of false positive calls. This complements the FisherStrand (FS) annotation.

```
MappingQualityRankSumTest (MQRankSum)
```

The rank sum test for mapping qualities. Note that the mapping quality rank sum test can not be calculated for sites without a mixture of reads showing both the reference and alternate alleles.

```
ReadPosRankSumTest (ReadPosRankSum)
```

The rank sum test for the distance from the end of the reads. 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 can not be calculated for sites without a mixture of reads showing both the reference and alternate alleles.

```
RMSMappingQuality (MQ)
```

Estimation of the overall mapping quality of reads supporting a variant call.

```
Coverage, InbreedingCoeff
```

Evidence of inbreeding in a population. See caveats regarding population size and composition detailed in the VQSR arguments FAQ doc. "The InbreedingCoeff is a population level statistic that requires at least 10 samples in order to be computed. For projects with fewer samples, or that includes many closely related samples (such as a family) please omit this annotation from the command lin"

InbreedingCoeff

Depth of coverage (the DP annotation invoked by Coverage) should not be used when working with exome datasets since there is extreme variation in the depth to which targets are captured. In whole genome experiments this variation is indicative of error but that is not the case in capture experiments.

InbreedingCoeff is a population level statistic that requires at least 10 samples in order to be computed. For projects with fewer samples, or that includes many closely related samples (such as a family) please omit this annotation from the command line.

DP QD FS SOR MQRankSum ReadPosRankSum InbreedingCoeff MQ

One of the last steps in the germline short variant calling process is the calculation of the QUAL score for each candidate variant. The **stand-call-conf**, which is the standard caller confidence cut off filter based on QUAL scores was lowered in the latest versions of gatk.Once that's done, a threshold is applied on the QUAL score and we discard any variants that scored lower than the given threshold value.

In [15]:

```
!time gatk HaplotypeCaller -R ../Data/Genome/chr1.fa \
-I ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam \
-A ReadPosRankSumTest \
-A FisherStrand \
-A StrandOddsRatio \
-A MappingQualityRankSumTest \
-A QualByDepth \
-A InbreedingCoeff \
-A RMSMappingQuality \
--genotyping-mode DISCOVERY -stand-call-conf 10 \
-O ../Results/Variants/gatk/setr5_out_raw_variants.vcf
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar HaplotypeCal
ler -R ../Data/Genome/chr1.fa -I ../Results/BQSR/set5_out_aln_dedup_
pass1.adjusted.bam -A ReadPosRankSumTest -A FisherStrand -A StrandOd
dsRatio -A MappingQualityRankSumTest -A QualByDepth -A InbreedingCoe
ff -A RMSMappingQuality --genotyping-mode DISCOVERY -stand-call-conf
10 -O ../Results/Variants/gatk/setr5_out_raw_variants.vcf
12:55:39.148 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
12:55:39.288 INFO  HaplotypeCaller - ----------------------------
----------------------------
12:55:39.289 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GA
TK) v4.0.8.1
12:55:39.289 INFO  HaplotypeCaller - For support and documentation g
o to https://software.broadinstitute.org/gatk/
12:55:39.289 INFO  HaplotypeCaller - Executing as caleb@hpc01.icipe.
org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
12:55:39.289 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Se
rver VM v1.8.0_121-b15
12:55:39.290 INFO  HaplotypeCaller - Start Date/Time: September 28,
2018 12:55:39 PM EAT
12:55:39.290 INFO  HaplotypeCaller - ----------------------------
----------------------------
12:55:39.290 INFO  HaplotypeCaller - ----------------------------
----------------------------
12:55:39.291 INFO  HaplotypeCaller - HTSJDK Version: 2.16.0
12:55:39.291 INFO  HaplotypeCaller - Picard Version: 2.18.7
12:55:39.291 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEV
EL : 2
12:55:39.291 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_RE
AD_FOR_SAMTOOLS : false
12:55:39.291 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WR
ITE_FOR_SAMTOOLS : true
12:55:39.291 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WR
ITE_FOR_TRIBBLE : false
12:55:39.291 INFO  HaplotypeCaller - Deflater: IntelDeflater
12:55:39.291 INFO  HaplotypeCaller - Inflater: IntelInflater
12:55:39.291 INFO  HaplotypeCaller - GCS max retries/reopens: 20
12:55:39.292 INFO  HaplotypeCaller - Using google-cloud-java fork ht
tps://github.com/broadinstitute/google-cloud-java/releases/tag/0.20.
5-alpha-GCS-RETRY-FIX
12:55:39.292 INFO  HaplotypeCaller - Initializing engine
12:55:39.900 INFO  HaplotypeCaller - Done initializing engine
12:55:39.909 INFO  HaplotypeCallerEngine - Disabling physical phasin
g, which is supported only for reference-model confidence output
12:55:39.925 INFO  NativeLibraryLoader - Loading libgkl_utils.so fro
m jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.0.8.1
-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgkl_util
s.so
12:55:39.927 INFO  NativeLibraryLoader - Loading libgkl_pairhmm_omp.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_pairhmm_omp.so
12:55:39.990 WARN  IntelPairHmm - Flush-to-zero (FTZ) is enabled whe
```

```
n running PairHMM
12:55:39.991 INFO  IntelPairHmm - Available threads: 64
12:55:39.991 INFO  IntelPairHmm - Requested threads: 4
12:55:39.991 INFO  PairHMM - Using the OpenMP multi-threaded AVX-acc
elerated native PairHMM implementation
12:55:40.036 INFO  ProgressMeter - Starting traversal
12:55:40.036 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes     Regions Processed   Regions/Minute
12:55:50.063 INFO  ProgressMeter -         chr1:1396976
0.2                  5210          31182.0
12:56:00.073 INFO  ProgressMeter -         chr1:2558303
0.3                  9760          29228.8
12:56:10.086 INFO  ProgressMeter -         chr1:5224356
0.5                 19760          39455.6
12:56:20.094 INFO  ProgressMeter -         chr1:7427590
0.7                 28080          42060.1
12:56:30.095 INFO  ProgressMeter -         chr1:9819358
0.8                 37100          44468.4
12:56:40.118 INFO  ProgressMeter -        chr1:11849485
1.0                 44780          44719.6
12:56:50.135 INFO  ProgressMeter -        chr1:13943611
1.2                 52430          44877.2
12:57:00.138 INFO  ProgressMeter -        chr1:16273990
1.3                 61250          45879.6
12:57:10.472 INFO  ProgressMeter -        chr1:16977019
1.5                 64170          42574.2
12:57:20.473 INFO  ProgressMeter -        chr1:18830977
1.7                 71230          42552.5
12:57:30.505 INFO  ProgressMeter -        chr1:21051596
1.8                 79650          43261.4
12:57:40.518 INFO  ProgressMeter -        chr1:23236864
2.0                 87900          43774.5
12:57:50.535 INFO  ProgressMeter -        chr1:25697256
2.2                 97150          44667.7
12:58:00.540 INFO  ProgressMeter -        chr1:27972613
2.3                105720          45146.4
12:58:10.551 INFO  ProgressMeter -        chr1:31193954
2.5                117630          46891.3
12:58:20.588 INFO  ProgressMeter -        chr1:33549562
2.7                126480          47267.2
12:58:30.597 INFO  ProgressMeter -        chr1:36129866
2.8                136160          47898.7
12:58:40.598 INFO  ProgressMeter -        chr1:38739948
3.0                145880          48475.9
12:58:50.603 INFO  ProgressMeter -        chr1:41315875
3.2                155510          48962.6
12:59:00.603 INFO  ProgressMeter -        chr1:43955712
3.3                165370          49471.0
12:59:10.604 INFO  ProgressMeter -        chr1:46540718
3.5                174970          49856.8
12:59:20.618 INFO  ProgressMeter -        chr1:49368140
3.7                185470          50449.5
12:59:30.660 INFO  ProgressMeter -        chr1:52941585
3.8                198520          51647.9
12:59:40.664 INFO  ProgressMeter -        chr1:55487384
4.0                208070          51882.0
12:59:50.671 INFO  ProgressMeter -        chr1:58596005
4.2                219590          52568.3
13:00:00.673 INFO  ProgressMeter -        chr1:61974225
4.3                232030          53414.9
13:00:10.691 INFO  ProgressMeter -        chr1:64669605
```

```
4.5                242090          53667.8
13:00:20.701 INFO  ProgressMeter -        chr1:67579724
4.7                252920          54068.9
13:00:30.713 INFO  ProgressMeter -        chr1:71037300
4.8                265570          54819.8
13:00:40.710 INFO  ProgressMeter -        chr1:74837626
5.0                279440          55762.9
13:00:50.743 INFO  ProgressMeter -        chr1:78041846
5.2                291280          56248.7
13:01:00.746 INFO  ProgressMeter -        chr1:81436865
5.3                303790          56834.7
13:01:10.750 INFO  ProgressMeter -        chr1:85100102
5.5                317200          57548.4
13:01:20.752 INFO  ProgressMeter -        chr1:87643374
5.7                326730          57537.2
13:01:30.754 INFO  ProgressMeter -        chr1:90922951
5.8                338830          57966.4
13:01:40.760 INFO  ProgressMeter -        chr1:93596673
6.0                348870          58028.5
13:01:50.765 INFO  ProgressMeter -        chr1:96475463
6.2                359580          58195.8
13:02:00.780 INFO  ProgressMeter -       chr1:100195946
6.3                373180          58808.2
13:02:10.791 INFO  ProgressMeter -       chr1:103007128
6.5                383680          58913.9
13:02:20.794 INFO  ProgressMeter -       chr1:106846170
6.7                397650          59534.8
13:02:30.810 INFO  ProgressMeter -       chr1:109729015
6.8                408390          59651.9
13:02:40.831 INFO  ProgressMeter -       chr1:111864232
7.0                416460          59382.0
13:02:50.837 INFO  ProgressMeter -       chr1:114509487
7.2                426360          59381.6
13:03:01.170 INFO  ProgressMeter -       chr1:117157752
7.4                436230          59333.1
13:03:11.173 INFO  ProgressMeter -       chr1:120304286
7.5                447820          59559.1
13:03:21.172 INFO  ProgressMeter -       chr1:141399406
7.7                518450          67457.5
13:03:31.242 INFO  ProgressMeter -       chr1:144911320
7.9                531050          67620.3
13:03:41.249 INFO  ProgressMeter -       chr1:147066406
8.0                539100          67217.8
13:03:51.284 INFO  ProgressMeter -       chr1:150532298
8.2                551630          67375.1
13:04:01.286 INFO  ProgressMeter -       chr1:152279547
8.4                558330          66832.7
13:04:11.290 INFO  ProgressMeter -       chr1:154036096
8.5                565070          66315.9
13:04:21.294 INFO  ProgressMeter -       chr1:155880354
8.7                572120          65854.7
13:04:31.304 INFO  ProgressMeter -       chr1:157909055
8.9                579820          65483.5
13:04:41.325 INFO  ProgressMeter -       chr1:160100468
9.0                588100          65189.1
13:04:51.324 INFO  ProgressMeter -       chr1:161777686
9.2                594570          64710.8
13:05:01.326 INFO  ProgressMeter -       chr1:165092201
9.4                606820          64867.1
13:05:11.338 INFO  ProgressMeter -       chr1:167868024
9.5                617190          64819.4
```

```
13:05:21.346 INFO  ProgressMeter -        chr1:170455995
9.7                 626870          64702.7
13:05:31.356 INFO  ProgressMeter -        chr1:173457862
9.9                 637970          64733.6
13:05:41.378 INFO  ProgressMeter -        chr1:176289640
10.0                 648530          64708.5
13:05:51.390 INFO  ProgressMeter -        chr1:179354072
10.2                 659860          64760.6
13:06:01.402 INFO  ProgressMeter -        chr1:182430956
10.4                 671230          64815.0
13:06:11.415 INFO  ProgressMeter -        chr1:185099661
10.5                 681220          64736.5
13:06:21.437 INFO  ProgressMeter -        chr1:187956459
10.7                 691800          64714.7
13:06:31.437 INFO  ProgressMeter -        chr1:191940271
10.9                 706340          65060.5
13:06:41.438 INFO  ProgressMeter -        chr1:195723571
11.0                 720180          65332.2
13:06:51.450 INFO  ProgressMeter -        chr1:198371098
11.2                 730040          65239.1
13:07:01.495 INFO  ProgressMeter -        chr1:201176534
11.4                 740490          65197.6
13:07:11.541 INFO  ProgressMeter -        chr1:203454418
11.5                 749030          64991.5
13:07:21.545 INFO  ProgressMeter -        chr1:205802607
11.7                 757900          64823.2
13:07:31.589 INFO  ProgressMeter -        chr1:208214630
11.9                 766900          64667.2
13:07:41.589 INFO  ProgressMeter -        chr1:211240988
12.0                 778140          64705.5
13:07:51.627 INFO  ProgressMeter -        chr1:214506677
12.2                 790190          64806.0
13:08:01.631 INFO  ProgressMeter -        chr1:217668687
12.4                 801850          64875.1
13:08:11.644 INFO  ProgressMeter -        chr1:220864100
12.5                 813620          64950.4
13:08:21.647 INFO  ProgressMeter -        chr1:223953552
12.7                 825050          64997.8
13:08:31.679 INFO  ProgressMeter -        chr1:226568195
12.9                 834860          64915.6
13:08:41.703 INFO  ProgressMeter -        chr1:229250655
13.0                 844860          64850.7
13:08:51.707 INFO  ProgressMeter -        chr1:231685476
13.2                 854010          64724.8
13:09:01.714 INFO  ProgressMeter -        chr1:234638788
13.4                 865020          64740.8
13:09:11.718 INFO  ProgressMeter -        chr1:236950080
13.5                 873760          64588.9
13:09:21.735 INFO  ProgressMeter -        chr1:239892449
13.7                 884710          64601.2
13:09:31.737 INFO  ProgressMeter -        chr1:243008402
13.9                 896270          64658.2
13:09:41.742 INFO  ProgressMeter -        chr1:246294607
14.0                 908370          64752.1
13:09:51.865 INFO  ProgressMeter -        chr1:248437066
14.2                 916440          64551.1
13:09:57.412 INFO  HaplotypeCaller - 150847 read(s) filtered by:
(((((((MappingQualityReadFilter AND MappingQualityAvailableReadFilt
er) AND MappedReadFilter) AND NotSecondaryAlignmentReadFilter) AND N
otDuplicateReadFilter) AND PassesVendorQualityCheckReadFilter) AND N
onZeroReferenceLengthAlignmentReadFilter) AND GoodCigarReadFilter) A
```

```
ND WellformedReadFilter)
  150847 read(s) filtered by: (((((((MappingQualityReadFilter AND Ma
ppingQualityAvailableReadFilter) AND MappedReadFilter) AND NotSecond
aryAlignmentReadFilter) AND NotDuplicateReadFilter) AND PassesVendor
QualityCheckReadFilter) AND NonZeroReferenceLengthAlignmentReadFilte
r) AND GoodCigarReadFilter)
      150847 read(s) filtered by: ((((((MappingQualityReadFilter AND
MappingQualityAvailableReadFilter) AND MappedReadFilter) AND NotSeco
ndaryAlignmentReadFilter) AND NotDuplicateReadFilter) AND PassesVend
orQualityCheckReadFilter) AND NonZeroReferenceLengthAlignmentReadFil
ter)
          150847 read(s) filtered by: (((((MappingQualityReadFilter
AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND Not
SecondaryAlignmentReadFilter) AND NotDuplicateReadFilter) AND Passes
VendorQualityCheckReadFilter)
              150847 read(s) filtered by: ((((MappingQualityReadFilt
er AND MappingQualityAvailableReadFilter) AND MappedReadFilter) AND
NotSecondaryAlignmentReadFilter) AND NotDuplicateReadFilter)
                  148740 read(s) filtered by: (((MappingQualityReadF
ilter AND MappingQualityAvailableReadFilter) AND MappedReadFilter) A
ND NotSecondaryAlignmentReadFilter)
                      148740 read(s) filtered by: ((MappingQualityRe
adFilter AND MappingQualityAvailableReadFilter) AND MappedReadFilte
r)
                          148740 read(s) filtered by: (MappingQualit
yReadFilter AND MappingQualityAvailableReadFilter)
                              148740 read(s) filtered by: MappingQua
lityReadFilter
                  2107 read(s) filtered by: NotDuplicateReadFilter

13:09:57.413 INFO  ProgressMeter -       chr1:249248204
14.3                919438         64343.2
13:09:57.413 INFO  ProgressMeter - Traversal complete. Processed 919
438 total regions in 14.3 minutes.
13:09:57.671 INFO  VectorLoglessPairHMM - Time spent in setup for JN
I call : 0.341741218
13:09:57.671 INFO  PairHMM - Total compute time in PairHMM computeLo
gLikelihoods() : 12.907999106
13:09:57.671 INFO  SmithWatermanAligner - Total compute time in java
Smith-Waterman : 77.15 sec
13:09:57.671 INFO  HaplotypeCaller - Shutting down engine
[September 28, 2018 1:09:57 PM EAT] org.broadinstitute.hellbender.to
ols.walkers.haplotypecaller.HaplotypeCaller done. Elapsed time: 14.3
1 minutes.
Runtime.totalMemory()=2811232256

real    14m22.202s
user    19m15.461s
sys     0m43.613s
```

Checking the number of variants discovered

In [17]:

```
!grep -c '^chr1' ../Results/Variants/gatk/setr5_out_raw_variants.vcf
```

```
28315
```

Generate the statistics from the vcf file

In [18]:

```
!bcftools stats ../Results/Variants/gatk/setr5_out_raw_variants.vcf  \
> ../Results/Variants/gatk/setr5_out_raw_variants.vcf.stats
```

In [19]:

```
!plot-vcfstats ../Results/Variants/gatk/setr5_out_raw_variants.vcf.stats \
-p ../Results/Variants/gatk/plots
```

```
Parsing bcftools stats output: ../Results/Variants/gatk/setr5_out_ra
w_variants.vcf.stats
Plotting graphs: python plot.py
Creating PDF: pdflatex summary.tex >plot-vcfstats.log 2>&1
Finished: ../Results/Variants/gatk/plots/summary.pdf
```

VQSR

The details of this step are available here (https://software.broadinstitute.org/gatk/documentation/article?id=11084). This is a post-variant calling step that we are yet to full understand.

First, we need to convert the files to remove chr

1. Edit the chr file

In [84]:

```
!sed 's/chr//g' ../Data/Genome/chr1.fa > ../Data/Genome/chr1_edited.fa
```

2. Then we create a dictionary of the edited file

In [86]:

```
!rm ../Data/Genome/chr1_edited.dict
```

In [87]:

```
!time picard CreateSequenceDictionary \
R=../Data/Genome/chr1_edited.fa \
O=../Data/Genome/chr1_edited.dict
```

```
18:05:28.259 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/picard-
2.18.11-0/picard.jar!/com/intel/gkl/native/libgkl_compression.so
[Mon Sep 24 18:05:28 EAT 2018] CreateSequenceDictionary OUTPUT=../Da
ta/Genome/chr1_edited.dict REFERENCE=../Data/Genome/chr1_edited.fa
TRUNCATE_NAMES_AT_WHITESPACE=true NUM_SEQUENCES=2147483647 VERBOSITY
=INFO QUIET=false VALIDATION_STRINGENCY=STRICT COMPRESSION_LEVEL=5 M
AX_RECORDS_IN_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA
4GH_CLIENT_SECRETS=client_secrets.json USE_JDK_DEFLATER=false USE_JD
K_INFLATER=false
[Mon Sep 24 18:05:28 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is not a
vailable; Picard version: 2.18.11-SNAPSHOT
[Mon Sep 24 18:05:29 EAT 2018] picard.sam.CreateSequenceDictionary d
one. Elapsed time: 0.03 minutes.
Runtime.totalMemory()=514850816

real    0m2.742s
user    0m3.827s
sys     0m0.528s
```

3. We index the edited file using samtools

In [24]:

```
!time samtools faidx ../Data/Genome/chr1_edited.fa
```

```
real    0m3.133s
user    0m2.831s
sys     0m0.112s
```

4. We also edit chr from the variant file

In [20]:

```
!sed 's/chr//g' ../Results/Variants/gatk/setr5_out_raw_variants.vcf \
> ../Results/Variants/gatk/setr5_out_raw_variants_edited.vcf
```

5. We then edit the alignemnt file:

a) First Convert the alignment to sam

In [21]:

```
!samtools view -h -o ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.sam \
../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam
```

b) change chr to 1

In [22]:

```
!sed 's/chr//g' ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.sam \
> ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.edited.sam
```

c) Finally Convert alignment back to the bam format

In [23]:

```
!samtools view -h -o ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.edited.ba
m \
../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.edited.sam
```

Variant Quality Analysis with VQSR

I am not sure if I need to run this step in the current gatk verson,so I may need to explre and figure out. This may not be arequred step, a long as we request for annotation during variant calling step.

https://gatkforums.broadinstitute.org/gatk/discussion/11187/has-variantannotator-tool-been-removed-from-gatk (https://gatkforums.broadinstitute.org/gatk/discussion/11187/has-variantannotator-tool-been-removed-from-gatk)

https://software.broadinstitute.org/gatk/documentation/article?id=6022 (https://software.broadinstitute.org/gatk/documentation/article?id=6022)

Refer to this documentation (https://gatkforums.broadinstitute.org/gatk/discussion/2805/) for detailed information on how the tools should be run.

NB: The sequences were aligned to b37 version of the genome. We need to identify the best genome we should be using in this analysis.

So we can see that there is so many variations in the files we were given and the final file we are using. First: We need to check with the data and confirm that we are actually aligning to the correct genome.

First round I got no output, following the recommendations in this thread (https://gatkforums.broadinstitute.org/gatk/discussion/11384/values-for-qd-annotation-not-detected-for-any-training-variant-in-the-input-callset), I decided to change the max-gaussian to 4.

The reasone for this is aptly explained here (http://discussions4562.rssing.com/chan-67237868/all_p22.html): "maxGaussians is the maximum number of different "clusters" (=Gaussians) of variants the program is "allowed" to try to identify. Lowering this number forces the program to group variants into a smaller number of clusters, which means there will be more variants in each cluster -- hopefully enough to satisfy the statistical requirements. Of course, this decreases the level of discrimination that you can achieve between variant profiles/error modes. It's all about trade-offs; and unfortunately if you don't have a lot of variants you can't afford to be very demanding in terms of resolution."

Let's start with SNPs

1. Variant Recalibration

This step...

We may need to think how we can apply the same to SNPs and Indels.

In [24]:

```
!time gatk VariantRecalibrator \
-R ../Data/Genome/chr1_edited.fa \
-V ../Results/Variants/gatk/setr5_out_raw_variants_edited.vcf \
-O ../Results/VQSR/setr5_out_raw_variants_SNP.recal \
-resource hapmap,known=false,training=true,truth=true,prior=15.0:/opt/data/accre
ditation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/hapmap_3.3.b37.vcf
\
-resource omni,known=false,training=true,truth=false,prior=12.0:/opt/data/accred
itation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/1000G_omni2.5.b37.v
cf \
-resource 1000G,known=false,training=true,truth=false,prior=10.0:/opt/data/accre
ditation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/1000G_phase1.snps.
high_confidence.b37.vcf \
-resource dbsnp,known=true,training=false,truth=false,prior=2.0:/opt/data/accred
itation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/dbsnp_138.b37.vcf \
-an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum -an MQ \
-mode SNP \
--max-gaussians 4 \
--tranches-file ../Results/VQSR/setr5_out_raw_variants_SNP.tranches \
--rscript-file ../Results/VQSR/setr5_out_SNP.plots.R
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar VariantRecal
ibrator -R ../Data/Genome/chr1_edited.fa -V ../Results/Variants/gat
k/setr5_out_raw_variants_edited.vcf -O ../Results/VQSR/setr5_out_raw
_variants_SNP.recal -resource hapmap,known=false,training=true,truth
=true,prior=15.0:/opt/data/accreditation/test/NextGenVariantCalling_
set5/Data/VcfDatabase/b37/hapmap_3.3.b37.vcf -resource omni,known=fa
lse,training=true,truth=false,prior=12.0:/opt/data/accreditation/tes
t/NextGenVariantCalling_set5/Data/VcfDatabase/b37/1000G_omni2.5.b37.
vcf -resource 1000G,known=false,training=true,truth=false,prior=10.
0:/opt/data/accreditation/test/NextGenVariantCalling_set5/Data/VcfDa
tabase/b37/1000G_phase1.snps.high_confidence.b37.vcf -resource dbsn
p,known=true,training=false,truth=false,prior=2.0:/opt/data/accredit
ation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/dbsnp_13
8.b37.vcf -an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum -an
MQ -mode SNP --max-gaussians 4 --tranches-file ../Results/VQSR/setr5
_out_raw_variants_SNP.tranches --rscript-file ../Results/VQSR/setr5_
out_SNP.plots.R
16:26:47.943 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
16:26:48.074 INFO  VariantRecalibrator - ----------------------------
--------------------------------
16:26:48.074 INFO  VariantRecalibrator - The Genome Analysis Toolkit
(GATK) v4.0.8.1
16:26:48.075 INFO  VariantRecalibrator - For support and documentati
on go to https://software.broadinstitute.org/gatk/
16:26:48.075 INFO  VariantRecalibrator - Executing as caleb@hpc01.ic
ipe.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
16:26:48.075 INFO  VariantRecalibrator - Java runtime: OpenJDK 64-Bi
t Server VM v1.8.0_121-b15
16:26:48.075 INFO  VariantRecalibrator - Start Date/Time: September
28, 2018 4:26:47 PM EAT
16:26:48.075 INFO  VariantRecalibrator - ----------------------------
--------------------------------
16:26:48.075 INFO  VariantRecalibrator - ----------------------------
--------------------------------
16:26:48.076 INFO  VariantRecalibrator - HTSJDK Version: 2.16.0
16:26:48.076 INFO  VariantRecalibrator - Picard Version: 2.18.7
16:26:48.076 INFO  VariantRecalibrator - HTSJDK Defaults.COMPRESSION
_LEVEL : 2
16:26:48.077 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_READ_FOR_SAMTOOLS : false
16:26:48.077 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_WRITE_FOR_SAMTOOLS : true
16:26:48.077 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_WRITE_FOR_TRIBBLE : false
16:26:48.077 INFO  VariantRecalibrator - Deflater: IntelDeflater
16:26:48.077 INFO  VariantRecalibrator - Inflater: IntelInflater
16:26:48.077 INFO  VariantRecalibrator - GCS max retries/reopens: 20
16:26:48.077 INFO  VariantRecalibrator - Using google-cloud-java for
k https://github.com/broadinstitute/google-cloud-java/releases/tag/
0.20.5-alpha-GCS-RETRY-FIX
16:26:48.077 INFO  VariantRecalibrator - Initializing engine
16:26:48.706 INFO  FeatureManager - Using codec VCFCodec to read fil
```

```
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/hapmap_3.3.b37.vcf
16:26:48.785 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/1000G_omni2.5.b37.vcf
16:26:48.811 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/1000G_phase1.snps.high_confidence.b37.vcf
16:26:48.915 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/dbsnp_138.b37.vcf
16:26:49.042 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/Variants/gatk/setr5_out_raw_variants_edited.vcf
16:26:49.071 WARN  IndexUtils - Feature file "/opt/data/accreditatio
n/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/1000G_phase1.
snps.high_confidence.b37.vcf" appears to contain no sequence diction
ary. Attempting to retrieve a sequence dictionary from the associate
d index file
16:26:49.171 INFO  VariantRecalibrator - Done initializing engine
16:26:49.174 INFO  TrainingSet - Found hapmap track:    Known = fals
e       Training = true         Truth = true    Prior = Q15.0
16:26:49.175 INFO  TrainingSet - Found omni track:      Known = fals
e       Training = true         Truth = false   Prior = Q12.0
16:26:49.175 INFO  TrainingSet - Found 1000G track:     Known = fals
e       Training = true         Truth = false   Prior = Q10.0
16:26:49.175 INFO  TrainingSet - Found dbsnp track:     Known = true
Training = false        Truth = false   Prior = Q2.0
16:26:49.183 WARN  GATKVariantContextUtils - Can't determine output
variant file format from output file extension "recal". Defaulting t
o VCF.
16:26:49.206 INFO  ProgressMeter - Starting traversal
16:26:49.206 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes    Variants Processed  Variants/Minute
16:27:00.040 INFO  ProgressMeter -           1:47717245
0.2                  8000          44305.0
16:27:11.156 INFO  ProgressMeter -          1:102843599
0.4                 13000          35535.3
16:27:22.692 INFO  ProgressMeter -          1:183920452
0.6                 21000          37627.7
16:27:34.070 INFO  ProgressMeter -          1:237711797
0.7                 27000          36109.1
16:27:36.547 INFO  ProgressMeter -          1:247768812
0.8                 28315          35886.4
16:27:36.547 INFO  ProgressMeter - Traversal complete. Processed 283
15 total variants in 0.8 minutes.
16:27:36.556 INFO  VariantDataManager - QD:      mean = 20.84    sta
ndard deviation = 8.55
16:27:36.566 INFO  VariantDataManager - FS:      mean = 0.84     sta
ndard deviation = 2.29
16:27:36.576 INFO  VariantDataManager - SOR:     mean = 1.61     sta
ndard deviation = 1.25
16:27:36.583 INFO  VariantDataManager - MQRankSum:       mean = -0.0
3        standard deviation = 0.33
16:27:36.592 INFO  VariantDataManager - ReadPosRankSum:          mea
n = 0.03         standard deviation = 1.00
16:27:36.597 INFO  VariantDataManager - MQ:      mean = 59.86    sta
ndard deviation = 1.36
16:27:36.642 INFO  VariantDataManager - Annotations are now ordered
by their information content: [MQ, QD, SOR, FS, MQRankSum, ReadPosRa
nkSum]
```

```
16:27:36.649 INFO  VariantDataManager - Training with 15748 variants
after standard deviation thresholding.
16:27:36.653 INFO  GaussianMixtureModel - Initializing model with 10
0 k-means iterations...
16:27:37.221 INFO  VariantRecalibratorEngine - Finished iteration 0.
16:27:37.609 INFO  VariantRecalibratorEngine - Finished iteration 5.
Current change in mixture coefficients = 0.16156
16:27:37.856 INFO  VariantRecalibratorEngine - Finished iteration 1
0.      Current change in mixture coefficients = 0.00512
16:27:37.901 INFO  VariantRecalibratorEngine - Convergence after 11
iterations!
16:27:37.951 INFO  VariantRecalibratorEngine - Evaluating full set o
f 26875 variants...
16:27:39.114 INFO  VariantDataManager - Selected worst 663 scoring v
ariants --> variants with LOD <= -5.0000.
16:27:39.114 INFO  GaussianMixtureModel - Initializing model with 10
0 k-means iterations...
16:27:39.124 INFO  VariantRecalibratorEngine - Finished iteration 0.
16:27:39.129 INFO  VariantRecalibratorEngine - Finished iteration 5.
Current change in mixture coefficients = 0.00654
16:27:39.130 INFO  VariantRecalibratorEngine - Convergence after 6 i
terations!
16:27:39.133 INFO  VariantRecalibratorEngine - Evaluating full set o
f 26875 variants...
16:27:39.993 INFO  TrancheManager - Finding 4 tranches for 26875 var
iants
16:27:40.022 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 100.00 => selection metric threshold 0.000
16:27:40.039 INFO  TrancheManager -   Found tranche for 100.000: 0.0
00 threshold starting with variant 0; running score is 0.000
16:27:40.039 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=100.00 minVQSLod=-13.323
1 known=(18584 @ 2.2688) novel=(8291 @ 2.1987) truthSites(8620 acces
sible, 8620 called), name=anonymous]
16:27:40.040 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 99.90 => selection metric threshold 0.001
16:27:40.048 INFO  TrancheManager -   Found tranche for 99.900: 0.00
1 threshold starting with variant 1207; running score is 0.001
16:27:40.048 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=99.90 minVQSLod=-1.8894
known=(17659 @ 2.2932) novel=(8009 @ 2.2074) truthSites(8620 accessi
ble, 8611 called), name=anonymous]
16:27:40.048 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 99.00 => selection metric threshold 0.010
16:27:40.055 INFO  TrancheManager -   Found tranche for 99.000: 0.01
0 threshold starting with variant 2085; running score is 0.010
16:27:40.055 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=99.00 minVQSLod=4.2506 k
nown=(16940 @ 2.3052) novel=(7850 @ 2.2028) truthSites(8620 accessib
le, 8533 called), name=anonymous]
16:27:40.055 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 90.00 => selection metric threshold 0.100
16:27:40.059 INFO  TrancheManager -   Found tranche for 90.000: 0.10
0 threshold starting with variant 13137; running score is 0.100
16:27:40.060 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=90.00 minVQSLod=21.2578
known=(12736 @ 2.3487) novel=(1002 @ 2.2961) truthSites(8620 accessi
ble, 7758 called), name=anonymous]
16:27:40.071 INFO  VariantRecalibrator - Writing out recalibration t
able...
16:27:40.518 INFO  VariantRecalibrator - Writing out visualization R
```

```
script file...
16:27:40.537 INFO  VariantRecalibrator - Building MQ x QD plot...
16:27:40.542 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:41.124 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:41.515 INFO  VariantRecalibrator - Building MQ x SOR plot...
16:27:41.519 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:42.097 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:42.451 INFO  VariantRecalibrator - Building MQ x FS plot...
16:27:42.452 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:43.045 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:43.394 INFO  VariantRecalibrator - Building MQ x MQRankSum plo
t...
16:27:43.396 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:43.987 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:44.341 INFO  VariantRecalibrator - Building MQ x ReadPosRankSu
m plot...
16:27:44.342 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:44.921 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:45.273 INFO  VariantRecalibrator - Building QD x SOR plot...
16:27:45.274 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:45.888 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:46.228 INFO  VariantRecalibrator - Building QD x FS plot...
16:27:46.229 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:46.819 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:47.160 INFO  VariantRecalibrator - Building QD x MQRankSum plo
t...
16:27:47.161 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:47.751 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:48.106 INFO  VariantRecalibrator - Building QD x ReadPosRankSu
m plot...
16:27:48.106 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:48.687 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:49.032 INFO  VariantRecalibrator - Building SOR x FS plot...
16:27:49.033 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:49.622 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:49.959 INFO  VariantRecalibrator - Building SOR x MQRankSum pl
ot...
16:27:49.959 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:50.553 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
```

```
16:27:50.980 INFO  VariantRecalibrator - Building SOR x ReadPosRankS
um plot...
16:27:50.981 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:51.560 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
16:27:51.898 INFO  VariantRecalibrator - Building FS x MQRankSum plo
t...
16:27:51.899 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3721 variants...
16:27:52.499 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3721 variants...
16:27:52.844 INFO  VariantRecalibrator - Building FS x ReadPosRankSu
m plot...
16:27:52.844 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:53.435 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:53.775 INFO  VariantRecalibrator - Building MQRankSum x ReadPo
sRankSum plot...
16:27:53.776 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:54.368 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
16:27:54.819 INFO  VariantRecalibrator - Executing: Rscript /opt/dat
a/accreditation/test/NextGenVariantCalling_set5/Code/../Results/VQS
R/setr5_out_SNP.plots.R
16:28:33.246 INFO  VariantRecalibrator - Executing: Rscript (resourc
e)org/broadinstitute/hellbender/tools/walkers/vqsr/plot_Tranches.R /
opt/data/accreditation/test/NextGenVariantCalling_set5/Code/../Resul
ts/VQSR/setr5_out_raw_variants_SNP.tranches 2.15
16:28:33.800 INFO  VariantRecalibrator - Shutting down engine
[September 28, 2018 4:28:33 PM EAT] org.broadinstitute.hellbender.to
ols.walkers.vqsr.VariantRecalibrator done. Elapsed time: 1.76 minute
s.
Runtime.totalMemory()=6556221440
Tool returned:
true

real    1m48.985s
user    2m29.300s
sys     0m27.109s
```

2. Apply VQSR

Details about this is available here (https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/org_broadinstitute_hellbender_tools_w

I have just changed the filtering level to 99.5 to meet the standards recommednded by gatk

In [25]:

```
!time gatk ApplyVQSR \
   -R ../Data/Genome/chr1_edited.fa \
   -V ../Results/Variants/gatk/setr5_out_raw_variants_edited.vcf \
   -O ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps.vcf \
   --ts-filter-level 99.5 \
   --tranches-file ../Results/VQSR/setr5_out_raw_variants_SNP.tranches \
   --recal-file ../Results/VQSR/setr5_out_raw_variants_SNP.recal \
   -mode SNP
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar ApplyVQSR -R
../Data/Genome/chr1_edited.fa -V ../Results/Variants/gatk/setr5_out_
raw_variants_edited.vcf -O ../Results/Variants/gatk/setr5_out_raw_va
riants_recalibrated_snps.vcf --ts-filter-level 99.5 --tranches-file
../Results/VQSR/setr5_out_raw_variants_SNP.tranches --recal-file ../
Results/VQSR/setr5_out_raw_variants_SNP.recal -mode SNP
17:01:34.504 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
17:01:34.658 INFO  ApplyVQSR - ------------------------------------
----------------------
17:01:34.658 INFO  ApplyVQSR - The Genome Analysis Toolkit (GATK) v
4.0.8.1
17:01:34.658 INFO  ApplyVQSR - For support and documentation go to h
ttps://software.broadinstitute.org/gatk/
17:01:34.659 INFO  ApplyVQSR - Executing as caleb@hpc01.icipe.org on
Linux v2.6.32-696.30.1.el6.x86_64 amd64
17:01:34.659 INFO  ApplyVQSR - Java runtime: OpenJDK 64-Bit Server V
M v1.8.0_121-b15
17:01:34.659 INFO  ApplyVQSR - Start Date/Time: September 28, 2018
5:01:34 PM EAT
17:01:34.660 INFO  ApplyVQSR - ------------------------------------
----------------------
17:01:34.660 INFO  ApplyVQSR - ------------------------------------
----------------------
17:01:34.660 INFO  ApplyVQSR - HTSJDK Version: 2.16.0
17:01:34.661 INFO  ApplyVQSR - Picard Version: 2.18.7
17:01:34.661 INFO  ApplyVQSR - HTSJDK Defaults.COMPRESSION_LEVEL : 2
17:01:34.661 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR
_SAMTOOLS : false
17:01:34.661 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_SAMTOOLS : true
17:01:34.661 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_TRIBBLE : false
17:01:34.661 INFO  ApplyVQSR - Deflater: IntelDeflater
17:01:34.661 INFO  ApplyVQSR - Inflater: IntelInflater
17:01:34.662 INFO  ApplyVQSR - GCS max retries/reopens: 20
17:01:34.662 INFO  ApplyVQSR - Using google-cloud-java fork https://
github.com/broadinstitute/google-cloud-java/releases/tag/0.20.5-alph
a-GCS-RETRY-FIX
17:01:34.662 INFO  ApplyVQSR - Initializing engine
17:01:35.283 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/VQSR/setr5_out_raw_variants_SNP.recal
17:01:35.313 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/Variants/gatk/setr5_out_raw_variants_edited.vcf
17:01:35.349 INFO  ApplyVQSR - Done initializing engine
17:01:35.353 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=90.00 minVQSLod=21.2578 known=(12736 @ 2.348
7) novel=(1002 @ 2.2961) truthSites(8620 accessible, 7758 called), n
ame=VQSRTrancheSNP0.00to90.00]
17:01:35.353 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=99.00 minVQSLod=4.2506 known=(16940 @ 2.3052)
```

```
novel=(7850 @ 2.2028) truthSites(8620 accessible, 8533 called), name
=VQSRTrancheSNP90.00to99.00]
17:01:35.354 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=99.90 minVQSLod=-1.8894 known=(17659 @ 2.293
2) novel=(8009 @ 2.2074) truthSites(8620 accessible, 8611 called), n
ame=VQSRTrancheSNP99.00to99.90]
17:01:35.355 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=100.00 minVQSLod=-13.3231 known=(18584 @ 2.26
88) novel=(8291 @ 2.1987) truthSites(8620 accessible, 8620 called),
name=VQSRTrancheSNP99.90to100.00]
17:01:35.374 INFO  ApplyVQSR - Keeping all variants in tranche Truth
SensitivityTranche targetTruthSensitivity=99.90 minVQSLod=-1.8894 kn
own=(17659 @ 2.2932) novel=(8009 @ 2.2074) truthSites(8620 accessibl
e, 8611 called), name=VQSRTrancheSNP99.00to99.90]
17:01:35.394 INFO  ProgressMeter - Starting traversal
17:01:35.394 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes    Variants Processed  Variants/Minute
17:01:36.908 INFO  ProgressMeter -          1:247768812
0.0                 28315        1123611.1
17:01:36.908 INFO  ProgressMeter - Traversal complete. Processed 283
15 total variants in 0.0 minutes.
17:01:37.144 INFO  ApplyVQSR - Shutting down engine
[September 28, 2018 5:01:37 PM EAT] org.broadinstitute.hellbender.to
ols.walkers.vqsr.ApplyVQSR done. Elapsed time: 0.04 minutes.
Runtime.totalMemory()=2463105024

real    0m6.016s
user    0m27.874s
sys     0m1.442s
```

Then move on to INDELS

1. Variant callibration

In this step, we sue the vcf file generated from snps recalibration

Most annotations related to mapping quality (MQ) have been removed since there is a conflation with the length of an indel in a read and the degradation in **mapping quality** that is assigned to the read by the aligner.

https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set (https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set)

In [26]:

```
!time gatk VariantRecalibrator \
-R ../Data/Genome/chr1_edited.fa \
-V ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps.vcf \
-O ../Results/VQSR/setr5_out_raw_variants_INDEL.recal \
-resource mills,known=false,training=true,truth=true,prior=12.0:/opt/data/accred
itation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/Mills_and_1000G_gol
d_standard.indels.b37.vcf \
-resource dbsnp,known=true,training=false,truth=false,prior=2.0:/opt/data/accred
itation/test/NextGenVariantCalling_set5/Data/VcfDatabase/b37/dbsnp_138.b37.vcf \
-an QD -an FS -an SOR -an MQRankSum -an ReadPosRankSum \
-mode INDEL \
--max-gaussians 4 \
--tranches-file ../Results/VQSR/setr5_out_raw_variants_INDEL.tranches \
--rscript-file ../Results/VQSR/setr5_out_INDEL.plots.R
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar VariantRecal
ibrator -R ../Data/Genome/chr1_edited.fa -V ../Results/Variants/gat
k/setr5_out_raw_variants_recalibrated_snps.vcf -O ../Results/VQSR/se
tr5_out_raw_variants_INDEL.recal -resource mills,known=false,trainin
g=true,truth=true,prior=12.0:/opt/data/accreditation/test/NextGenVar
iantCalling_set5/Data/VcfDatabase/b37/Mills_and_1000G_gold_standard.
indels.b37.vcf -resource dbsnp,known=true,training=false,truth=fals
e,prior=2.0:/opt/data/accreditation/test/NextGenVariantCalling_set5/
Data/VcfDatabase/b37/dbsnp_138.b37.vcf -an QD -an FS -an SOR -an MQR
ankSum -an ReadPosRankSum -mode INDEL --max-gaussians 4 --tranches-f
ile ../Results/VQSR/setr5_out_raw_variants_INDEL.tranches --rscript-
file ../Results/VQSR/setr5_out_INDEL.plots.R
17:01:48.660 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
17:01:48.790 INFO  VariantRecalibrator - ----------------------------
--------------------------------
17:01:48.790 INFO  VariantRecalibrator - The Genome Analysis Toolkit
(GATK) v4.0.8.1
17:01:48.790 INFO  VariantRecalibrator - For support and documentati
on go to https://software.broadinstitute.org/gatk/
17:01:48.790 INFO  VariantRecalibrator - Executing as caleb@hpc01.ic
ipe.org on Linux v2.6.32-696.30.1.el6.x86_64 amd64
17:01:48.790 INFO  VariantRecalibrator - Java runtime: OpenJDK 64-Bi
t Server VM v1.8.0_121-b15
17:01:48.791 INFO  VariantRecalibrator - Start Date/Time: September
28, 2018 5:01:48 PM EAT
17:01:48.791 INFO  VariantRecalibrator - ----------------------------
--------------------------------
17:01:48.791 INFO  VariantRecalibrator - ----------------------------
--------------------------------
17:01:48.791 INFO  VariantRecalibrator - HTSJDK Version: 2.16.0
17:01:48.792 INFO  VariantRecalibrator - Picard Version: 2.18.7
17:01:48.792 INFO  VariantRecalibrator - HTSJDK Defaults.COMPRESSION
_LEVEL : 2
17:01:48.792 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_READ_FOR_SAMTOOLS : false
17:01:48.792 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_WRITE_FOR_SAMTOOLS : true
17:01:48.792 INFO  VariantRecalibrator - HTSJDK Defaults.USE_ASYNC_I
O_WRITE_FOR_TRIBBLE : false
17:01:48.792 INFO  VariantRecalibrator - Deflater: IntelDeflater
17:01:48.792 INFO  VariantRecalibrator - Inflater: IntelInflater
17:01:48.792 INFO  VariantRecalibrator - GCS max retries/reopens: 20
17:01:48.792 INFO  VariantRecalibrator - Using google-cloud-java for
k https://github.com/broadinstitute/google-cloud-java/releases/tag/
0.20.5-alpha-GCS-RETRY-FIX
17:01:48.792 INFO  VariantRecalibrator - Initializing engine
17:01:49.383 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/Mills_and_1000G_gold_standard.indels.b37.vcf
17:01:49.453 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Dat
a/VcfDatabase/b37/dbsnp_138.b37.vcf
```

```
17:01:49.589 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps.
vcf
17:01:49.622 INFO  VariantRecalibrator - Done initializing engine
17:01:49.625 INFO  TrainingSet - Found mills track:     Known = fals
e       Training = true         Truth = true    Prior = Q12.0
17:01:49.625 INFO  TrainingSet - Found dbsnp track:     Known = true
Training = false        Truth = false   Prior = Q2.0
17:01:49.631 WARN  GATKVariantContextUtils - Can't determine output
variant file format from output file extension "recal". Defaulting t
o VCF.
17:01:49.656 INFO  ProgressMeter - Starting traversal
17:01:49.656 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes    Variants Processed  Variants/Minute
17:01:59.951 INFO  ProgressMeter -          1:230678478
0.2                 26000         151544.6
17:02:00.906 INFO  ProgressMeter -          1:247768812
0.2                 28315         151013.3
17:02:00.906 INFO  ProgressMeter - Traversal complete. Processed 283
15 total variants in 0.2 minutes.
17:02:00.907 INFO  VariantDataManager - QD:      mean = 20.73    sta
ndard deviation = 9.04
17:02:00.908 INFO  VariantDataManager - FS:      mean = 1.11     sta
ndard deviation = 2.63
17:02:00.909 INFO  VariantDataManager - SOR:     mean = 1.72     sta
ndard deviation = 1.41
17:02:00.909 INFO  VariantDataManager - MQRankSum:       mean = -0.0
3        standard deviation = 0.28
17:02:00.910 INFO  VariantDataManager - ReadPosRankSum:          mea
n = 0.06         standard deviation = 0.99
17:02:00.921 INFO  VariantDataManager - Annotations are now ordered
by their information content: [QD, SOR, FS, ReadPosRankSum, MQRankSu
m]
17:02:00.921 INFO  VariantDataManager - Training with 664 variants a
fter standard deviation thresholding.
17:02:00.921 WARN  VariantDataManager - WARNING: Training with very
few variant sites! Please check the model reporting PDF to ensure th
e quality of the model is reliable.
17:02:00.925 INFO  GaussianMixtureModel - Initializing model with 10
0 k-means iterations...
17:02:01.012 INFO  VariantRecalibratorEngine - Finished iteration 0.
17:02:01.046 INFO  VariantRecalibratorEngine - Finished iteration 5.
Current change in mixture coefficients = 0.23175
17:02:01.067 INFO  VariantRecalibratorEngine - Finished iteration 1
0.      Current change in mixture coefficients = 0.07408
17:02:01.087 INFO  VariantRecalibratorEngine - Finished iteration 1
5.      Current change in mixture coefficients = 0.03830
17:02:01.094 INFO  VariantRecalibratorEngine - Convergence after 17
iterations!
17:02:01.105 INFO  VariantRecalibratorEngine - Evaluating full set o
f 1440 variants...
17:02:01.180 INFO  VariantDataManager - Selected worst 24 scoring va
riants --> variants with LOD <= -5.0000.
17:02:01.180 INFO  GaussianMixtureModel - Initializing model with 10
0 k-means iterations...
17:02:01.181 INFO  VariantRecalibratorEngine - Finished iteration 0.
17:02:01.182 INFO  VariantRecalibratorEngine - Finished iteration 5.
Current change in mixture coefficients = 0.10349
17:02:01.183 INFO  VariantRecalibratorEngine - Convergence after 8 i
terations!
```

```
17:02:01.184 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:01.205 INFO  TrancheManager - Finding 4 tranches for 1440 vari
ants
17:02:01.210 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 100.00 => selection metric threshold 0.000
17:02:01.219 INFO  TrancheManager -   Found tranche for 100.000: 0.0
00 threshold starting with variant 0; running score is 0.000
17:02:01.219 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=100.00 minVQSLod=-50355.
8881 known=(970 @ 0.0000) novel=(470 @ 0.0000) truthSites(665 access
ible, 665 called), name=anonymous]
17:02:01.219 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 99.90 => selection metric threshold 0.001
17:02:01.220 INFO  TrancheManager -   Found tranche for 99.900: 0.00
1 threshold starting with variant 3; running score is 0.002
17:02:01.221 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=99.90 minVQSLod=-23665.9
999 known=(967 @ 0.0000) novel=(470 @ 0.0000) truthSites(665 accessi
ble, 664 called), name=anonymous]
17:02:01.221 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 99.00 => selection metric threshold 0.010
17:02:01.221 INFO  TrancheManager -   Found tranche for 99.000: 0.01
0 threshold starting with variant 37; running score is 0.011
17:02:01.222 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=99.00 minVQSLod=-2.9505
known=(948 @ 0.0000) novel=(455 @ 0.0000) truthSites(665 accessible,
658 called), name=anonymous]
17:02:01.222 INFO  TrancheManager -   TruthSensitivityTranche thresh
old 90.00 => selection metric threshold 0.100
17:02:01.222 INFO  TrancheManager -   Found tranche for 90.000: 0.10
0 threshold starting with variant 170; running score is 0.101
17:02:01.223 INFO  TrancheManager -   TruthSensitivityTranche is Tru
thSensitivityTranche targetTruthSensitivity=90.00 minVQSLod=-0.5943
known=(867 @ 0.0000) novel=(403 @ 0.0000) truthSites(665 accessible,
598 called), name=anonymous]
17:02:01.225 INFO  VariantRecalibrator - Writing out recalibration t
able...
17:02:01.330 INFO  VariantRecalibrator - Writing out visualization R
script file...
17:02:01.335 INFO  VariantRecalibrator - Building QD x SOR plot...
17:02:01.340 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
17:02:02.003 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:02.100 INFO  VariantRecalibrator - Building QD x FS plot...
17:02:02.104 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:02.748 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:02.812 INFO  VariantRecalibrator - Building QD x ReadPosRankSu
m plot...
17:02:02.816 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:03.427 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:03.484 INFO  VariantRecalibrator - Building QD x MQRankSum plo
t...
17:02:03.485 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:04.098 WARN  VariantRecalibratorEngine - Model could not pre-c
```

```
ompute denominators.
17:02:04.147 INFO  VariantRecalibrator - Building SOR x FS plot...
17:02:04.147 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
17:02:04.774 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:04.820 INFO  VariantRecalibrator - Building SOR x ReadPosRankS
um plot...
17:02:04.821 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
17:02:05.444 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:05.488 INFO  VariantRecalibrator - Building SOR x MQRankSum pl
ot...
17:02:05.489 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3660 variants...
17:02:06.114 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:06.159 INFO  VariantRecalibrator - Building FS x ReadPosRankSu
m plot...
17:02:06.160 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:06.781 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:06.827 INFO  VariantRecalibrator - Building FS x MQRankSum plo
t...
17:02:06.828 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:07.446 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:07.491 INFO  VariantRecalibrator - Building ReadPosRankSum x M
QRankSum plot...
17:02:07.491 INFO  VariantRecalibratorEngine - Evaluating full set o
f 3600 variants...
17:02:08.106 WARN  VariantRecalibratorEngine - Model could not pre-c
ompute denominators.
17:02:08.299 INFO  VariantRecalibrator - Executing: Rscript /opt/dat
a/accreditation/test/NextGenVariantCalling_set5/Code/../Results/VQS
R/setr5_out_INDEL.plots.R
17:02:34.720 INFO  VariantRecalibrator - Tranches plot will not be g
enerated since we are running in INDEL mode
17:02:34.817 INFO  VariantRecalibrator - Shutting down engine
[September 28, 2018 5:02:34 PM EAT] org.broadinstitute.hellbender.to
ols.walkers.vqsr.VariantRecalibrator done. Elapsed time: 0.77 minute
s.
Runtime.totalMemory()=3656908800
Tool returned:
true

real    0m49.246s
user    1m12.829s
sys     0m9.846s
```

2. Apply VQSR

In [27]:

```
!time gatk ApplyVQSR \
   -R ../Data/Genome/chr1_edited.fa \
   -V ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps.vcf \
   -O ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels.v
cf \
   --ts-filter-level 99.0 \
   --tranches-file ../Results/VQSR/setr5_out_raw_variants_INDEL.tranches \
   --recal-file ../Results/VQSR/setr5_out_raw_variants_INDEL.recal \
   -mode INDEL
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar ApplyVQSR -R
../Data/Genome/chr1_edited.fa -V ../Results/Variants/gatk/setr5_out_
raw_variants_recalibrated_snps.vcf -O ../Results/Variants/gatk/setr5
_out_raw_variants_recalibrated_snps_indels.vcf --ts-filter-level 99.
0 --tranches-file ../Results/VQSR/setr5_out_raw_variants_INDEL.tranc
hes --recal-file ../Results/VQSR/setr5_out_raw_variants_INDEL.recal
-mode INDEL
17:09:46.839 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
17:09:46.982 INFO  ApplyVQSR - ------------------------------------
----------------------
17:09:46.983 INFO  ApplyVQSR - The Genome Analysis Toolkit (GATK) v
4.0.8.1
17:09:46.983 INFO  ApplyVQSR - For support and documentation go to h
ttps://software.broadinstitute.org/gatk/
17:09:46.983 INFO  ApplyVQSR - Executing as caleb@hpc01.icipe.org on
Linux v2.6.32-696.30.1.el6.x86_64 amd64
17:09:46.983 INFO  ApplyVQSR - Java runtime: OpenJDK 64-Bit Server V
M v1.8.0_121-b15
17:09:46.984 INFO  ApplyVQSR - Start Date/Time: September 28, 2018
5:09:46 PM EAT
17:09:46.984 INFO  ApplyVQSR - ------------------------------------
----------------------
17:09:46.984 INFO  ApplyVQSR - ------------------------------------
----------------------
17:09:46.985 INFO  ApplyVQSR - HTSJDK Version: 2.16.0
17:09:46.985 INFO  ApplyVQSR - Picard Version: 2.18.7
17:09:46.985 INFO  ApplyVQSR - HTSJDK Defaults.COMPRESSION_LEVEL : 2
17:09:46.985 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR
_SAMTOOLS : false
17:09:46.985 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_SAMTOOLS : true
17:09:46.985 INFO  ApplyVQSR - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FO
R_TRIBBLE : false
17:09:46.986 INFO  ApplyVQSR - Deflater: IntelDeflater
17:09:46.986 INFO  ApplyVQSR - Inflater: IntelInflater
17:09:46.986 INFO  ApplyVQSR - GCS max retries/reopens: 20
17:09:46.986 INFO  ApplyVQSR - Using google-cloud-java fork https://
github.com/broadinstitute/google-cloud-java/releases/tag/0.20.5-alph
a-GCS-RETRY-FIX
17:09:46.986 INFO  ApplyVQSR - Initializing engine
17:09:47.592 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/VQSR/setr5_out_raw_variants_INDEL.recal
17:09:47.617 INFO  FeatureManager - Using codec VCFCodec to read fil
e file:///opt/data/accreditation/test/NextGenVariantCalling_set5/Cod
e/../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps.
vcf
17:09:47.660 INFO  ApplyVQSR - Done initializing engine
17:09:47.664 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=90.00 minVQSLod=-0.5943 known=(867 @ 0.0000)
novel=(403 @ 0.0000) truthSites(665 accessible, 598 called), name=VQ
SRTrancheINDEL0.00to90.00]
```

```
17:09:47.666 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=99.00 minVQSLod=-2.9505 known=(948 @ 0.0000)
novel=(455 @ 0.0000) truthSites(665 accessible, 658 called), name=VQ
SRTrancheINDEL90.00to99.00]
17:09:47.667 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=99.90 minVQSLod=-23665.9999 known=(967 @ 0.00
00) novel=(470 @ 0.0000) truthSites(665 accessible, 664 called), nam
e=VQSRTrancheINDEL99.00to99.90]
17:09:47.667 INFO  ApplyVQSR - Read tranche TruthSensitivityTranche
targetTruthSensitivity=100.00 minVQSLod=-50355.8881 known=(970 @ 0.0
000) novel=(470 @ 0.0000) truthSites(665 accessible, 665 called), na
me=VQSRTrancheINDEL99.90to100.00]
17:09:47.691 INFO  ApplyVQSR - Keeping all variants in tranche Truth
SensitivityTranche targetTruthSensitivity=99.00 minVQSLod=-2.9505 kn
own=(948 @ 0.0000) novel=(455 @ 0.0000) truthSites(665 accessible, 6
58 called), name=VQSRTrancheINDEL90.00to99.00]
17:09:47.706 INFO  ProgressMeter - Starting traversal
17:09:47.707 INFO  ProgressMeter -        Current Locus  Elapsed Min
utes    Variants Processed  Variants/Minute
17:09:49.033 INFO  ProgressMeter -          1:247768812
0.0                 28315        1281221.7
17:09:49.034 INFO  ProgressMeter - Traversal complete. Processed 283
15 total variants in 0.0 minutes.
17:09:49.257 INFO  ApplyVQSR - Shutting down engine
[September 28, 2018 5:09:49 PM EAT] org.broadinstitute.hellbender.to
ols.walkers.vqsr.ApplyVQSR done. Elapsed time: 0.04 minutes.
Runtime.totalMemory()=2433220608

real    0m5.579s
user    0m25.221s
sys     0m1.455s
```

Using freebayes for variant calling

First we install it:

```
conda install -c bioconda freebayes
```

As stipulated by the SOP, we only include variants with over 30 quality score.

1. call the variaants

In [28]:

```
!time freebayes -f ../Data/Genome/chr1.fa --pooled-continuous -m 30 \
-b ../Results/BQSR/set5_out_aln_dedup_pass1.adjusted.bam \
|vcffilter -f "QUAL > 30" >../Results/Variants/freebayes/set5_out_filtered_varia
nts.vcf
```

```
real    38m12.418s
user    37m56.081s
sys     0m1.960s
```

2. Compress the variants

In [48]:

```
!time bcftools view -O z \
-o ../Results/Variants/freebayes/set5_out_filtered_variants.vcf.gz \
../Results/Variants/freebayes/set5_out_filtered_variants.vcf
```

```
real    0m0.894s
user    0m0.777s
sys     0m0.022s
```

3. Finally, we create an index file for the called, compresssed variants

In [49]:

```
!bcftools index ../Results/Variants/freebayes/set5_out_filtered_variants.vcf.gz
```

In [50]:

```
!bcftools stats ../Results/Variants/freebayes/set5_out_filtered_variants.vcf  \
> ../Results/Variants/freebayes/set5_out_filtered_variants.vcf.stats
```

In [51]:

```
!plot-vcfstats ../Results/Variants/freebayes/set5_out_filtered_variants.vcf.stat
s \
-p ../Results/Variants/freebayes/plots
```

```
Parsing bcftools stats output: ../Results/Variants/freebayes/set5_ou
t_filtered_variants.vcf.stats
Plotting graphs: python plot.py
Creating PDF: pdflatex summary.tex >plot-vcfstats.log 2>&1
Finished: ../Results/Variants/freebayes/plots/summary.pdf
```

Analysis

Now that the pipeline is complete and working, how are the results generated interpreted?

Read this GATK documnent on Variant Evaluation for more information on how to determine if the variants were properly called. The quality of the calls. How do we know that the calls were successful?

Variants that are above the threshold pass the filter, so the FILTER field will contain PASS. Variants that are below the threshold will be filtered out; they will be written to the output file, but in the FILTER field they will have the name of the tranche they belonged to. So VQSRTrancheSNP99.90to100.00 means that the variant was in the range of VQSLODs corresponding to the remaining 0.1% of the truth set, which are considered false positives. Yes, we accept the possibility that some small number of variant calls in the truth set are wrong...

Generate the statistics from the final vcf file

In [19]:

```
!bcftools stats ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snp
s_indels.vcf  \
> ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels.vcf.s
tats
```

In [20]:

```
!plot-vcfstats ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps
_indels.vcf.stats \
-p ../Results/Variants/gatk/recalibrated_plots
```

```
Parsing bcftools stats output: ../Results/Variants/gatk/setr5_out_ra
w_variants_recalibrated_snps_indels.vcf.stats
Plotting graphs: python plot.py
Creating PDF: pdflatex summary.tex >plot-vcfstats.log 2>&1
Finished: ../Results/Variants/gatk/recalibrated_plots/summary.pdf
```

It appear the number of variant and the corresponding sumary have not changed much. This makes sense since, the recalibration or filtering stage does not eliminate any sequence. Rather, the variants which do not meet the set level are marked with the tranche level. "

In [33]:

```
!grep -c 'PASS' ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snp
s_indels.vcf
```

```
27071
```

In [34]:

```
!grep -c 'VQSRTranche' ../Results/Variants/gatk/setr5_out_raw_variants_recalibra
ted_snps_indels.vcf |head
```

```
1249
```

In [34]:

```
!sed 's/^1/chr1/g' ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_
snps_indels.vcf \
> ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels_chr1.
vcf
```

In [53]:

```
!time gatk CollectVariantCallingMetrics \
--INPUT ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels
_chr1.vcf \
--OUTPUT ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indel
s.vcf.metrics \
--DBSNP ../Data/VcfDatabase/hg38/dbsnp_146.hg38.vcf
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar CollectVaria
ntCallingMetrics --INPUT ../Results/Variants/gatk/setr5_out_raw_vari
ants_recalibrated_snps_indels_chr1.vcf --OUTPUT ../Results/Variants/
gatk/setr5_out_raw_variants_recalibrated_snps_indels.vcf.metrics --D
BSNP ../Data/VcfDatabase/hg38/dbsnp_146.hg38.vcf
16:23:37.578 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
[Tue Sep 25 16:23:37 EAT 2018] CollectVariantCallingMetrics  --INPUT
../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_in
dels_chr1.vcf --OUTPUT ../Results/Variants/gatk/setr5_out_raw_varian
ts_recalibrated_snps_indels.vcf.metrics --DBSNP ../Data/VcfDatabase/
hg38/dbsnp_146.hg38.vcf  --GVCF_INPUT false --THREAD_COUNT 1 --VERBO
SITY INFO --QUIET false --VALIDATION_STRINGENCY STRICT --COMPRESSION
_LEVEL 2 --MAX_RECORDS_IN_RAM 500000 --CREATE_INDEX false --CREATE_M
D5_FILE false --GA4GH_CLIENT_SECRETS client_secrets.json --help fals
e --version false --showHidden false --USE_JDK_DEFLATER false --USE_
JDK_INFLATER false
[Tue Sep 25 16:23:37 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is avail
able; Picard version: Version:4.0.8.1
INFO    2018-09-25 16:23:37     CollectVariantCallingMetrics    Load
ing dbSNP file ...
INFO    2018-09-25 16:23:38     CollectVariantCallingMetrics    Read
100,000 variants.  Elapsed time: 00:00:01s.  Time for last 100,000:
1s.  Last read position: chr1:1,935,266
INFO    2018-09-25 16:23:39     CollectVariantCallingMetrics    Read
200,000 variants.  Elapsed time: 00:00:01s.  Time for last 100,000:
0s.  Last read position: chr1:3,412,155
INFO    2018-09-25 16:23:40     CollectVariantCallingMetrics    Read
300,000 variants.  Elapsed time: 00:00:02s.  Time for last 100,000:
0s.  Last read position: chr1:4,990,449
INFO    2018-09-25 16:23:41     CollectVariantCallingMetrics    Read
400,000 variants.  Elapsed time: 00:00:03s.  Time for last 100,000:
0s.  Last read position: chr1:6,597,075
INFO    2018-09-25 16:23:41     CollectVariantCallingMetrics    Read
500,000 variants.  Elapsed time: 00:00:04s.  Time for last 100,000:
0s.  Last read position: chr1:8,405,290
INFO    2018-09-25 16:23:42     CollectVariantCallingMetrics    Read
600,000 variants.  Elapsed time: 00:00:04s.  Time for last 100,000:
0s.  Last read position: chr1:10,249,930
INFO    2018-09-25 16:23:43     CollectVariantCallingMetrics    Read
700,000 variants.  Elapsed time: 00:00:05s.  Time for last 100,000:
0s.  Last read position: chr1:11,916,890
INFO    2018-09-25 16:23:43     CollectVariantCallingMetrics    Read
800,000 variants.  Elapsed time: 00:00:05s.  Time for last 100,000:
0s.  Last read position: chr1:14,005,701
INFO    2018-09-25 16:23:44     CollectVariantCallingMetrics    Read
900,000 variants.  Elapsed time: 00:00:06s.  Time for last 100,000:
0s.  Last read position: chr1:15,774,826
INFO    2018-09-25 16:23:45     CollectVariantCallingMetrics    Read
1,000,000 variants.  Elapsed time: 00:00:07s.  Time for last 100,00
0:    0s.  Last read position: chr1:17,396,821
```

```
INFO    2018-09-25 16:23:45     CollectVariantCallingMetrics    Read
1,100,000 variants.  Elapsed time: 00:00:07s.  Time for last 100,00
0:    0s.  Last read position: chr1:19,148,583
INFO    2018-09-25 16:23:46     CollectVariantCallingMetrics    Read
1,200,000 variants.  Elapsed time: 00:00:08s.  Time for last 100,00
0:    0s.  Last read position: chr1:20,898,315
INFO    2018-09-25 16:23:47     CollectVariantCallingMetrics    Read
1,300,000 variants.  Elapsed time: 00:00:09s.  Time for last 100,00
0:    0s.  Last read position: chr1:22,660,400
INFO    2018-09-25 16:23:47     CollectVariantCallingMetrics    Read
1,400,000 variants.  Elapsed time: 00:00:10s.  Time for last 100,00
0:    0s.  Last read position: chr1:24,571,760
INFO    2018-09-25 16:23:48     CollectVariantCallingMetrics    Read
1,500,000 variants.  Elapsed time: 00:00:10s.  Time for last 100,00
0:    0s.  Last read position: chr1:26,432,265
INFO    2018-09-25 16:23:49     CollectVariantCallingMetrics    Read
1,600,000 variants.  Elapsed time: 00:00:11s.  Time for last 100,00
0:    0s.  Last read position: chr1:28,359,437
INFO    2018-09-25 16:23:49     CollectVariantCallingMetrics    Read
1,700,000 variants.  Elapsed time: 00:00:12s.  Time for last 100,00
0:    0s.  Last read position: chr1:30,286,054
INFO    2018-09-25 16:23:50     CollectVariantCallingMetrics    Read
1,800,000 variants.  Elapsed time: 00:00:12s.  Time for last 100,00
0:    0s.  Last read position: chr1:32,106,354
INFO    2018-09-25 16:23:51     CollectVariantCallingMetrics    Read
1,900,000 variants.  Elapsed time: 00:00:13s.  Time for last 100,00
0:    0s.  Last read position: chr1:33,967,034
INFO    2018-09-25 16:23:51     CollectVariantCallingMetrics    Read
2,000,000 variants.  Elapsed time: 00:00:14s.  Time for last 100,00
0:    0s.  Last read position: chr1:35,902,091
INFO    2018-09-25 16:23:52     CollectVariantCallingMetrics    Read
2,100,000 variants.  Elapsed time: 00:00:14s.  Time for last 100,00
0:    0s.  Last read position: chr1:37,733,317
INFO    2018-09-25 16:23:53     CollectVariantCallingMetrics    Read
2,200,000 variants.  Elapsed time: 00:00:15s.  Time for last 100,00
0:    0s.  Last read position: chr1:39,657,222
INFO    2018-09-25 16:23:53     CollectVariantCallingMetrics    Read
2,300,000 variants.  Elapsed time: 00:00:15s.  Time for last 100,00
0:    0s.  Last read position: chr1:41,559,094
INFO    2018-09-25 16:23:54     CollectVariantCallingMetrics    Read
2,400,000 variants.  Elapsed time: 00:00:16s.  Time for last 100,00
0:    0s.  Last read position: chr1:43,452,963
INFO    2018-09-25 16:23:55     CollectVariantCallingMetrics    Read
2,500,000 variants.  Elapsed time: 00:00:17s.  Time for last 100,00
0:    0s.  Last read position: chr1:45,338,272
INFO    2018-09-25 16:23:55     CollectVariantCallingMetrics    Read
2,600,000 variants.  Elapsed time: 00:00:17s.  Time for last 100,00
0:    0s.  Last read position: chr1:47,217,763
INFO    2018-09-25 16:23:56     CollectVariantCallingMetrics    Read
2,700,000 variants.  Elapsed time: 00:00:18s.  Time for last 100,00
0:    0s.  Last read position: chr1:49,359,156
INFO    2018-09-25 16:23:57     CollectVariantCallingMetrics    Read
2,800,000 variants.  Elapsed time: 00:00:19s.  Time for last 100,00
0:    0s.  Last read position: chr1:51,599,947
INFO    2018-09-25 16:23:57     CollectVariantCallingMetrics    Read
2,900,000 variants.  Elapsed time: 00:00:19s.  Time for last 100,00
0:    0s.  Last read position: chr1:53,509,254
INFO    2018-09-25 16:23:58     CollectVariantCallingMetrics    Read
3,000,000 variants.  Elapsed time: 00:00:20s.  Time for last 100,00
0:    0s.  Last read position: chr1:55,309,861
INFO    2018-09-25 16:23:59     CollectVariantCallingMetrics    Read
```

```
3,100,000 variants.  Elapsed time: 00:00:21s.  Time for last 100,00
0:    0s.  Last read position: chr1:57,260,075
INFO    2018-09-25 16:23:59     CollectVariantCallingMetrics    Read
3,200,000 variants.  Elapsed time: 00:00:21s.  Time for last 100,00
0:    0s.  Last read position: chr1:59,297,816
INFO    2018-09-25 16:24:00     CollectVariantCallingMetrics    Read
3,300,000 variants.  Elapsed time: 00:00:22s.  Time for last 100,00
0:    0s.  Last read position: chr1:61,334,556
INFO    2018-09-25 16:24:01     CollectVariantCallingMetrics    Read
3,400,000 variants.  Elapsed time: 00:00:23s.  Time for last 100,00
0:    0s.  Last read position: chr1:63,314,732
INFO    2018-09-25 16:24:01     CollectVariantCallingMetrics    Read
3,500,000 variants.  Elapsed time: 00:00:23s.  Time for last 100,00
0:    0s.  Last read position: chr1:65,376,771
INFO    2018-09-25 16:24:02     CollectVariantCallingMetrics    Read
3,600,000 variants.  Elapsed time: 00:00:24s.  Time for last 100,00
0:    0s.  Last read position: chr1:67,374,491
INFO    2018-09-25 16:24:03     CollectVariantCallingMetrics    Read
3,700,000 variants.  Elapsed time: 00:00:25s.  Time for last 100,00
0:    0s.  Last read position: chr1:69,396,703
INFO    2018-09-25 16:24:03     CollectVariantCallingMetrics    Read
3,800,000 variants.  Elapsed time: 00:00:25s.  Time for last 100,00
0:    0s.  Last read position: chr1:71,508,361
INFO    2018-09-25 16:24:04     CollectVariantCallingMetrics    Read
3,900,000 variants.  Elapsed time: 00:00:26s.  Time for last 100,00
0:    0s.  Last read position: chr1:73,651,283
INFO    2018-09-25 16:24:05     CollectVariantCallingMetrics    Read
4,000,000 variants.  Elapsed time: 00:00:27s.  Time for last 100,00
0:    0s.  Last read position: chr1:75,668,429
INFO    2018-09-25 16:24:05     CollectVariantCallingMetrics    Read
4,100,000 variants.  Elapsed time: 00:00:27s.  Time for last 100,00
0:    0s.  Last read position: chr1:77,715,766
INFO    2018-09-25 16:24:06     CollectVariantCallingMetrics    Read
4,200,000 variants.  Elapsed time: 00:00:28s.  Time for last 100,00
0:    0s.  Last read position: chr1:79,694,131
INFO    2018-09-25 16:24:07     CollectVariantCallingMetrics    Read
4,300,000 variants.  Elapsed time: 00:00:29s.  Time for last 100,00
0:    0s.  Last read position: chr1:81,623,926
INFO    2018-09-25 16:24:07     CollectVariantCallingMetrics    Read
4,400,000 variants.  Elapsed time: 00:00:29s.  Time for last 100,00
0:    0s.  Last read position: chr1:83,709,854
INFO    2018-09-25 16:24:08     CollectVariantCallingMetrics    Read
4,500,000 variants.  Elapsed time: 00:00:30s.  Time for last 100,00
0:    0s.  Last read position: chr1:85,755,741
INFO    2018-09-25 16:24:09     CollectVariantCallingMetrics    Read
4,600,000 variants.  Elapsed time: 00:00:31s.  Time for last 100,00
0:    0s.  Last read position: chr1:87,780,391
INFO    2018-09-25 16:24:09     CollectVariantCallingMetrics    Read
4,700,000 variants.  Elapsed time: 00:00:31s.  Time for last 100,00
0:    0s.  Last read position: chr1:89,837,664
INFO    2018-09-25 16:24:10     CollectVariantCallingMetrics    Read
4,800,000 variants.  Elapsed time: 00:00:32s.  Time for last 100,00
0:    0s.  Last read position: chr1:91,900,960
INFO    2018-09-25 16:24:11     CollectVariantCallingMetrics    Read
4,900,000 variants.  Elapsed time: 00:00:33s.  Time for last 100,00
0:    0s.  Last read position: chr1:93,986,940
INFO    2018-09-25 16:24:11     CollectVariantCallingMetrics    Read
5,000,000 variants.  Elapsed time: 00:00:33s.  Time for last 100,00
0:    0s.  Last read position: chr1:95,992,452
INFO    2018-09-25 16:24:12     CollectVariantCallingMetrics    Read
5,100,000 variants.  Elapsed time: 00:00:34s.  Time for last 100,00
```

```
0:    0s.  Last read position: chr1:98,080,046
INFO    2018-09-25 16:24:13     CollectVariantCallingMetrics    Read
5,200,000 variants.  Elapsed time: 00:00:35s.  Time for last 100,00
0:    0s.  Last read position: chr1:100,097,444
INFO    2018-09-25 16:24:13     CollectVariantCallingMetrics    Read
5,300,000 variants.  Elapsed time: 00:00:35s.  Time for last 100,00
0:    0s.  Last read position: chr1:102,050,317
INFO    2018-09-25 16:24:14     CollectVariantCallingMetrics    Read
5,400,000 variants.  Elapsed time: 00:00:36s.  Time for last 100,00
0:    0s.  Last read position: chr1:104,196,501
INFO    2018-09-25 16:24:14     CollectVariantCallingMetrics    Read
5,500,000 variants.  Elapsed time: 00:00:37s.  Time for last 100,00
0:    0s.  Last read position: chr1:106,018,964
INFO    2018-09-25 16:24:15     CollectVariantCallingMetrics    Read
5,600,000 variants.  Elapsed time: 00:00:37s.  Time for last 100,00
0:    0s.  Last read position: chr1:108,016,132
INFO    2018-09-25 16:24:16     CollectVariantCallingMetrics    Read
5,700,000 variants.  Elapsed time: 00:00:38s.  Time for last 100,00
0:    0s.  Last read position: chr1:110,018,451
INFO    2018-09-25 16:24:16     CollectVariantCallingMetrics    Read
5,800,000 variants.  Elapsed time: 00:00:39s.  Time for last 100,00
0:    0s.  Last read position: chr1:111,915,373
INFO    2018-09-25 16:24:17     CollectVariantCallingMetrics    Read
5,900,000 variants.  Elapsed time: 00:00:39s.  Time for last 100,00
0:    0s.  Last read position: chr1:113,909,958
INFO    2018-09-25 16:24:18     CollectVariantCallingMetrics    Read
6,000,000 variants.  Elapsed time: 00:00:40s.  Time for last 100,00
0:    0s.  Last read position: chr1:115,849,432
INFO    2018-09-25 16:24:18     CollectVariantCallingMetrics    Read
6,100,000 variants.  Elapsed time: 00:00:41s.  Time for last 100,00
0:    0s.  Last read position: chr1:117,813,366
INFO    2018-09-25 16:24:19     CollectVariantCallingMetrics    Read
6,200,000 variants.  Elapsed time: 00:00:41s.  Time for last 100,00
0:    0s.  Last read position: chr1:119,749,517
INFO    2018-09-25 16:24:20     CollectVariantCallingMetrics    Read
6,300,000 variants.  Elapsed time: 00:00:42s.  Time for last 100,00
0:    0s.  Last read position: chr1:144,817,323
INFO    2018-09-25 16:24:20     CollectVariantCallingMetrics    Read
6,400,000 variants.  Elapsed time: 00:00:42s.  Time for last 100,00
0:    0s.  Last read position: chr1:148,487,539
INFO    2018-09-25 16:24:21     CollectVariantCallingMetrics    Read
6,500,000 variants.  Elapsed time: 00:00:43s.  Time for last 100,00
0:    0s.  Last read position: chr1:151,193,960
INFO    2018-09-25 16:24:22     CollectVariantCallingMetrics    Read
6,600,000 variants.  Elapsed time: 00:00:44s.  Time for last 100,00
0:    0s.  Last read position: chr1:152,911,770
INFO    2018-09-25 16:24:22     CollectVariantCallingMetrics    Read
6,700,000 variants.  Elapsed time: 00:00:44s.  Time for last 100,00
0:    0s.  Last read position: chr1:154,641,503
INFO    2018-09-25 16:24:23     CollectVariantCallingMetrics    Read
6,800,000 variants.  Elapsed time: 00:00:45s.  Time for last 100,00
0:    0s.  Last read position: chr1:156,371,300
INFO    2018-09-25 16:24:23     CollectVariantCallingMetrics    Read
6,900,000 variants.  Elapsed time: 00:00:46s.  Time for last 100,00
0:    0s.  Last read position: chr1:158,105,930
INFO    2018-09-25 16:24:24     CollectVariantCallingMetrics    Read
7,000,000 variants.  Elapsed time: 00:00:46s.  Time for last 100,00
0:    0s.  Last read position: chr1:159,871,974
INFO    2018-09-25 16:24:25     CollectVariantCallingMetrics    Read
7,100,000 variants.  Elapsed time: 00:00:47s.  Time for last 100,00
0:    0s.  Last read position: chr1:161,591,906
```

```
INFO    2018-09-25 16:24:25     CollectVariantCallingMetrics    Read
7,200,000 variants.  Elapsed time: 00:00:48s.  Time for last 100,00
0:    0s.  Last read position: chr1:163,552,339
INFO    2018-09-25 16:24:26     CollectVariantCallingMetrics    Read
7,300,000 variants.  Elapsed time: 00:00:48s.  Time for last 100,00
0:    0s.  Last read position: chr1:165,541,529
INFO    2018-09-25 16:24:27     CollectVariantCallingMetrics    Read
7,400,000 variants.  Elapsed time: 00:00:49s.  Time for last 100,00
0:    0s.  Last read position: chr1:167,481,174
INFO    2018-09-25 16:24:27     CollectVariantCallingMetrics    Read
7,500,000 variants.  Elapsed time: 00:00:50s.  Time for last 100,00
0:    0s.  Last read position: chr1:169,398,403
INFO    2018-09-25 16:24:28     CollectVariantCallingMetrics    Read
7,600,000 variants.  Elapsed time: 00:00:50s.  Time for last 100,00
0:    0s.  Last read position: chr1:171,347,576
INFO    2018-09-25 16:24:29     CollectVariantCallingMetrics    Read
7,700,000 variants.  Elapsed time: 00:00:51s.  Time for last 100,00
0:    0s.  Last read position: chr1:173,439,018
INFO    2018-09-25 16:24:29     CollectVariantCallingMetrics    Read
7,800,000 variants.  Elapsed time: 00:00:52s.  Time for last 100,00
0:    0s.  Last read position: chr1:175,475,850
INFO    2018-09-25 16:24:30     CollectVariantCallingMetrics    Read
7,900,000 variants.  Elapsed time: 00:00:52s.  Time for last 100,00
0:    0s.  Last read position: chr1:177,526,132
INFO    2018-09-25 16:24:31     CollectVariantCallingMetrics    Read
8,000,000 variants.  Elapsed time: 00:00:53s.  Time for last 100,00
0:    0s.  Last read position: chr1:179,525,166
INFO    2018-09-25 16:24:31     CollectVariantCallingMetrics    Read
8,100,000 variants.  Elapsed time: 00:00:54s.  Time for last 100,00
0:    0s.  Last read position: chr1:181,481,764
INFO    2018-09-25 16:24:32     CollectVariantCallingMetrics    Read
8,200,000 variants.  Elapsed time: 00:00:54s.  Time for last 100,00
0:    0s.  Last read position: chr1:183,422,017
INFO    2018-09-25 16:24:33     CollectVariantCallingMetrics    Read
8,300,000 variants.  Elapsed time: 00:00:55s.  Time for last 100,00
0:    0s.  Last read position: chr1:185,473,446
INFO    2018-09-25 16:24:33     CollectVariantCallingMetrics    Read
8,400,000 variants.  Elapsed time: 00:00:56s.  Time for last 100,00
0:    0s.  Last read position: chr1:187,405,241
INFO    2018-09-25 16:24:34     CollectVariantCallingMetrics    Read
8,500,000 variants.  Elapsed time: 00:00:56s.  Time for last 100,00
0:    0s.  Last read position: chr1:189,366,459
INFO    2018-09-25 16:24:35     CollectVariantCallingMetrics    Read
8,600,000 variants.  Elapsed time: 00:00:57s.  Time for last 100,00
0:    0s.  Last read position: chr1:191,337,294
INFO    2018-09-25 16:24:35     CollectVariantCallingMetrics    Read
8,700,000 variants.  Elapsed time: 00:00:57s.  Time for last 100,00
0:    0s.  Last read position: chr1:193,426,343
INFO    2018-09-25 16:24:36     CollectVariantCallingMetrics    Read
8,800,000 variants.  Elapsed time: 00:00:58s.  Time for last 100,00
0:    0s.  Last read position: chr1:195,377,882
INFO    2018-09-25 16:24:36     CollectVariantCallingMetrics    Read
8,900,000 variants.  Elapsed time: 00:00:59s.  Time for last 100,00
0:    0s.  Last read position: chr1:197,294,716
INFO    2018-09-25 16:24:37     CollectVariantCallingMetrics    Read
9,000,000 variants.  Elapsed time: 00:00:59s.  Time for last 100,00
0:    0s.  Last read position: chr1:199,418,411
INFO    2018-09-25 16:24:38     CollectVariantCallingMetrics    Read
9,100,000 variants.  Elapsed time: 00:01:00s.  Time for last 100,00
0:    0s.  Last read position: chr1:201,320,533
INFO    2018-09-25 16:24:38     CollectVariantCallingMetrics    Read
```

```
9,200,000 variants.  Elapsed time: 00:01:01s.  Time for last 100,00
0:    0s.  Last read position: chr1:203,161,570
INFO    2018-09-25 16:24:39     CollectVariantCallingMetrics    Read
9,300,000 variants.  Elapsed time: 00:01:01s.  Time for last 100,00
0:    0s.  Last read position: chr1:204,981,769
INFO    2018-09-25 16:24:40     CollectVariantCallingMetrics    Read
9,400,000 variants.  Elapsed time: 00:01:02s.  Time for last 100,00
0:    0s.  Last read position: chr1:207,031,097
INFO    2018-09-25 16:24:40     CollectVariantCallingMetrics    Read
9,500,000 variants.  Elapsed time: 00:01:03s.  Time for last 100,00
0:    0s.  Last read position: chr1:209,008,495
INFO    2018-09-25 16:24:41     CollectVariantCallingMetrics    Read
9,600,000 variants.  Elapsed time: 00:01:03s.  Time for last 100,00
0:    0s.  Last read position: chr1:210,956,198
INFO    2018-09-25 16:24:42     CollectVariantCallingMetrics    Read
9,700,000 variants.  Elapsed time: 00:01:04s.  Time for last 100,00
0:    0s.  Last read position: chr1:212,927,390
INFO    2018-09-25 16:24:42     CollectVariantCallingMetrics    Read
9,800,000 variants.  Elapsed time: 00:01:05s.  Time for last 100,00
0:    0s.  Last read position: chr1:214,929,218
INFO    2018-09-25 16:24:43     CollectVariantCallingMetrics    Read
9,900,000 variants.  Elapsed time: 00:01:05s.  Time for last 100,00
0:    0s.  Last read position: chr1:216,899,509
INFO    2018-09-25 16:24:44     CollectVariantCallingMetrics    Read
10,000,000 variants.  Elapsed time: 00:01:06s.  Time for last 100,00
0:    0s.  Last read position: chr1:218,868,907
INFO    2018-09-25 16:24:44     CollectVariantCallingMetrics    Read
10,100,000 variants.  Elapsed time: 00:01:07s.  Time for last 100,00
0:    0s.  Last read position: chr1:220,882,246
INFO    2018-09-25 16:24:45     CollectVariantCallingMetrics    Read
10,200,000 variants.  Elapsed time: 00:01:07s.  Time for last 100,00
0:    0s.  Last read position: chr1:222,946,827
INFO    2018-09-25 16:24:46     CollectVariantCallingMetrics    Read
10,300,000 variants.  Elapsed time: 00:01:08s.  Time for last 100,00
0:    0s.  Last read position: chr1:224,996,084
INFO    2018-09-25 16:24:46     CollectVariantCallingMetrics    Read
10,400,000 variants.  Elapsed time: 00:01:09s.  Time for last 100,00
0:    0s.  Last read position: chr1:226,957,402
INFO    2018-09-25 16:24:47     CollectVariantCallingMetrics    Read
10,500,000 variants.  Elapsed time: 00:01:09s.  Time for last 100,00
0:    0s.  Last read position: chr1:228,823,452
INFO    2018-09-25 16:24:48     CollectVariantCallingMetrics    Read
10,600,000 variants.  Elapsed time: 00:01:10s.  Time for last 100,00
0:    0s.  Last read position: chr1:230,670,951
INFO    2018-09-25 16:24:48     CollectVariantCallingMetrics    Read
10,700,000 variants.  Elapsed time: 00:01:10s.  Time for last 100,00
0:    0s.  Last read position: chr1:232,541,298
INFO    2018-09-25 16:24:49     CollectVariantCallingMetrics    Read
10,800,000 variants.  Elapsed time: 00:01:11s.  Time for last 100,00
0:    0s.  Last read position: chr1:234,450,417
INFO    2018-09-25 16:24:50     CollectVariantCallingMetrics    Read
10,900,000 variants.  Elapsed time: 00:01:12s.  Time for last 100,00
0:    0s.  Last read position: chr1:236,316,897
INFO    2018-09-25 16:24:50     CollectVariantCallingMetrics    Read
11,000,000 variants.  Elapsed time: 00:01:12s.  Time for last 100,00
0:    0s.  Last read position: chr1:238,098,222
INFO    2018-09-25 16:24:51     CollectVariantCallingMetrics    Read
11,100,000 variants.  Elapsed time: 00:01:13s.  Time for last 100,00
0:    0s.  Last read position: chr1:240,050,624
INFO    2018-09-25 16:24:52     CollectVariantCallingMetrics    Read
11,200,000 variants.  Elapsed time: 00:01:14s.  Time for last 100,00
```

```
0:    0s.  Last read position: chr1:241,904,845
INFO    2018-09-25 16:24:52     CollectVariantCallingMetrics    Read
11,300,000 variants.  Elapsed time: 00:01:14s.  Time for last 100,00
0:    0s.  Last read position: chr1:243,913,701
INFO    2018-09-25 16:24:53     CollectVariantCallingMetrics    Read
11,400,000 variants.  Elapsed time: 00:01:15s.  Time for last 100,00
0:    0s.  Last read position: chr1:245,728,332
INFO    2018-09-25 16:24:54     CollectVariantCallingMetrics    Read
11,500,000 variants.  Elapsed time: 00:01:16s.  Time for last 100,00
0:    0s.  Last read position: chr1:247,480,298
[Tue Sep 25 16:24:54 EAT 2018] picard.vcf.CollectVariantCallingMetri
cs done. Elapsed time: 1.28 minutes.
Runtime.totalMemory()=4955045888
To get help, see http://broadinstitute.github.io/picard/index.html#G
ettingHelp
java.lang.NullPointerException
        at picard.util.DbSnpBitSetUtil.loadVcf(DbSnpBitSetUtil.java:
163)
        at picard.util.DbSnpBitSetUtil.createSnpAndIndelBitSets(DbSn
pBitSetUtil.java:131)
        at picard.vcf.CollectVariantCallingMetrics.doWork(CollectVar
iantCallingMetrics.java:105)
        at picard.cmdline.CommandLineProgram.instanceMain(CommandLin
eProgram.java:282)
        at org.broadinstitute.hellbender.cmdline.PicardCommandLinePr
ogramExecutor.instanceMain(PicardCommandLineProgramExecutor.java:25)
        at org.broadinstitute.hellbender.Main.runCommandLineProgram
(Main.java:160)
        at org.broadinstitute.hellbender.Main.mainEntry(Main.java:20
3)
        at org.broadinstitute.hellbender.Main.main(Main.java:289)

real    1m19.930s
user    1m36.108s
sys     0m6.320s
```

In [35]:

```
!gatk3 -T VariantEval \
-R ../Data/Genome/chr1_edited.fa \
-eval ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels.v
cf \
--dbsnp ../Data/VcfDatabase/b37/dbsnp_138.b37.vcf \
--known_names ../Data/VcfDatabase/b37/hapmap_3.3.b37.vcf \
--known_names ../Data/VcfDatabase/b37/1000G_omni2.5.b37.vcf \
-noEV -EV CompOverlap -EV IndelSummary -EV TiTvVariantEvaluator -EV CountVariant
s -EV MultiallelicSummary \
-o ../Results/Variants/SampleVariants_Evaluation_dbsnp_hapmap_omni2.eval.grp
```

```
INFO  21:46:47,399 HelpFormatter - --------------------------------
-----------------------------------------------
INFO  21:46:47,402 HelpFormatter - The Genome Analysis Toolkit (GAT
K) v3.8-0-ge9d806836, Compiled 2017/07/28 21:26:50
INFO  21:46:47,402 HelpFormatter - Copyright (c) 2010-2016 The Broad
Institute
INFO  21:46:47,402 HelpFormatter - For support and documentation go
to https://software.broadinstitute.org/gatk
INFO  21:46:47,403 HelpFormatter - [Fri Sep 28 21:46:47 EAT 2018] Ex
ecuting on Linux 2.6.32-696.30.1.el6.x86_64 amd64
INFO  21:46:47,403 HelpFormatter - OpenJDK 64-Bit Server VM 1.8.0_12
1-b15
INFO  21:46:47,407 HelpFormatter - Program Args: -T VariantEval -R
../Data/Genome/chr1_edited.fa -eval ../Results/Variants/gatk/setr5_o
ut_raw_variants_recalibrated_snps_indels.vcf --dbsnp ../Data/VcfData
base/b37/dbsnp_138.b37.vcf --known_names ../Data/VcfDatabase/b37/hap
map_3.3.b37.vcf --known_names ../Data/VcfDatabase/b37/1000G_omni2.5.
b37.vcf -noEV -EV CompOverlap -EV IndelSummary -EV TiTvVariantEvalua
tor -EV CountVariants -EV MultiallelicSummary -o ../Results/Variant
s/SampleVariants_Evaluation_dbsnp_hapmap_omni2.eval.grp
INFO  21:46:47,425 HelpFormatter - Executing as caleb@hpc01.icipe.or
g on Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server V
M 1.8.0_121-b15.
INFO  21:46:47,425 HelpFormatter - Date/Time: 2018/09/28 21:46:47
INFO  21:46:47,426 HelpFormatter - --------------------------------
-----------------------------------------------
INFO  21:46:47,426 HelpFormatter - --------------------------------
-----------------------------------------------
ERROR StatusLogger Unable to create class org.apache.logging.log4j.c
ore.impl.Log4jContextFactory specified in jar:file:/home/caleb/minic
onda3/envs/icipe-env/opt/gatk-3.8/GenomeAnalysisTK.jar!/META-INF/log
4j-provider.properties
ERROR StatusLogger Log4j2 could not find a logging implementation. P
lease add log4j-core to the classpath. Using SimpleLogger to log to
the console...
INFO  21:46:47,591 GenomeAnalysisEngine - Deflater: IntelDeflater
INFO  21:46:47,591 GenomeAnalysisEngine - Inflater: IntelInflater
INFO  21:46:47,592 GenomeAnalysisEngine - Strictness is SILENT
INFO  21:46:47,687 GenomeAnalysisEngine - Downsampling Settings: Met
hod: BY_SAMPLE, Target Coverage: 1000
INFO  21:46:47,974 GenomeAnalysisEngine - Preparing for traversal
INFO  21:46:47,978 GenomeAnalysisEngine - Done preparing for travers
al
INFO  21:46:47,979 ProgressMeter - [INITIALIZATION COMPLETE; STARTIN
G PROCESSING]
INFO  21:46:47,979 ProgressMeter -                 | processed |
time |    per 1M |           |   total | remaining
INFO  21:46:47,980 ProgressMeter -        Location |     sites | ela
psed |     sites | completed | runtime |   runtime
INFO  21:46:47,998 VariantEval - Creating 3 combinatorial stratifica
tion states
INFO  21:47:17,983 ProgressMeter -     1:150268323   2929469.0    3
0.0 s      10.0 s       60.3%    49.0 s      19.0 s
INFO  21:47:41,228 VariantEval - Finalizing variant report
INFO  21:47:41,576 ProgressMeter -            done   5284204.0    5
3.0 s      10.0 s      100.0%    53.0 s       0.0 s
INFO  21:47:41,577 ProgressMeter - Total runtime 53.60 secs, 0.89 mi
n, 0.01 hours
------------------------------------------------------------------
----------------------
Done. There were no warn messages.
```

```
------------------------------------------------------------------------
----------------------
```

Comparing the output of the two varinat callers

In [ ]:

```
!vcftools --vcf --diff
```

In [8]:

```
!time vcftools --vcf ../Results/Variants/gatk/setr5_out_raw_variants_recalibrate
d_snps_indels_chr1.vcf \
--diff ../Results/Variants/freebayes/set5_out_raw_variants.vcf \
--diff-site --out ../Results/Variants/gatk/setr5_out_raw_variants.diff
```

```
VCFtools - 0.1.16
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
        --vcf ../Results/Variants/gatk/setr5_out_raw_variants_recali
brated_snps_indels_chr1.vcf
        --out ../Results/Variants/gatk/setr5_out_raw_variants.diff
        --diff ../Results/Variants/freebayes/set5_out_raw_variants.v
cf
        --diff-site

Warning: Expected at least 2 parts in FORMAT entry: ID=PL,Number=G,T
ype=Integer,Description="Normalized, Phred-scaled likelihoods for ge
notypes as defined in the VCF specification">
Warning: Expected at least 2 parts in INFO entry: ID=AC,Number=A,Typ
e=Integer,Description="Allele count in genotypes, for each ALT allel
e, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AC,Number=A,Typ
e=Integer,Description="Allele count in genotypes, for each ALT allel
e, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Allele Frequency, for each ALT allele, in the s
ame order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Allele Frequency, for each ALT allele, in the s
ame order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAC,Number=A,
Type=Integer,Description="Maximum likelihood expectation (MLE) for t
he allele counts (not necessarily the same as the AC), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAC,Number=A,
Type=Integer,Description="Maximum likelihood expectation (MLE) for t
he allele counts (not necessarily the same as the AC), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAF,Number=A,
Type=Float,Description="Maximum likelihood expectation (MLE) for the
allele frequency (not necessarily the same as the AF), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAF,Number=A,
Type=Float,Description="Maximum likelihood expectation (MLE) for the
allele frequency (not necessarily the same as the AF), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=culprit,Number=
1,Type=String,Description="The annotation which was the worst perfor
ming in the Gaussian mixture model, likely the reason why the varian
t was filtered out">
After filtering, kept 1 out of 1 Individuals
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Estimated allele frequency in the range (0,1]">
Warning: Expected at least 2 parts in INFO entry: ID=PRO,Number=1,Ty
pe=Float,Description="Reference allele observation count, with parti
al observations recorded fractionally">
Warning: Expected at least 2 parts in INFO entry: ID=PAO,Number=A,Ty
pe=Float,Description="Alternate allele observations, with partial ob
servations recorded fractionally">
Warning: Expected at least 2 parts in INFO entry: ID=SRP,Number=1,Ty
pe=Float,Description="Strand balance probability for the reference a
llele: Phred-scaled upper-bounds estimate of the probability of obse
rving the deviation between SRF and SRR given E(SRF/SRR) ~ 0.5, deri
ved using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=SAP,Number=A,Ty
```

```
pe=Float,Description="Strand balance probability for the alternate a
llele: Phred-scaled upper-bounds estimate of the probability of obse
rving the deviation between SAF and SAR given E(SAF/SAR) ~ 0.5, deri
ved using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=AB,Number=A,Typ
e=Float,Description="Allele balance at heterozygous sites: a number 
between 0 and 1 representing the ratio of reads showing the referenc
e allele to all reads, considering only reads from individuals calle
d as heterozygous">
Warning: Expected at least 2 parts in INFO entry: ID=ABP,Number=A,Ty
pe=Float,Description="Allele balance probability at heterozygous sit
es: Phred-scaled upper-bounds estimate of the probability of observi
ng the deviation between ABR and ABA given E(ABR/ABA) ~ 0.5, derived
using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=RPP,Number=A,Ty
pe=Float,Description="Read Placement Probability: Phred-scaled upper
-bounds estimate of the probability of observing the deviation betwe
en RPL and RPR given E(RPL/RPR) ~ 0.5, derived using Hoeffding's ine
quality">
Warning: Expected at least 2 parts in INFO entry: ID=RPPR,Number=1,T
ype=Float,Description="Read Placement Probability for reference obse
rvations: Phred-scaled upper-bounds estimate of the probability of o
bserving the deviation between RPL and RPR given E(RPL/RPR) ~ 0.5, d
erived using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=EPP,Number=A,Ty
pe=Float,Description="End Placement Probability: Phred-scaled upper-
bounds estimate of the probability of observing the deviation betwee
n EL and ER given E(EL/ER) ~ 0.5, derived using Hoeffding's inequali
ty">
Warning: Expected at least 2 parts in INFO entry: ID=EPPR,Number=1,T
ype=Float,Description="End Placement Probability for reference obser
vations: Phred-scaled upper-bounds estimate of the probability of ob
serving the deviation between EL and ER given E(EL/ER) ~ 0.5, derive
d using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=CIGAR,Number=A,
Type=String,Description="The extended CIGAR representation of each a
lternate allele, with the exception that '=' is replaced by 'M' to e
ase VCF parsing.  Note that INDEL alleles do not have the first matc
hed base (which is provided by default, per the spec) referred to by
the CIGAR.">
Warning: Expected at least 2 parts in FORMAT entry: ID=GQ,Number=1,T
ype=Float,Description="Genotype Quality, the Phred-scaled marginal
(or unconditional) probability of the called genotype">
Warning: Expected at least 2 parts in FORMAT entry: ID=GL,Number=G,T
ype=Float,Description="Genotype Likelihood, log10-scaled likelihoods
of the data given the called genotype for each possible genotype gen
```

```
erated from the reference and alternate alleles given the sample plo
idy">
Comparing sites in VCF files...
Found 25578 sites common to both files.
Found 959 sites only in main file.
Found 23220 sites only in second file.
Found 1778 non-matching overlapping sites.
After filtering, kept 28315 out of a possible 28315 Sites
Run Time = 1.00 seconds

real    0m1.472s
user    0m1.213s
sys     0m0.191s
```

Quality filtering for freebayes

In [47]:

```
!vcffilter -f "QUAL > 30" ../Results/Variants/freebayes/set5_out_raw_variants.vc
f  >../Results/Variants/freebayes/set5_out_filtered_variants.vcf
```

In [52]:

```
!time vcftools --vcf ../Results/Variants/gatk/setr5_out_raw_variants_recalibrate
d_snps_indels_chr1.vcf \
--diff ../Results/Variants/freebayes/set5_out_filtered_variants.vcf \
--diff-site --out ../Results/Variants/setr5_out_filtered_variants.diff
```

```
VCFtools - 0.1.16
(C) Adam Auton and Anthony Marcketta 2009

Parameters as interpreted:
        --vcf ../Results/Variants/gatk/setr5_out_raw_variants_recali
brated_snps_indels_chr1.vcf
        --out ../Results/Variants/setr5_out_filtered_variants.diff
        --diff ../Results/Variants/freebayes/set5_out_filtered_varia
nts.vcf
        --diff-site

Warning: Expected at least 2 parts in FORMAT entry: ID=PL,Number=G,T
ype=Integer,Description="Normalized, Phred-scaled likelihoods for ge
notypes as defined in the VCF specification">
Warning: Expected at least 2 parts in INFO entry: ID=AC,Number=A,Typ
e=Integer,Description="Allele count in genotypes, for each ALT allel
e, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AC,Number=A,Typ
e=Integer,Description="Allele count in genotypes, for each ALT allel
e, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Allele Frequency, for each ALT allele, in the s
ame order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Allele Frequency, for each ALT allele, in the s
ame order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAC,Number=A,
Type=Integer,Description="Maximum likelihood expectation (MLE) for t
he allele counts (not necessarily the same as the AC), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAC,Number=A,
Type=Integer,Description="Maximum likelihood expectation (MLE) for t
he allele counts (not necessarily the same as the AC), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAF,Number=A,
Type=Float,Description="Maximum likelihood expectation (MLE) for the
allele frequency (not necessarily the same as the AF), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=MLEAF,Number=A,
Type=Float,Description="Maximum likelihood expectation (MLE) for the
allele frequency (not necessarily the same as the AF), for each ALT
allele, in the same order as listed">
Warning: Expected at least 2 parts in INFO entry: ID=culprit,Number=
1,Type=String,Description="The annotation which was the worst perfor
ming in the Gaussian mixture model, likely the reason why the varian
t was filtered out">
After filtering, kept 1 out of 1 Individuals
Warning: Expected at least 2 parts in INFO entry: ID=AF,Number=A,Typ
e=Float,Description="Estimated allele frequency in the range (0,1]">
Warning: Expected at least 2 parts in INFO entry: ID=PRO,Number=1,Ty
pe=Float,Description="Reference allele observation count, with parti
al observations recorded fractionally">
Warning: Expected at least 2 parts in INFO entry: ID=PAO,Number=A,Ty
pe=Float,Description="Alternate allele observations, with partial ob
servations recorded fractionally">
Warning: Expected at least 2 parts in INFO entry: ID=SRP,Number=1,Ty
pe=Float,Description="Strand balance probability for the reference a
llele: Phred-scaled upper-bounds estimate of the probability of obse
rving the deviation between SRF and SRR given E(SRF/SRR) ~ 0.5, deri
ved using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=SAP,Number=A,Ty
```

```
pe=Float,Description="Strand balance probability for the alternate a
llele: Phred-scaled upper-bounds estimate of the probability of obse
rving the deviation between SAF and SAR given E(SAF/SAR) ~ 0.5, deri
ved using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=AB,Number=A,Typ
e=Float,Description="Allele balance at heterozygous sites: a number
between 0 and 1 representing the ratio of reads showing the referenc
e allele to all reads, considering only reads from individuals calle
d as heterozygous">
Warning: Expected at least 2 parts in INFO entry: ID=ABP,Number=A,Ty
pe=Float,Description="Allele balance probability at heterozygous sit
es: Phred-scaled upper-bounds estimate of the probability of observi
ng the deviation between ABR and ABA given E(ABR/ABA) ~ 0.5, derived
using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=RPP,Number=A,Ty
pe=Float,Description="Read Placement Probability: Phred-scaled upper
-bounds estimate of the probability of observing the deviation betwe
en RPL and RPR given E(RPL/RPR) ~ 0.5, derived using Hoeffding's ine
quality">
Warning: Expected at least 2 parts in INFO entry: ID=RPPR,Number=1,T
ype=Float,Description="Read Placement Probability for reference obse
rvations: Phred-scaled upper-bounds estimate of the probability of o
bserving the deviation between RPL and RPR given E(RPL/RPR) ~ 0.5, d
erived using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=EPP,Number=A,Ty
pe=Float,Description="End Placement Probability: Phred-scaled upper-
bounds estimate of the probability of observing the deviation betwee
n EL and ER given E(EL/ER) ~ 0.5, derived using Hoeffding's inequali
ty">
Warning: Expected at least 2 parts in INFO entry: ID=EPPR,Number=1,T
ype=Float,Description="End Placement Probability for reference obser
vations: Phred-scaled upper-bounds estimate of the probability of ob
serving the deviation between EL and ER given E(EL/ER) ~ 0.5, derive
d using Hoeffding's inequality">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=TYPE,Number=A,T
ype=String,Description="The type of allele, either snp, mnp, ins, de
l, or complex.">
Warning: Expected at least 2 parts in INFO entry: ID=CIGAR,Number=A,
Type=String,Description="The extended CIGAR representation of each a
lternate allele, with the exception that '=' is replaced by 'M' to e
ase VCF parsing.  Note that INDEL alleles do not have the first matc
hed base (which is provided by default, per the spec) referred to by
the CIGAR.">
Warning: Expected at least 2 parts in FORMAT entry: ID=GQ,Number=1,T
ype=Float,Description="Genotype Quality, the Phred-scaled marginal
(or unconditional) probability of the called genotype">
Warning: Expected at least 2 parts in FORMAT entry: ID=GL,Number=G,T
ype=Float,Description="Genotype Likelihood, log10-scaled likelihoods
of the data given the called genotype for each possible genotype gen
```

```
erated from the reference and alternate alleles given the sample plo
idy">
Comparing sites in VCF files...
Found 22357 sites common to both files.
Found 4508 sites only in main file.
Found 1923 sites only in second file.
Found 1450 non-matching overlapping sites.
After filtering, kept 28315 out of a possible 28315 Sites
Run Time = 0.00 seconds

real    0m0.927s
user    0m0.791s
sys     0m0.116s
```

In [53]:

```
22357/28315
```

Out[53]:

```
0.7895814939078227
```

In [4]:

```
!time gatk GenotypeConcordance \
-CV ../Results/Variants/gatk/setr5_out_raw_variants_recalibrated_snps_indels.vcf
\
-TV ../Data/VcfDatabase/b37/dbsnp_138.b37.vcf \
-O ../Results/Variants/set5_metrics_concordance
```

```
Using GATK jar /home/caleb/miniconda3/envs/icipe-env/share/gatk4-4.
0.8.1-0/gatk-package-4.0.8.1-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_asyn
c_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -
Dsamjdk.compression_level=2 -jar /home/caleb/miniconda3/envs/icipe-e
nv/share/gatk4-4.0.8.1-0/gatk-package-4.0.8.1-local.jar GenotypeConc
ordance -CV ../Results/Variants/gatk/setr5_out_raw_variants_recalibr
ated_snps_indels.vcf -TV ../Data/VcfDatabase/b37/dbsnp_138.b37.vcf -
O ../Results/Variants/set5_metrics_concordance
18:01:52.393 INFO  NativeLibraryLoader - Loading libgkl_compression.
so from jar:file:/home/caleb/miniconda3/envs/icipe-env/share/gatk4-
4.0.8.1-0/gatk-package-4.0.8.1-local.jar!/com/intel/gkl/native/libgk
l_compression.so
[Sun Sep 30 18:01:52 EAT 2018] GenotypeConcordance  --TRUTH_VCF ../D
ata/VcfDatabase/b37/dbsnp_138.b37.vcf --CALL_VCF ../Results/Variant
s/gatk/setr5_out_raw_variants_recalibrated_snps_indels.vcf --OUTPUT
../Results/Variants/set5_metrics_concordance  --OUTPUT_VCF false --I
NTERSECT_INTERVALS true --MIN_GQ 0 --MIN_DP 0 --OUTPUT_ALL_ROWS fals
e --USE_VCF_INDEX false --MISSING_SITES_HOM_REF false --IGNORE_FILTE
R_STATUS false --VERBOSITY INFO --QUIET false --VALIDATION_STRINGENC
Y STRICT --COMPRESSION_LEVEL 2 --MAX_RECORDS_IN_RAM 500000 --CREATE_
INDEX false --CREATE_MD5_FILE false --GA4GH_CLIENT_SECRETS client_se
crets.json --help false --version false --showHidden false --USE_JDK
_DEFLATER false --USE_JDK_INFLATER false
[Sun Sep 30 18:01:52 EAT 2018] Executing as caleb@hpc01.icipe.org on
Linux 2.6.32-696.30.1.el6.x86_64 amd64; OpenJDK 64-Bit Server VM 1.
8.0_121-b15; Deflater: Intel; Inflater: Intel; Provider GCS is avail
able; Picard version: Version:4.0.8.1
[Sun Sep 30 18:01:52 EAT 2018] picard.vcf.GenotypeConcordance done.
Elapsed time: 0.00 minutes.
Runtime.totalMemory()=1556611072
To get help, see http://broadinstitute.github.io/picard/index.html#G
ettingHelp
java.lang.IndexOutOfBoundsException: Index: 0, Size: 0
        at java.util.ArrayList.rangeCheck(ArrayList.java:653)
        at java.util.ArrayList.get(ArrayList.java:429)
        at picard.vcf.GenotypeConcordance.doWork(GenotypeConcordanc
e.java:331)
        at picard.cmdline.CommandLineProgram.instanceMain(CommandLin
eProgram.java:282)
        at org.broadinstitute.hellbender.cmdline.PicardCommandLinePr
ogramExecutor.instanceMain(PicardCommandLineProgramExecutor.java:25)
        at org.broadinstitute.hellbender.Main.runCommandLineProgram
(Main.java:160)
        at org.broadinstitute.hellbender.Main.mainEntry(Main.java:20
3)
        at org.broadinstitute.hellbender.Main.main(Main.java:289)

real    0m2.955s
user    0m7.197s
sys     0m0.741s
```