A Two-level Scheme for Quality Score Compression

Supplementary Material

Jan Voges, Ali Fotouhi, Jörn Ostermann and M. Oğuzhan Külekci

Contents

1 Tools

Table 1 lists the tools, including QScomp, that were selected for the evaluation in this work.

Tool name	Tool version	Lossless (Y/N)	Lossy (Y/N)
QScomp	ec5c61b	Y	Y
Crumble	0.5	N	Y
Quartz	0.2.2	N	Y
QVZ 2	d5383c6	Y	Y

Table 1: Tools selected for the evaluation.

1.1 QScomp

QScomp can be downloaded from https://github.com/voges/QScomp. Build and usage information as well as the supplementary scripts are available on the respective website.

QScomp compresses quality scores extracted from e.g. a FASTQ, SAM, or BAM file.

The quality scores were extracted from a SAM file `file.sam` with the following command.

```
$ python xtract_field_sam.py file.sam 10 1> file.qual
```

The quality scores are then stored in the file `file.qual`. Compression of the quality scores was performed with the following commands.

```
$ QScomp file.qual
$ bzip2 -9 -c file.qual.dim1 > file.qual.dim1.bz2
$ for f in file.qual.dim2.*; do
$     bzip2 -9 -c $f > $f.bz2
$ done
```

QScomp produces the file `file.qual.dim1` which contains the lossy representation of the quality scores. The file `file.qual.dim1_rc` contains the reconstructed quality scores. The files `file.qual.dim2.*` contain the necessary information for the lossless reconstruction of the quality scores. The file `file.qual.dim2_a` contains all second-level residues (i.e., all data that is distributed among the files `file.qual.dim2.*`).

Finally, a SAM file containing the reconstructed quality scores was produced with the following command.

```
$ python replace_qual_sam.py file.sam file.qual.dim1_rc 1> file.recon_qual.
    sam
```

This produces a new SAM file `file.recon_qual.sam` which contains the reconstructed quality scores.

1.2 Crumble

Crumble 0.5 was downloaded from https://github.com/jkbonfield/crumble. BAM-to-BAM compression of a BAM file `file.bam` with Crumble for the two compression levels `-1` and `-9` was performed with the following commands.

```
$ crumble -v -1 file.bam file.bam.crumble-1.bam
$ crumble -v -9 file.bam file.bam.crumble-9.bam
```

Subsequently, the resulting BAM files, which contain the modified quality values, were compressed with Scramble 1.14.6 [Bon14] using the following commands. Scramble was downloaded from https://github.com/jkbonfield/io_lib.

```
$ scramble -r ref.fa file.bam.crumble-1.bam file.bam.crumble-1.bam.cram
$ scramble -r ref.fa file.bam.crumble-9.bam file.bam.crumble-9.bam.cram
```

The compressed size of the quality values in the resulting CRAM files was determined using the tool cram_size which is included in the Scramble package.

```
$ cram_size file.bam.crumble-1.bam.cram
$ cram_size file.bam.crumble-9.bam.cram
```

Cram_size outputs the sizes of numerous data classes contained in a CRAM file. The data class named "QS" corresponds to the compressed size of the quality values.

1.3 Quartz

Quartz 0.2.2 [YYPB15] was downloaded from https://github.com/yunwilliamyu/quartz. For the working of Quartz, a sequence dictionary is necessary. The sequence dictionary `dec200.bin.sorted` used in this work was downloaded from http://giant.csail.mit.edu/quartz/dec200.bin.sorted.gz. The modification of quality values of a FASTQ file `reads.fastq` with Quartz and the extraction of the quality values were then performed with the following two commands.

```
$ quartz dec200.bin.sorted "quartz" 8 0 reads.fastq
$ python xtract_qual_fastq.py reads.fastq.filtered_quartz \
      2> reads.fastq.filtered_quartz.qual
```

As recommended by the Quartz authors, we subsequently applied bzip2 compression on the modified quality values with the following command.

```
$ bzip2 reads.fastq.filtered_quartz.qual
```

The size of the compressed quality values was determined with the following command.

```
$ wc -c reads.fastq.filtered_quartz.qual.bz2
```

1.4 QVZ 2

QVZ 2 [HOW16] was downloaded from https://github.com/mikelhernaez/qvz2. Compression of quality values of a SAM file `file.sam` with QVZ 2 was performed with the following two commands. We performed compression for the targeted mean square error (MSE) distortions $T \in \{1, 2, 4, 8, 16\}$.

```
$ python xtract_qual_sam.py file.sam 2> file.sam.qual
$ qvz2 -t $T -v -u file.sam.qvz2.qual \
      file.sam.qual file.sam.qual.qvz2
$ python replace_qual_sam.py file.sam file.sam.qvz2.qual
```

The size of the compressed quality values was determined with the following command.

```
$ wc -c file.sam.qual.qvz2
```

2 Datasets

To evaluate the performance of the chosen compression tools, we used the datasets shown in Table 2. For the dataset H01 we selected the first pair of FASTQ files, namely `ERR174324_1.fastq` and `ERR174324_2.fastq`.

ID	Name	Sequencing technology	Coverage
H01	ERR174324	Illumina HiSeq 2000	14×
H11	SRR1238539	Ion Torrent	10×
H12	Garvan replicate	Illumina HiSeq X	49×

Table 2: Datasets selected for the evaluation.

The data were downloaded from the following locations.

- H01: `http://www.ebi.ac.uk/ena/data/view/ERP001775/`
- H11: `ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA096/SRA096885/SRX517292/SRR1238539.fastq.bz2`
- H12, file 1: `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R1.fastq.gz`
- H12, file 2: `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R2.fastq.gz`

3 Performance measurements

We measured the maximum memory usage of each tool with GNU time 1.7. Furthermore, we measured the execution time of each tool also with GNU time 1.7. For example, to measure the performance of QScomp, we used the following command.

```
$ time -v -o file.qual.QScomp_mem+time QScomp file.qual
```

The complete performance results for all tools and datasets are shown in Figure 1.

The maximum RAM usage results for all tools and datasets are shown in Figure 2. Note that we applied a logarithmic scaling to the y-axis.

The running times for all tools and datasets are shown in Figure 3.

4 Variant calling pipelines

This section provides information on the specific configurations of the variant calling pipelines used to assess the performance of QScomp and of the previously proposed compression tools. The alignment and preprocessing is common to all pipelines.

Specifically, we used three different pipelines.

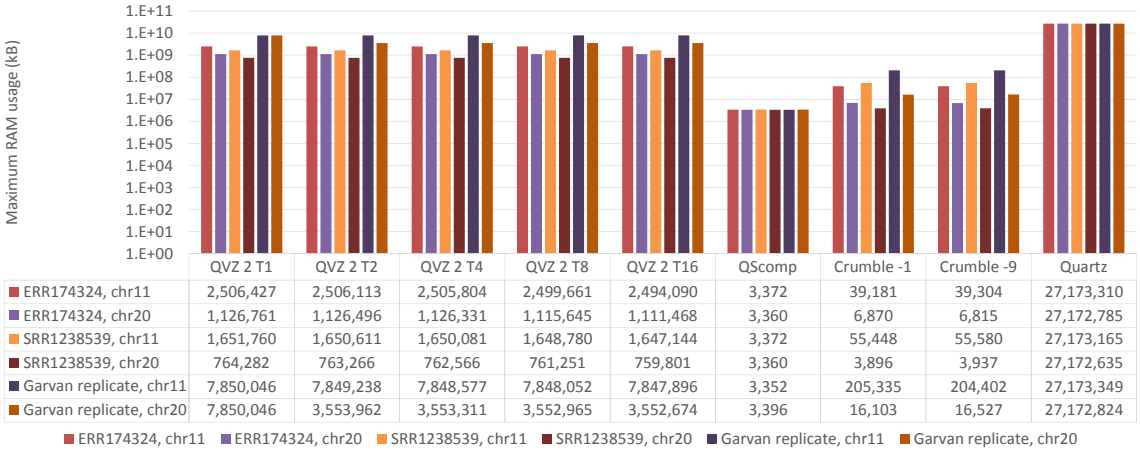
- GATK + VQSR: GATK variant calling and SNP extraction with subsequent filtering of variants using GATK Vector Quality Score Recalibration (VQSR) with four different filter values.
- GATK + hard filtration: GATK variant calling and SNP extraction with the more traditional subsequent hard filtration of variants.
- Platypus: Platypus variant calling as recommended by the authors.

H01 (ERR174324)									
	Chromosome 11				Chromosome 20				Platform
	RAM usage (kB)	Time (s)			RAM usage (kB)	Time (s)			
	Max	User	System	Total	Max	User	System	Total	
QVZ 2 T1	2,506,427	272	6	278	1,126,761	126	1	127	Intel Xeon E5-2680 v3 CPU (2.50 GHz);270 GB RAM
QVZ 2 T2	2,506,113	237	3	240	1,126,496	110	1	111	
QVZ 2 T4	2,505,804	223	6	229	1,126,331	96	1	97	
QVZ 2 T8	2,499,661	186	2	188	1,115,645	89	1	90	
QVZ 2 T16	2,494,090	183	2	185	1,111,468	83	1	84	
QScomp	3,372	131	5	136	3,360	57	2	59	
Crumble -1	39,181	976	8	984	6,870	359	3	362	
Crumble -9	39,304	697	4	701	6,815	289	3	292	
Quartz	27,173,310	1,118	261	1,379	27,172,785	456	238	694	

H11 (SRR1238539)									
	Chromosome 11				Chromosome 20				Platform
	RAM usage (kB)	Time (s)			RAM usage (kB)	Time (s)			
	Max	User	System	Total	Max	User	System	Total	
QVZ 2 T1	1,651,760	312	3	315	764,282	254	2	256	Intel Xeon E5-2680 v3 CPU (2.50 GHz); 270 GB RAM
QVZ 2 T2	1,650,611	303	4	307	763,266	239	2	241	
QVZ 2 T4	1,650,081	287	3	290	762,566	251	3	254	
QVZ 2 T8	1,648,780	285	3	288	761,251	117	1	118	
QVZ 2 T16	1,647,144	280	4	284	759,801	170	2	172	
QScomp	3,372	131	4	135	3,360	57	2	59	
Crumble -1	55,448	3,078	9	3,087	3,896	1,282	3	1,285	
Crumble -9	55,580	916	5	921	3,937	414	2	416	
Quartz	27,173,165	618	244	862	27,172,635	260	190	450	

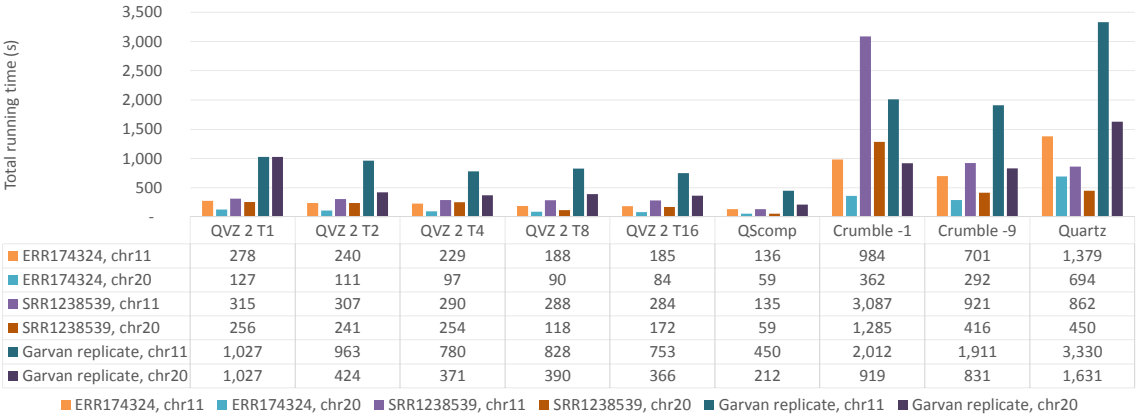
H12 (Garvan replicate)									
	Chromosome 11				Chromosome 20				Platform
	RAM usage (kB)	Time (s)			RAM usage (kB)	Time (s)			
	Max	User	System	Total	Max	User	System	Total	
QVZ 2 T1	7,850,046	1,017	10	1,027	7,850,046	1,017	10	1,027	Intel Xeon E5-2680 v3 CPU (2.50 GHz); 270 GB RAM
QVZ 2 T2	7,849,238	946	17	963	3,553,962	420	4	424	
QVZ 2 T4	7,848,577	766	14	780	3,553,311	367	4	371	
QVZ 2 T8	7,848,052	812	16	828	3,552,965	382	8	390	
QVZ 2 T16	7,847,896	736	17	753	3,552,674	362	4	366	
QScomp	3,352	436	14	450	3,396	205	7	212	
Crumble -1	205,335	1,998	14	2,012	16,103	911	8	919	
Crumble -9	204,402	1,897	14	1,911	16,527	825	6	831	
Quartz	27,173,349	2,943	387	3,330	27,172,824	1,346	285	1,631	

Figure 1: Performance measurements results.



	QVZ 2 T1	QVZ 2 T2	QVZ 2 T4	QVZ 2 T8	QVZ 2 T16	QScomp	Crumble -1	Crumble -9	Quartz
ERR174324, chr11	2,506,427	2,506,113	2,505,804	2,499,661	2,494,090	3,372	39,181	39,304	27,173,310
ERR174324, chr20	1,126,761	1,126,496	1,126,331	1,115,645	1,111,468	3,360	6,870	6,815	27,172,785
SRR1238539, chr11	1,651,760	1,650,611	1,650,081	1,648,780	1,647,144	3,372	55,448	55,580	27,173,165
SRR1238539, chr20	764,282	763,266	762,566	761,251	759,801	3,360	3,896	3,937	27,172,635
Garvan replicate, chr11	7,850,046	7,849,238	7,848,577	7,848,052	7,847,896	3,352	205,335	204,402	27,173,349
Garvan replicate, chr20	7,850,046	3,553,962	3,553,311	3,552,965	3,552,674	3,396	16,103	16,527	27,172,824

Figure 2: Maximum RAM usage results.



	QVZ 2 T1	QVZ 2 T2	QVZ 2 T4	QVZ 2 T8	QVZ 2 T16	QScomp	Crumble -1	Crumble -9	Quartz
ERR174324, chr11	278	240	229	188	185	136	984	701	1,379
ERR174324, chr20	127	111	97	90	84	59	362	292	694
SRR1238539, chr11	315	307	290	288	284	135	3,087	921	862
SRR1238539, chr20	256	241	254	118	172	59	1,285	416	450
Garvan replicate, chr11	1,027	963	780	828	753	450	2,012	1,911	3,330
Garvan replicate, chr20	1,027	424	371	390	366	212	919	831	1,631

Figure 3: Total running time results.

The following tool versions were used.

- Bowtie 2 2.2.5 [LS12]
- Picard 2.4.1
- SAMtools 1.3 (built with HTSlib version 1.3) [LHW+09]
- GATK 3.6 [MHB+10]
- Platypus (latest stable release downloaded from `http://www.well.ox.ac.uk/platypus`) [RPM+14]

To perform the GATK variant calling procedure, the following additional file from the GATK resource bundle (`https://software.broadinstitute.org/gatk/download/bundle`) is needed.

- `dbsnp_138.b37.vcf`

To perform the GATK VQSR procedure and the alignment and preprocessing, the following additional files from the GATK resource bundle (`https://software.broadinstitute.org/gatk/download/bundle`) are needed.

- `Mills_and_1000G_gold_standard.indels.b37.vcf`
- `1000G_phase1.indels.b37.vcf`
- `dbsnp_138.b37.vcf`
- `hapmap_3.3.b37.vcf`
- `1000G_omni2.5.b37.vcf`
- `1000G_phase1.snps.high_confidence.b37.vcf`

For the purpose of this evaluation, we used the GATK resource bundle version 2.8.

4.1 Alignment and preprocessing

4.1.1 Alignment with Bowtie 2

The first step consists in building the reference indexes. We used the file `human_g1k_v37.fasta` from the GATK resource bundle as reference `ref.fa`.

```
$ bowtie2-build ref.fa $idx
```

Once completed, the current directory contains new files that all start with `$idx` and end with `.1.bt2`, `.2.bt2`, `.3.bt2`, `.4.bt2`, `.rev.1.bt2`, and `.rev.2.bt2`. These files constitute the index. At this point alignment can take place.

```
$ bowtie2 -x $idx -1 reads_1.fastq -2 reads_2.fastq -S aln.sam
```

4.1.2 Sorting and indexing

We converted the SAM file to the BAM format using SAMtools.

```
$ samtools view -bh aln.sam > aln.bam
```

Then we sorted and indexed the BAM file.

```
$ samtools sort aln.bam > sorted.bam
$ samtools index sorted.bam
```

4.1.3 Duplicate marking

The duplicates were marked in the BAM file using Picard.

```
$ java -jar picard.jar MarkDuplicates \
      I=sorted.bam \
      O=dupmark.bam \
      M=metrics.txt \
      ASSUME_SORTED=true
```

Subsequently, we used Picard to label the BAM headers.

```
$ java -jar picard.jar AddOrReplaceReadGroups \
      I=dupmark.bam \
      O=label.bam \
      RGID=1 \
      RGLB=Library \
      RGPL=Illumna\
      RGPU=PlatformUnit \
      RGSM=SampleName
```

Then we indexed the resulting file.

```
$ samtools index label.bam
```

4.1.4 Indel realignment

We created the target list of intervals with GATK.

```
$ java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
      -R ref.fa \
      -I label.bam \
```

```
      --known Mills_and_1000G_gold_standard.indels.b37.vcf \
      -o target_intervals.list
```

The following command performs the realignment.

```
$ java -jar GenomeAnalysisTK.jar -T IndelRealigner \
      -R ref.fa \
      -I label.bam \
      -targetIntervals target_intervals.list \
      -o realign.bam
```

4.1.5 Base quality score recalibration (BQSR)

A recalibration of the quality values was performed using the following two commands.

```
$ java -jar GenomeAnalysisTK.jar -T BaseRecalibrator \
      -R ref.fa \
      -I realign.bam \
      -knownSites dbsnp_138.b37.vcf \
      -knownSites Mills_and_1000G_gold_standard.indels.b37.vcf \
      -knownSites 1000G_phase1.indels.b37.vcf \
      -o recal.data

$ java -jar GenomeAnalysisTK.jar -T PrintReads \
      -R ref.fa \
      -I realign.bam \
      -BQSR recal.data \
      -o recal.bam
```

4.2 Variant calling

4.2.1 SNP calling with GATK

We consider the tool HaplotypeCaller as the variant caller for the GATK pipeline to call variants on chromosome `$chr`.

```
$ java -jar GenomeAnalysisTK.jar -T HaplotypeCaller \
      -R ref.fa \
      -L $chr \
      -I recal.bam \
      --dbsnp dbsnp_138.b37.vcf \
      --genotyping_mode DISCOVERY \
      -stand_emit_conf 10 \
      -stand_call_conf 30 \
      -o gatk_calls.vcf
```

Once the calls are made, SNPs extraction was performed using the following command.

```
$ java -jar GenomeAnalysisTK.jar -T SelectVariants \
      -R ref.fa \
      -L $chr \
      -V gatk_calls.vcf \
      -selectType SNP \
      -o gatk_snps.vcf
```

4.2.1.1 Variant quality score recalibration (VQSR) Call filtering was performed using the VQSR command. First, the SNP recalibration model was built where 100.0, 99.9, 99.0 and 90.0 are the thresholds used:

```
$ java -jar GenomeAnalysisTK.jar -T VariantRecalibrator \
      -R ref.fa \
      -L $chr \
      -input gatk_snps.vcf \
      -resource:hapmap,known=false,training=true,truth=true,prior=15.0 \
          hapmap_3.3.b37.vcf \
      -resource:omni,known=false,training=true,truth=true,prior=12.0 \
          1000G_omni2.5.b37.vcf \
      -resource:1000G,known=false,training=true,truth=false,prior=10.0 \
          1000G_phase1.snps.high_confidence.b37.vcf \
      -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 \
          dbsnp_138.b37.vcf \
      -an DP -an QD -an FS -an SOR -an MQ -an MQRankSum \
      -an ReadPosRankSum \
      -mode SNP \
      -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
      -recalFile gatk_snps.recal \
      -tranchesFile gatk_snps.tranches \
      -rscriptFile gatk_snps.r
```

Then the desired level of recalibration was applied. Note that the variable `$recal_level` should be 100.0, 99.9, 99.0, and 90.0.

```
$ java -jar GenomeAnalysisTK.jar -T ApplyRecalibration \
      -R ref.fa \
      -L $chr \
      -input gatk_snps.vcf \
      -mode SNP \
      --ts_filter_level $recal_level \
      -recalFile gatk_snps.recal \
      -tranchesFile gatk_snps.tranches \
      -o recal.vcf
```

4.2.1.2 Hard filtration

Hard filtration of variants was performed with the following command as recommended by the GATK authors in http://gatkforums.broadinstitute.org/gatk/discussion/2806/.

```
$ java -jar GenomeAnalysisTK.jar -T VariantFiltration \
      -R ref.fa \
      -L $chr \
      -V gatk_snps.vcf \
      --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || \
          MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
      --filterName "GATK_Recommended" \
      -o filtered.vcf
```

4.2.2 SNP calling with Platypus

SNP calling with Platypus was performed with the following commands.

```
$ python Platypus.py callVariants \
      --bamFiles=recal.bam \
      --refFile=ref.fa \
      --regions=$chr \
      --output=platypus_calls.vcf \
      --logFileName=log.txt
$ java -jar GenomeAnalysisTK.jar -T SelectVariants \
      -R ref.fa \
      -L $chr \
```

```
    -V platypus_calls.vcf \
    -selectType SNP \
    -o platypus_snps.vcf
```

4.3 Benchmarking tools

We used the benchmarking tools proposed by the Global Alliance for Genomics and Health (GA4GH). The tools were downloaded from the following locations.

- https://github.com/ga4gh/benchmarking-tools
- https://github.com/Illumina/hap.py

The benchmarking is mainly based on the haplotype comparison tool hap.py, developed by Illumina. Hap.py requires the following files from the Genome in a Bottle (GIAB) high-confidence variant call set:

- the VCF file containing the "golden reference" (`gt.vcf`),
- the BED file containing the confident regions of the golden reference (`gt.bed`),
- the VCF file generated after running the variant calling pipeline (`input.vcf`),
- the FASTA file containing the reference sequence(s) used for alignment (`ref.fa`).

We used the GIAB high-confidence variant call set version 3.2.2 which was downloaded from https://www.nist.gov/programs-projects/genome-bottle. Specifically, we used the following command to run hap.py.

```
$ python hap.py gt.vcf input.vcf \
      -f gt.bed \
      -o happy_root \
      -r ref.fa \
      --roc VQLSOD
```

We used the benchmarking tool rep.py from the GA4GH to summarize the output files in an HTML file.

5 Variant calling results

The results shown in the main paper are the results of variant calling with the GATK + VQSR pipeline. In the main paper, we averaged the Recall and Precision metrics over the two chromosomes 11 and 20 and over the four VQSR filter values ($\theta \in \{90, 99, 99.9, 100\}$) which in total yielded 6 plots (i.e., 3 data sets $\times$ 2 metrics).

In this section, we show tables with all variant calling results (from all 3 pipelines) generated in this study.

5.1 GATK + VQSR

H01 (ERR174324)

Chromosome 11

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	1,986,669,293	8.0000	0.7846	0.9983	0.8786	0.9147	0.9978	0.9544	0.9351	0.9978	0.9654	0.9471	0.9972	0.9715	0.8954	0.9978	0.9425	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	180,808,513	0.7281	0.7847	0.9983	0.8787	0.9144	0.9977	0.9542	0.9349	0.9977	0.9653	0.9468	0.9971	0.9713	0.8952	0.9977	0.9424	-0.0002	-0.0001	-0.0001	N/A
QVZ 2 T2	93,357,561	0.3759	0.7959	0.9979	0.8855	0.9009	0.9978	0.9469	0.9311	0.9977	0.9633	0.9468	0.9968	0.9712	0.8937	0.9976	0.9417	-0.0017	-0.0002	-0.0008	N/A
QVZ 2 T4	32,931,170	0.1326	0.7840	0.9964	0.8775	0.9155	0.9959	0.9540	0.9368	0.9959	0.9654	0.9467	0.9944	0.9700	0.8958	0.9957	0.9417	0.0004	-0.0021	-0.0008	N/A
QVZ 2 T8	9,832,950	0.0396	0.7819	0.9902	0.8738	0.9136	0.9902	0.9504	0.9333	0.9897	0.9607	0.9468	0.9849	0.9655	0.8939	0.9888	0.9376	-0.0015	-0.0090	-0.0049	N/A
QVZ 2 T16	3,480,969	0.0140	0.7867	0.9856	0.8750	0.9163	0.9857	0.9497	0.9353	0.9852	0.9596	0.9465	0.9794	0.9627	0.8962	0.9840	0.9367	0.0008	-0.0138	-0.0058	N/A
QScomp (+ bzip2 -9)	36,732,800	0.1479	0.7848	0.9984	0.8788	0.9148	0.9977	0.9545	0.9342	0.9977	0.9649	0.9456	0.9972	0.9707	0.8949	0.9978	0.9422	-0.0005	0.0000	-0.0003	dim1
Crumble -1 (+ CRAM)	100,830,366	0.4060	0.7883	0.9981	0.8809	0.9166	0.9978	0.9555	0.9347	0.9978	0.9652	0.9471	0.9973	0.9716	0.8967	0.9978	0.9433	0.0013	0.0000	0.0008	N/A
Crumble -9 (+ CRAM)	58,407,264	0.2352	0.7878	0.9980	0.8805	0.9145	0.9977	0.9543	0.9337	0.9977	0.9646	0.9479	0.9972	0.9719	0.8960	0.9977	0.9428	0.0006	-0.0001	0.0003	N/A
Quartz (+ bzip2)	111,880,409	0.4505	0.7880	0.9980	0.8807	0.9174	0.9976	0.9558	0.9384	0.9975	0.9670	0.9503	0.9967	0.9729	0.8985	0.9975	0.9441	0.0031	-0.0003	0.0016	N/A

H11 (SRR1238539)

Chromosome 11

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	1,402,713,277	8.0000	0.4824	0.9379	0.6371	0.5392	0.9398	0.6852	0.5524	0.9393	0.6957	0.5555	0.9372	0.6975	0.5324	0.9386	0.6789	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	346,716,840	1.9774	0.4832	0.9386	0.6380	0.5380	0.9402	0.6844	0.5500	0.9402	0.6940	0.5544	0.9375	0.6968	0.5314	0.9391	0.6783	-0.0010	0.0006	-0.0006	N/A
QVZ 2 T2	273,266,584	1.5585	0.4870	0.9384	0.6412	0.5418	0.9405	0.6875	0.5541	0.9402	0.6973	0.5579	0.9378	0.6996	0.5352	0.9392	0.6814	0.0028	0.0007	0.0025	N/A
QVZ 2 T4	198,709,595	1.1333	0.4786	0.9365	0.6335	0.5369	0.9385	0.6830	0.5485	0.9375	0.6921	0.5517	0.9356	0.6941	0.5289	0.9370	0.6757	-0.0035	-0.0015	-0.0032	N/A
QVZ 2 T8	121,742,167	0.6943	0.4850	0.9364	0.6390	0.5382	0.9382	0.6840	0.5505	0.9378	0.6938	0.5541	0.9357	0.6960	0.5320	0.9370	0.6782	-0.0004	-0.0015	-0.0007	N/A
QVZ 2 T16	60,751,253	0.3465	0.4872	0.9382	0.6414	0.5449	0.9395	0.6898	0.5578	0.9384	0.6997	0.5612	0.9359	0.7017	0.5378	0.9380	0.6831	0.0054	-0.0006	0.0042	N/A
QScomp (+ bzip2 -9)	132,239,998	0.7542	0.3490	0.9029	0.5034	0.4016	0.9059	0.5565	0.4111	0.9047	0.5653	0.4135	0.9044	0.5675	0.3938	0.9045	0.5482	-0.1386	-0.0341	-0.1307	dim1
Crumble -1 (+ CRAM)	642,683,176	3.6654	0.4855	0.9381	0.6399	0.5405	0.9402	0.6864	0.5526	0.9400	0.6960	0.5564	0.9375	0.6983	0.5338	0.9390	0.6802	0.0014	0.0004	0.0013	N/A
Crumble -9 (+ CRAM)	570,828,844	3.2556	0.4922	0.9402	0.6461	0.5478	0.9421	0.6928	0.5600	0.9416	0.7023	0.5638	0.9392	0.7046	0.5410	0.9408	0.6865	0.0086	0.0022	0.0076	N/A
Quartz (+ bzip2)	247,127,738	1.4094	0.5083	0.9332	0.6581	0.5656	0.9332	0.7043	0.5749	0.9327	0.7113	0.5791	0.9297	0.7137	0.5570	0.9322	0.6969	0.0246	-0.0063	0.0180	N/A

H12 (Garvan replicate)

Chromosome 11

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	6,696,582,269	8.0000	0.7853	0.9999	0.8797	0.9367	0.9997	0.9672	0.9720	0.9990	0.9853	0.9861	0.9973	0.9917	0.9200	0.9990	0.9560	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	814,504,450	0.9730	0.8035	0.9998	0.8910	0.9361	0.9996	0.9668	0.9691	0.9992	0.9839	0.9861	0.9972	0.9916	0.9237	0.9990	0.9583	0.0037	0.0000	0.0024	N/A
QVZ 2 T2	583,280,511	0.6968	0.7907	0.9999	0.8831	0.9353	0.9997	0.9664	0.9730	0.9991	0.9859	0.9861	0.9971	0.9916	0.9213	0.9990	0.9567	0.0013	0.0000	0.0008	N/A
QVZ 2 T4	369,338,993	0.4412	0.8146	0.9998	0.8977	0.9306	0.9996	0.9639	0.9751	0.9989	0.9869	0.9860	0.9971	0.9915	0.9266	0.9989	0.9600	0.0065	-0.0001	0.0040	N/A
QVZ 2 T8	259,975,714	0.3106	0.7919	0.9999	0.8838	0.9256	0.9998	0.9613	0.9723	0.9992	0.9856	0.9860	0.9971	0.9915	0.9190	0.9990	0.9555	0.0011	0.0000	0.0004	N/A
QVZ 2 T16	152,385,057	0.1820	0.7943	0.9998	0.8853	0.9344	0.9996	0.9659	0.9716	0.9991	0.9852	0.9860	0.9966	0.9913	0.9216	0.9988	0.9569	0.0016	-0.0002	0.0009	N/A
QScomp (+ bzip2 -9)	447,290,216	0.5344	0.7971	0.9998	0.8870	0.9385	0.9995	0.9680	0.9720	0.9990	0.9853	0.9856	0.9973	0.9914	0.9233	0.9989	0.9579	0.0033	-0.0001	0.0020	dim1
Crumble -1 (+ CRAM)	328,644,318	0.3926	0.7966	0.9997	0.8867	0.9339	0.9994	0.9655	0.9723	0.9989	0.9854	0.9863	0.9972	0.9917	0.9223	0.9988	0.9573	0.0022	-0.0002	0.0014	N/A
Crumble -9 (+ CRAM)	246,283,161	0.2942	0.8006	0.9998	0.8892	0.9365	0.9994	0.9669	0.9707	0.9989	0.9846	0.9865	0.9969	0.9917	0.9236	0.9988	0.9581	0.0035	-0.0002	0.0021	N/A
Quartz (+ bzip2)	625,378,135	0.7471	0.8034	0.9997	0.8909	0.9339	0.9995	0.9656	0.9762	0.9988	0.9874	0.9852	0.9974	0.9913	0.9247	0.9989	0.9588	0.0046	-0.0001	0.0028	N/A

Figure 4: Variant calling results for chromosome 11 for the GATK+VQSR pipeline.

H01 (ERR174324)

Chromosome 20

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	880,786,761	8.0000	0.7964	0.9981	0.8859	0.9266	0.9979	0.9609	0.9458	0.9978	0.9711	0.9553	0.9970	0.9757	0.9060	0.9977	0.9484	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	84,415,660	0.7667	0.8031	0.9978	0.8899	0.9238	0.9978	0.9594	0.9469	0.9976	0.9716	0.9552	0.9970	0.9757	0.9073	0.9976	0.9491	0.0012	-0.0002	0.0007	N/A
QVZ 2 T2	46,732,315	0.4245	0.8024	0.9977	0.8895	0.9262	0.9975	0.9605	0.9420	0.9975	0.9690	0.9551	0.9966	0.9754	0.9064	0.9973	0.9486	0.0004	-0.0004	0.0002	N/A
QVZ 2 T4	17,191,826	0.1561	0.8056	0.9960	0.8907	0.9238	0.9960	0.9585	0.9468	0.9959	0.9707	0.9554	0.9940	0.9743	0.9079	0.9955	0.9486	0.0019	-0.0022	0.0002	N/A
QVZ 2 T8	5,722,879	0.0520	0.8021	0.9885	0.8856	0.9274	0.9890	0.9572	0.9477	0.9883	0.9676	0.9554	0.9823	0.9687	0.9082	0.9870	0.9448	0.0021	-0.0107	-0.0037	N/A
QVZ 2 T16	1,879,305	0.0171	0.8045	0.9820	0.8844	0.9295	0.9826	0.9553	0.9459	0.9823	0.9638	0.9544	0.9728	0.9635	0.9086	0.9799	0.9418	0.0025	-0.0178	-0.0067	N/A
QScomp (+ bzip2 -9)	18,097,457	0.1644	0.7989	0.9980	0.8874	0.9201	0.9980	0.9575	0.9416	0.9978	0.9689	0.9535	0.9972	0.9749	0.9035	0.9978	0.9472	-0.0025	0.0000	-0.0013	dim1
Crumble -1 (+ CRAM)	44,840,996	0.4073	0.7908	0.9983	0.8825	0.9272	0.9979	0.9613	0.9485	0.9974	0.9723	0.9554	0.9972	0.9759	0.9055	0.9977	0.9480	-0.0006	0.0000	-0.0004	N/A
Crumble -9 (+ CRAM)	23,642,566	0.2147	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	VQSR failed for chromosome 20
Quartz (+ bzip2)	46,634,066	0.4236	0.8102	0.9977	0.8942	0.9292	0.9974	0.9621	0.9490	0.9972	0.9725	0.9585	0.9963	0.9770	0.9117	0.9972	0.9515	0.0057	-0.0006	0.0031	N/A

H11 (SRR1238539)

Chromosome 20

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	622,751,899	8.0000	0.4690	0.9270	0.6229	0.5188	0.9296	0.6659	0.5310	0.9294	0.6759	0.5330	0.9267	0.6768	0.5130	0.9282	0.6604	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	155,988,650	2.0039	0.4689	0.9265	0.6227	0.5192	0.9292	0.6662	0.5285	0.9273	0.6733	0.5304	0.9255	0.6743	0.5118	0.9271	0.6591	-0.0012	-0.0010	-0.0012	N/A
QVZ 2 T2	122,833,357	1.5779	0.4680	0.9284	0.6223	0.5220	0.9299	0.6687	0.5317	0.9291	0.6764	0.5285	0.9238	0.6724	0.5126	0.9278	0.6599	-0.0004	-0.0004	-0.0004	N/A
QVZ 2 T4	89,402,714	1.1485	0.4677	0.9258	0.6215	0.5176	0.9277	0.6645	0.5267	0.9261	0.6715	0.5285	0.9238	0.6724	0.5101	0.9259	0.6574	-0.0028	-0.0023	-0.0029	N/A
QVZ 2 T8	55,420,782	0.7119	0.4638	0.9272	0.6183	0.5194	0.9283	0.6661	0.5299	0.9269	0.6743	0.5324	0.9241	0.6756	0.5114	0.9266	0.6586	-0.0016	-0.0015	-0.0018	N/A
QVZ 2 T16	28,272,006	0.3632	0.4731	0.9273	0.6265	0.5250	0.9286	0.6708	0.5358	0.9270	0.6791	0.5379	0.9241	0.6800	0.5180	0.9268	0.6641	0.0050	-0.0014	0.0037	N/A
QScomp (+ bzip2 -9)	59,339,662	0.7623	0.3361	0.8876	0.4876	0.3792	0.8921	0.5322	0.3847	0.8923	0.5376	0.3862	0.8902	0.5387	0.3716	0.8906	0.5240	-0.1414	-0.0376	-0.1363	dim1
Crumble -1 (+ CRAM)	286,005,859	3.6741	0.4706	0.9286	0.6246	0.5229	0.9300	0.6694	0.5316	0.9289	0.6762	0.5337	0.9268	0.6773	0.5147	0.9286	0.6619	0.0018	0.0004	0.0015	N/A
Crumble -9 (+ CRAM)	254,275,065	3.2665	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	VQSR failed for chromosome 20
Quartz (+ bzip2)	108,111,613	1.3888	0.4877	0.9243	0.6385	0.5427	0.9243	0.6839	0.5521	0.9231	0.6909	0.5540	0.9196	0.6914	0.5341	0.9228	0.6762	0.0212	-0.0053	0.0158	N/A

H12 (Garvan replicate)

Chromosome 20

	Size (B)	Bits/QS	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	3,017,194,085	8.0000	0.8030	0.9999	0.8907	0.9373	0.9998	0.9675	0.9824	0.9987	0.9905	0.9920	0.9974	0.9947	0.9287	0.9990	0.9609	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	383,733,965	1.0175	0.8066	0.9998	0.8929	0.9477	0.9996	0.9730	0.9857	0.9985	0.9921	0.9919	0.9972	0.9945	0.9330	0.9988	0.9631	0.0043	-0.0002	0.0023	N/A
QVZ 2 T2	277,566,711	0.7360	0.8264	0.9998	0.9049	0.9446	0.9996	0.9713	0.9823	0.9993	0.9907	0.9919	0.9971	0.9945	0.9363	0.9990	0.9654	0.0076	0.0000	0.0045	N/A
QVZ 2 T4	175,908,797	0.4664	0.8105	0.9999	0.8953	0.9526	0.9997	0.9756	0.9809	0.9991	0.9899	0.9920	0.9972	0.9946	0.9340	0.9990	0.9638	0.0053	0.0000	0.0030	N/A
QVZ 2 T8	125,973,247	0.3340	0.8201	0.9998	0.9011	0.9431	0.9996	0.9705	0.9788	0.9984	0.9885	0.9920	0.9971	0.9945	0.9335	0.9987	0.9637	0.0048	-0.0002	0.0028	N/A
QVZ 2 T16	74,727,311	0.1981	0.8215	0.9997	0.9019	0.9440	0.9995	0.9710	0.9835	0.9989	0.9911	0.9919	0.9964	0.9942	0.9352	0.9986	0.9645	0.0065	-0.0003	0.0037	N/A
QScomp (+ bzip2 -9)	210,050,219	0.5569	0.8133	0.9998	0.8970	0.9520	0.9996	0.9752	0.9795	0.9993	0.9893	0.9916	0.9977	0.9946	0.9341	0.9991	0.9640	0.0054	0.0001	0.0032	dim1
Crumble -1 (+ CRAM)	147,471,578	0.3910	0.8054	0.9999	0.8922	0.9506	0.9995	0.9744	0.9780	0.9985	0.9881	0.9921	0.9973	0.9947	0.9315	0.9988	0.9624	0.0028	-0.0001	0.0015	N/A
Crumble -9 (+ CRAM)	104,264,203	0.2765	0.8308	0.9997	0.9075	0.9422	0.9996	0.9701	0.9802	0.9993	0.9897	0.9923	0.9973	0.9948	0.9364	0.9990	0.9655	0.0077	0.0000	0.0046	N/A
Quartz (+ bzip2)	287,310,936	0.7618	0.8095	0.9997	0.8946	0.9451	0.9996	0.9716	0.9811	0.9991	0.9900	0.9917	0.9971	0.9944	0.9319	0.9989	0.9626	0.0032	-0.0001	0.0018	N/A

Figure 5: Variant calling results for chromosome 20 for the GATK + VQSR pipeline.

H01 (ERR174324)				
	Chromosomes 11 and 20			Remarks
	Bits/QS	Average differences		
		R	P	
Uncompressed	8.0000	0.0000	0.0000	N/A
QVZ 2 T1	0.7474	0.0005	-0.0001	N/A
QVZ 2 T2	0.4002	-0.0006	-0.0003	N/A
QVZ 2 T4	0.1444	0.0011	-0.0022	N/A
QVZ 2 T8	0.0458	0.0003	-0.0099	N/A
QVZ 2 T16	0.0155	0.0017	-0.0158	N/A
QScomp (+ bzip2 -9)	0.1561	-0.0015	0.0000	dim1
Crumble -1 (+ CRAM)	0.4067	0.0004	0.0000	N/A
Crumble -9 (+ CRAM)	0.2250	N/A	N/A	VQSR failed for chromosome 20
Quartz (+ bzip2)	0.4370	0.0044	-0.0004	N/A

H11 (SRR1238539)				
	Chromosomes 11 and 20			Remarks
	Bits/QS	Average differences		
		R	P	
Uncompressed	8.0000	0.0000	0.0000	N/A
QVZ 2 T1	1.9906	-0.0011	-0.0002	N/A
QVZ 2 T2	1.5682	0.0012	0.0001	N/A
QVZ 2 T4	1.1409	-0.0031	-0.0019	N/A
QVZ 2 T8	0.7031	-0.0010	-0.0015	N/A
QVZ 2 T16	0.3548	0.0052	-0.0010	N/A
QScomp (+ bzip2 -9)	0.7582	-0.1400	-0.0358	dim1
Crumble -1 (+ CRAM)	3.6697	0.0016	0.0004	N/A
Crumble -9 (+ CRAM)	3.2610	N/A	N/A	VQSR failed for chromosome 20
Quartz (+ bzip2)	1.3991	0.0229	-0.0058	N/A

H12 (Garvan replicate)				
	Chromosomes 11 and 20			Remarks
	Bits/QS	Average differences		
		R	P	
Uncompressed	8.0000	0.0000	0.00000	N/A
QVZ 2 T1	0.9952	0.0040	-0.00010	N/A
QVZ 2 T2	0.7164	0.0044	-0.00001	N/A
QVZ 2 T4	0.4538	0.0059	-0.00005	N/A
QVZ 2 T8	0.3223	0.0019	-0.00010	N/A
QVZ 2 T16	0.1901	0.0041	-0.00026	N/A
QScomp (+ bzip2 -9)	0.5456	0.0043	0.00004	dim1
Crumble -1 (+ CRAM)	0.3918	0.0025	-0.00016	N/A
Crumble -9 (+ CRAM)	0.2853	0.0056	-0.00010	N/A
Quartz (+ bzip2)	0.7544	0.0039	-0.00010	N/A

Figure 6: Variant calling results averaged over both chromosomes and over all four VQSR filter values for the GATK + VQSR pipeline.

5.2 GATK + hard filtration

H01 (ERR174324)

	Chromosome 11									Chromosome 20									Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences				
						R	P	F							R	P	F		
Uncompressed	1,986,669,293	8.0000	0.8662	0.9979	0.9274	0.0000	0.0000	0.0000		880,786,761	8.0000	0.8942	0.9977	0.9431	0.0000	0.0000	0.0000		N/A
QVZ 2 T1	180,808,513	0.7281	0.8660	0.9979	0.9273	-0.0002	0.0000	-0.0001		84,415,660	0.7667	0.8944	0.9976	0.9432	0.0002	-0.0001	0.0001		N/A
QVZ 2 T2	93,357,561	0.3759	0.8660	0.9976	0.9272	-0.0002	-0.0003	-0.0002		46,732,315	0.4245	0.8941	0.9972	0.9428	-0.0001	-0.0005	-0.0003		N/A
QVZ 2 T4	32,931,170	0.1326	0.8659	0.9953	0.9261	-0.0003	-0.0026	-0.0013		17,191,826	0.1561	0.8944	0.9946	0.9418	0.0002	-0.0031	-0.0013		N/A
QVZ 2 T8	9,832,950	0.0396	0.8661	0.9863	0.9223	-0.0001	-0.0116	-0.0051		5,722,879	0.0520	0.8946	0.9843	0.9373	0.0004	-0.0134	-0.0058		N/A
QVZ 2 T16	3,480,969	0.0140	0.8663	0.9805	0.9199	0.0001	-0.0174	-0.0075		1,879,305	0.0171	0.8943	0.9750	0.9329	0.0001	-0.0227	-0.0102		N/A
QScomp (+ bzip2 -9)	36,732,800	0.1479	0.8646	0.9979	0.9265	-0.0016	0.0000	-0.0009		18,097,457	0.1644	0.8926	0.9978	0.9423	-0.0016	0.0001	-0.0008		dim1
Crumble -1 (+ CRAM)	100,830,366	0.4060	0.8667	0.9980	0.9277	0.0005	0.0001	0.0003		44,840,996	0.4073	0.8952	0.9977	0.9437	0.0010	0.0000	0.0006		N/A
Crumble -9 (+ CRAM)	58,407,264	0.2352	0.8678	0.9977	0.9282	0.0016	-0.0002	0.0008		23,642,566	0.2147	0.8970	0.9971	0.9444	0.0028	-0.0006	0.0013		N/A
Quartz (+ bzip2)	111,880,409	0.4505	0.8707	0.9977	0.9299	0.0045	-0.0002	0.0025		46,634,066	0.4236	0.8983	0.9973	0.9452	0.0041	-0.0004	0.0021		N/A

H11 (SRR1238539)

	Chromosome 11									Chromosome 20									Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences				
						R	P	F							R	P	F		
Uncompressed	1,402,713,277	8.0000	0.3401	0.9380	0.4992	0.0000	0.0000	0.0000		622,751,899	8.0000	0.3296	0.9311	0.4869	0.0000	0.0000	0.0000		N/A
QVZ 2 T1	346,716,840	1.9774	0.3397	0.9381	0.4988	-0.0004	0.0001	-0.0004		155,988,650	2.0039	0.3277	0.9301	0.4846	-0.0019	-0.0010	-0.0022		N/A
QVZ 2 T2	273,266,584	1.5585	0.3428	0.9388	0.5022	0.0027	0.0008	0.0030		122,833,357	1.5779	0.3301	0.9305	0.4873	0.0005	-0.0006	0.0005		N/A
QVZ 2 T4	198,709,595	1.1333	0.3429	0.9363	0.5020	0.0028	-0.0017	0.0028		89,402,714	1.1485	0.3309	0.9275	0.4878	0.0013	-0.0036	0.0009		N/A
QVZ 2 T8	121,742,167	0.6943	0.3442	0.9365	0.5034	0.0041	-0.0015	0.0042		55,420,782	0.7119	0.3318	0.9276	0.4888	0.0022	-0.0035	0.0019		N/A
QVZ 2 T16	60,751,253	0.3465	0.3444	0.9372	0.5037	0.0043	-0.0008	0.0045		28,272,006	0.3632	0.3314	0.9284	0.4884	0.0018	-0.0027	0.0016		N/A
QScomp (+ bzip2 -9)	132,239,998	0.7542	0.2767	0.9019	0.4235	-0.0634	-0.0361	-0.0757		59,339,662	0.7623	0.2592	0.8928	0.4018	-0.0704	-0.0383	-0.0851		dim1
Crumble -1 (+ CRAM)	642,683,176	3.6654	0.3405	0.9385	0.4997	0.0004	0.0005	0.0005		286,005,859	3.6741	0.3299	0.9312	0.4872	0.0003	0.0001	0.0003		N/A
Crumble -9 (+ CRAM)	570,828,844	3.2556	0.3424	0.9407	0.5021	0.0023	0.0027	0.0029		254,275,065	3.2665	0.3324	0.9335	0.4902	0.0028	0.0024	0.0034		N/A
Quartz (+ bzip2)	247,127,738	1.4094	0.3763	0.9298	0.5358	0.0362	-0.0082	0.0366		108,111,613	1.3888	0.3623	0.9219	0.5202	0.0327	-0.0092	0.0333		N/A

H12 (Garvan replicate)

	Chromosome 11									Chromosome 20									Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences				
						R	P	F							R	P	F		
Uncompressed	6,696,582,269	8.0000	0.8676	0.9995	0.9289	0.0000	0.0000	0.0000		3,017,194,085	8.0000	0.8913	0.9994	0.9423	0.0000	0.0000	0.0000		N/A
QVZ 2 T1	814,504,450	0.9730	0.8676	0.9995	0.9289	0.0000	0.0000	0.0000		383,733,965	1.0175	0.8914	0.9993	0.9423	0.0001	-0.0001	0.0000		N/A
QVZ 2 T2	583,280,511	0.6968	0.8678	0.9995	0.9290	0.0002	0.0000	0.0001		277,566,711	0.7360	0.8914	0.9992	0.9422	0.0001	-0.0002	0.0000		N/A
QVZ 2 T4	369,338,993	0.4412	0.8679	0.9994	0.9290	0.0003	-0.0001	0.0001		175,908,797	0.4664	0.8916	0.9994	0.9424	0.0003	0.0000	0.0002		N/A
QVZ 2 T8	259,975,714	0.3106	0.8679	0.9994	0.9290	0.0003	-0.0001	0.0001		125,973,247	0.3340	0.8912	0.9993	0.9422	-0.0001	-0.0001	-0.0001		N/A
QVZ 2 T16	152,385,057	0.1820	0.8687	0.9990	0.9293	0.0011	-0.0005	0.0004		74,727,311	0.1981	0.8921	0.9990	0.9425	0.0008	-0.0004	0.0003		N/A
QScomp (+ bzip2 -9)	447,290,216	0.5344	0.8681	0.9995	0.9292	0.0005	0.0000	0.0003		210,050,219	0.5569	0.8915	0.9995	0.9424	0.0002	0.0001	0.0002		dim1
Crumble -1 (+ CRAM)	328,644,318	0.3926	0.8658	0.9995	0.9279	-0.0018	0.0000	-0.0010		147,471,578	0.3910	0.8881	0.9995	0.9405	-0.0032	0.0001	-0.0017		N/A
Crumble -9 (+ CRAM)	246,283,161	0.2942	0.8654	0.9994	0.9276	-0.0022	-0.0001	-0.0013		104,264,203	0.2765	0.8877	0.9994	0.9402	-0.0036	0.0000	-0.0020		N/A
Quartz (+ bzip2)	625,378,135	0.7471	0.8762	0.9993	0.9337	0.0086	-0.0002	0.0048		287,310,936	0.7618	0.9024	0.9993	0.9484	0.0111	-0.0001	0.0061		N/A

Figure 7: Variant calling results for the GATK + hard filtration pipeline.

5.3 Platypus

H01 (ERR174324)

	Chromosome 11									Chromosome 20								Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences			
						R	P	F							R	P	F	
Uncompressed	1,986,669,293	8.0000	0.7759	0.9984	0.8732	0.0000	0.0000	0.0000		880,786,761	8.0000	0.7947	0.9981	0.8849	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	180,808,513	0.7281	0.7755	0.9984	0.8729	-0.0004	0.0000	-0.0003		84,415,660	0.7667	0.7946	0.9982	0.8848	-0.0001	0.0001	0.0000	N/A
QVZ 2 T2	93,357,561	0.3759	0.7761	0.9984	0.8733	0.0002	0.0000	0.0001		46,732,315	0.4245	0.7951	0.9981	0.8851	0.0004	0.0000	0.0002	N/A
QVZ 2 T4	32,931,170	0.1326	0.7771	0.9974	0.8736	0.0012	-0.0010	0.0004		17,191,826	0.1561	0.7961	0.9971	0.8853	0.0014	-0.0010	0.0005	N/A
QVZ 2 T8	9,832,950	0.0396	0.7781	0.9922	0.8722	0.0022	-0.0062	-0.0010		5,722,879	0.0520	0.7976	0.9911	0.8839	0.0029	-0.0070	-0.0010	N/A
QVZ 2 T16	3,480,969	0.0140	0.7784	0.9837	0.8691	0.0025	-0.0147	-0.0041		1,879,305	0.0171	0.7978	0.9806	0.8798	0.0031	-0.0175	-0.0051	N/A
QScomp (+ bzip2 -9)	36,732,800	0.1479	0.7751	0.9985	0.8727	-0.0008	0.0001	-0.0005		18,097,457	0.1644	0.7939	0.9983	0.8844	-0.0008	0.0002	-0.0004	dim1
Crumble -1 (+ CRAM)	100,830,366	0.4060	0.7760	0.9985	0.8733	0.0001	0.0001	0.0001		44,840,996	0.4073	0.7948	0.9983	0.8850	0.0001	0.0002	0.0001	N/A
Crumble -9 (+ CRAM)	58,407,264	0.2352	0.7771	0.9987	0.8741	0.0012	0.0003	0.0009		23,642,566	0.2147	0.7961	0.9985	0.8859	0.0014	0.0004	0.0010	N/A
Quartz (+ bzip2)	111,880,409	0.4505	0.7829	0.9984	0.8776	0.0070	0.0000	0.0044		46,634,066	0.4236	0.8024	0.9982	0.8897	0.0077	0.0001	0.0048	N/A

H11 (SRR1238539)

	Chromosome 11									Chromosome 20								Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences			
						R	P	F							R	P	F	
Uncompressed	1,402,713,277	8.0000	0.0660	0.9616	0.1235	0.0000	0.0000	0.0000		622,751,899	8.0000	0.0605	0.9599	0.1138	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	346,716,840	1.9774	0.0656	0.9621	0.1228	-0.0004	0.0005	-0.0007		155,988,650	2.0039	0.0602	0.9601	0.1133	-0.0003	0.0002	-0.0005	N/A
QVZ 2 T2	273,266,584	1.5585	0.0661	0.9606	0.1237	0.0001	-0.0010	0.0002		122,833,357	1.5779	0.0605	0.9601	0.1138	0.0000	0.0002	0.0000	N/A
QVZ 2 T4	198,709,595	1.1333	0.0661	0.9612	0.1237	0.0001	-0.0004	0.0002		89,402,714	1.1485	0.0609	0.9599	0.1145	0.0004	0.0000	0.0007	N/A
QVZ 2 T8	121,742,167	0.6943	0.0664	0.9590	0.1242	0.0004	-0.0026	0.0007		55,420,782	0.7119	0.0609	0.9590	0.1145	0.0004	-0.0009	0.0007	N/A
QVZ 2 T16	60,751,253	0.3465	0.0670	0.9578	0.1252	0.0010	-0.0038	0.0017		28,272,006	0.3632	0.0615	0.9576	0.1156	0.0010	-0.0023	0.0018	N/A
QScomp (+ bzip2 -9)	132,239,998	0.7542	0.0632	0.9612	0.1186	-0.0028	-0.0004	-0.0049		59,339,662	0.7623	0.0582	0.9615	0.1098	-0.0023	0.0016	-0.0041	dim1
Crumble -1 (+ CRAM)	642,683,176	3.6654	0.0660	0.9617	0.1235	0.0000	0.0001	0.0000		286,005,859	3.6741	0.0605	0.9601	0.1138	0.0000	0.0002	0.0000	N/A
Crumble -9 (+ CRAM)	570,828,844	3.2556	0.0660	0.9616	0.1235	0.0000	0.0000	0.0000		254,275,065	3.2665	0.0606	0.9593	0.1140	0.0001	-0.0006	0.0002	N/A
Quartz (+ bzip2)	247,127,738	1.4094	0.0635	0.9629	0.1191	-0.0025	0.0013	-0.0044		108,111,613	1.3888	0.0595	0.9659	0.1121	-0.0010	0.0060	-0.0017	N/A

H12 (Garvan replicate)

	Chromosome 11									Chromosome 20								Remarks
	Size (B)	Bits/QS	R	P	F	Differences				Size (B)	Bits/QS	R	P	F	Differences			
						R	P	F							R	P	F	
Uncompressed	6,696,582,269	8.0000	0.7248	0.9996	0.8403	0.0000	0.0000	0.0000		3,017,194,085	8.0000	0.7297	0.9997	0.8436	0.0000	0.0000	0.0000	N/A
QVZ 2 T1	814,504,450	0.9730	0.7239	0.9996	0.8397	-0.0009	0.0000	-0.0006		383,733,965	1.0175	0.7287	0.9997	0.8430	-0.0010	0.0000	-0.0007	N/A
QVZ 2 T2	583,280,511	0.6968	0.7234	0.9996	0.8394	-0.0014	0.0000	-0.0009		277,566,711	0.7360	0.7276	0.9997	0.8422	-0.0021	0.0000	-0.0014	N/A
QVZ 2 T4	369,338,993	0.4412	0.7201	0.9996	0.8371	-0.0047	0.0000	-0.0032		175,908,797	0.4664	0.7229	0.9997	0.8391	-0.0068	0.0000	-0.0046	N/A
QVZ 2 T8	259,975,714	0.3106	0.7234	0.9996	0.8394	-0.0014	0.0000	-0.0009		125,973,247	0.3340	0.7270	0.9997	0.8418	-0.0027	0.0000	-0.0018	N/A
QVZ 2 T16	152,385,057	0.1820	0.7276	0.9995	0.8421	0.0028	-0.0001	0.0018		74,727,311	0.1981	0.7332	0.9996	0.8459	0.0035	-0.0001	0.0023	N/A
QScomp (+ bzip2 -9)	447,290,216	0.5344	0.7290	0.9995	0.8431	0.0042	-0.0001	0.0028		210,050,219	0.5569	0.7353	0.9996	0.8473	0.0056	-0.0001	0.0037	dim1
Crumble -1 (+ CRAM)	328,644,318	0.3926	0.7289	0.9996	0.8431	0.0041	0.0000	0.0027		147,471,578	0.3910	0.7352	0.9997	0.8473	0.0055	0.0000	0.0037	N/A
Crumble -9 (+ CRAM)	246,283,161	0.2942	0.7297	0.9996	0.8436	0.0049	0.0000	0.0033		104,264,203	0.2765	0.7363	0.9997	0.8480	0.0066	0.0000	0.0044	N/A
Quartz (+ bzip2)	625,378,135	0.7471	0.7640	0.9995	0.8660	0.0392	-0.0001	0.0257		287,310,936	0.7618	0.7754	0.9997	0.8734	0.0457	0.0000	0.0298	N/A

Figure 8: Variant calling results for the Platypus pipeline.

6 Compression ratios results

In this section, we show in Figure 9 the obtained compression ratios for all tools. We also show the results for the compression of the quality scores with gzip and bzip2.

H01 (ERR174324)					
	Chromosome 11		Chromosome 20		Remarks
	Size (B)	Bits/QS	Size (B)	Bits/QS	
Uncompressed	1,986,669,293	8.0000	880,786,761	8.0000	N/A
Entropy	N/A	3.1100	N/A	3.3800	N/A
gzip	769,263,571	3.0977	346,169,589	3.1442	N/A
bzip2 -9	711,098,163	2.8635	320,181,485	2.9081	N/A
QVZ 2 T1	180,808,513	0.7281	84,415,660	0.7667	N/A
QVZ 2 T2	93,357,561	0.3759	46,732,315	0.4245	N/A
QVZ 2 T4	32,931,170	0.1326	17,191,826	0.1561	N/A
QVZ 2 T8	9,832,950	0.0396	5,722,879	0.0520	N/A
QVZ 2 T16	3,480,969	0.0140	1,879,305	0.0171	N/A
QScomp dim1 (+ bzip2 -9)	36,732,800	0.1479	18,097,457	0.1644	N/A
QScomp dim1 and dim2.* (+ bzip2 -9)	705,912,327	2.8426	317,849,142	2.8870	N/A
QScomp dim1 and dim2_a (+ bzip2 -9)	721,242,819	2.9043	325,382,531	2.9554	N/A
Crumble -1 (+ CRAM)	100,830,366	0.4060	44,840,996	0.4073	N/A
Crumble -9 (+ CRAM)	58,407,264	0.2352	23,642,566	0.2147	N/A
Quartz (+ bzip2)	111,880,409	0.4505	46,634,066	0.4236	N/A

H11 (SRR1238539)					
	Chromosome 11		Chromosome 20		Remarks
	Size (B)	Bits/QS	Size (B)	Bits/QS	
Uncompressed	1,402,713,277	8.0000	622,751,899	8.0000	N/A
Entropy	N/A	4.3700	N/A	4.3800	N/A
gzip	777,249,613	4.4328	346,098,789	4.4461	N/A
bzip2 -9	710,043,433	4.0495	316,314,730	4.0634	N/A
QVZ 2 T1	346,716,840	1.9774	155,988,650	2.0039	N/A
QVZ 2 T2	273,266,584	1.5585	122,833,357	1.5779	N/A
QVZ 2 T4	198,709,595	1.1333	89,402,714	1.1485	N/A
QVZ 2 T8	121,742,167	0.6943	55,420,782	0.7119	N/A
QVZ 2 T16	60,751,253	0.3465	28,272,006	0.3632	N/A
QScomp dim1 (+ bzip2 -9)	132,239,998	0.7542	59,339,662	0.7623	N/A
QScomp dim1 and dim2.* (+ bzip2 -9)	737,175,434	4.2043	328,273,753	4.2171	N/A
QScomp dim1 and dim2_a (+ bzip2 -9)	789,666,804	4.5037	351,714,573	4.5182	N/A
Crumble -1 (+ CRAM)	642,683,176	3.6654	286,005,859	3.6741	N/A
Crumble -9 (+ CRAM)	570,828,844	3.2556	254,275,065	3.2665	N/A
Quartz (+ bzip2)	247,127,738	1.4094	108,111,613	1.3888	N/A

H12 (Garvan replicate)					
	Chromosome 11		Chromosome 20		Remarks
	Size (B)	Bits/QS	Size (B)	Bits/QS	
Uncompressed	6,696,582,269	8.0000	3,017,194,085	8.0000	N/A
Entropy	N/A	3.3300	N/A	3.1600	N/A
gzip	2,544,194,141	3.0394	1,169,607,791	3.1012	N/A
bzip2 -9	2,392,635,823	2.8583	1,099,838,401	2.9162	N/A
QVZ 2 T1	814,504,450	0.9730	383,733,965	1.0175	N/A
QVZ 2 T2	583,280,511	0.6968	277,566,711	0.7360	N/A
QVZ 2 T4	369,338,993	0.4412	175,908,797	0.4664	N/A
QVZ 2 T8	259,975,714	0.3106	125,973,247	0.3340	N/A
QVZ 2 T16	152,385,057	0.1820	74,727,311	0.1981	N/A
QScomp dim1 (+ bzip2 -9)	447,290,216	0.5344	210,050,219	0.5569	N/A
QScomp dim1 and dim2.* (+ bzip2 -9)	2,465,941,916	2.9459	1,133,452,875	3.0053	N/A
QScomp dim1 and dim2_a (+ bzip2 -9)	2,626,334,794	3.1375	1,208,519,425	3.2044	N/A
Crumble -1 (+ CRAM)	328,644,318	0.3926	147,471,578	0.3910	N/A
Crumble -9 (+ CRAM)	246,283,161	0.2942	104,264,203	0.2765	N/A
Quartz (+ bzip2)	625,378,135	0.7471	287,310,936	0.7618	N/A

Figure 9: Compression ratios results.

References

[Bon14] James K Bonfield. The Scramble conversion tool. *Bioinformatics*, 30(19):2818–2819, 2014.

[HOW16] Mikel Hernaez, Idoia Ochoa, and Tsachy Weissman. A Cluster-Based Approach to Compression of Quality Scores. In *2016 Data Compression Conference (DCC)*, pages 261–270, 2016.

[LHW+09] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[LS12] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[MHB+10] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[RPM+14] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918, 2014.

[YYPB15] Y William Yu, Deniz Yorukoglu, Jian Peng, and Bonnie Berger. Quality score compression improves genotyping accuracy. *Nature Biotechnology*, 33(3):240–243, 2015.