CALQ: compression of quality values of aligned sequencing data

Supplementary Data

Jan Voges, Jörn Ostermann and Mikel Hernaez

Contents

1 Tools

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

Table 1: Tools selected for the evaluation.

Tool name	**Version**	**Mode**	**Lossless (Y/N)**	**Lossy (Y/N)**
bgzip	1.3	default	Y	N
CALQ	0b74e3d	default	N	Y
Crumble (+ CRAM)	0.5	-1	N	Y
		-9	N	Y
DSRC IL8B (+ gzip)	2	default	N	Y
P-Block	da36a12	p=1	N	Y
		p=4	N	Y
Quartz (+ bzip2)	0.2.2	default	N	Y
QVZ 2	d5383c6	T1	N	Y
		T2	N	Y
		T4	N	Y
		T8	N	Y
		T16	N	Y
R-Block	da36a12	r=5	N	Y
		r=20	N	Y

1.1 bgzip

Bgzip is part of the HTSlib repository, which in turn is part of the SAMtools program suite. HTSlib 1.3 was downloaded from http://www.htslib.org/. The extraction of the quality values of a SAM file `file.sam` was performed with the following command.

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

Subsequently, the compression of the quality values with bgzip was performed with the following command.

```
$ bgzip -c file.sam.qual 1> file.sam.qual.bgz
```

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

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

1.2 CALQ

CALQ can be downloaded from https://github.com/voges/calq. Build and usage information is available on the respective website.

CALQ processes the input data in fixed-size blocks, where each block contains b sequence alignments. We empirically derived a suitable block size using dataset B16 from Table 2. Figure 1 shows the CALQ compression ratio in bits per quality value versus the block size in number of sequence alignments for this dataset. As seen in the figure, the compression ratio asymptotically approaches a limit at around 0.29 bits per quality value. With CALQ, we want to give the user the opportunity to quickly perform selective access on the compressed data. Therefore, we chose a block size of $b = 10,000$ sequence alignments for our further experiments.

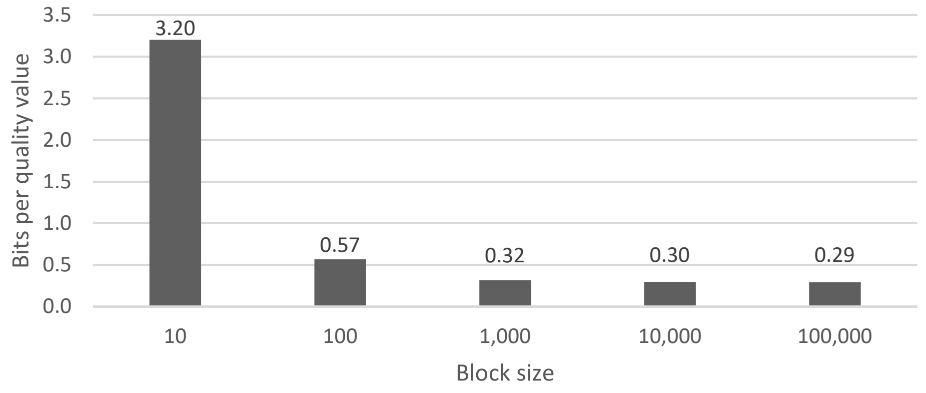


Figure 1: Bits per quality value versus block size in number of sequence alignments.

The following command was used to perform compression of the quality values in a SAM file `file.sam` with CALQ. The string given to the argument `-q` indicates the quality value type (i.e., quality value range) to be assumed by CALQ. Furthermore, the argument `-p2` conveys to the CALQ encoder the information that the sequenced organism possesses a diploid chromosome set. Finally, with `-b10000`, the CALQ encoder is instructed to process the input data in blocks which may contain at most 10,000 sequence alignments. The compressed quality value information is stored in the output file `file.sam.cq`.

```
$ calq -q Illumina-1.8+ -p 2 -b 10000 file.sam -o file.sam.cq
```

The following command was used to perform the decompression of the compressed quality values stored in the file `file.sam.cq`. To perform the decompression procedure, the CALQ decoder requires the alignment information, namely the mapping positions (POS), the CIGAR strings, and the reference sequence name(s) (RNAME). We pass this information to the CALQ decoder with the argument `-sfile.sam`. The reconstructed quality values are stored in the output file `file.sam.cq.qual`.

```
$ calq -d -s file.sam file.sam.cq -o file.sam.cq.qual
```

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

```
$ python replace_qual_sam.py file.sam file.sam.cq.qual
```

The CALQ codec structure is shown in Figure 2.

1.3 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
```

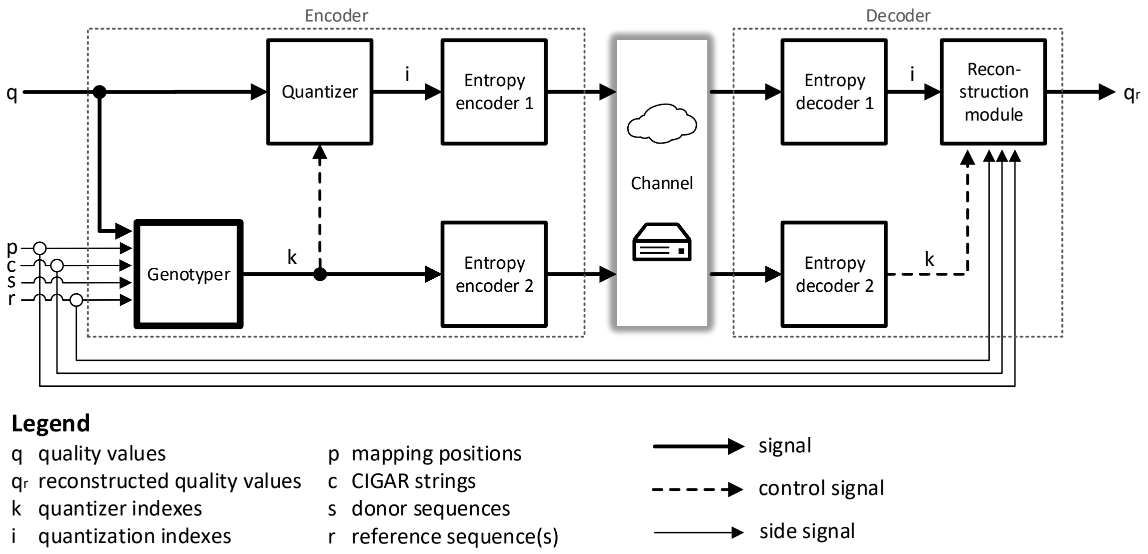


Figure 2: CALQ codec structure. The encoder gets as input quality values q, mapping positions p, CIGAR strings c, nucleotide sequences s, and optionally the reference sequence(s) r. The computation of the index k (as a function of the "genotype uncertainty") is performed by the Genotyper block. The Quantizer block uses then the index k to select a specific quantizer from a previously computed set of quantizers, which transforms the input quality values into quantization indexes i. Both indexes k and i are encoded by the two entropy encoder blocks, respectively. For decompression, first, indexes k and indexes i are decompressed. Then, after the alignment information, i.e., the mapping positions, the CIGAR strings, and the reference sequence name(s), has been transmitted as side information to the decoder, the reconstruction module reconstructs the quality values q_r.

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.

1.4 DSRC

We used DSRC [DG11, RD14] to mimic the 8-level mapping introduced by Illumina. The specific mapping performed by DSRC may differ slightly from the actual Illumina binning, as the latter might depend on the sequencing machine.

DSRC was downloaded from https://github.com/refresh-bio/DSRC. The compression of quality values in a FASTQ file `file.fastq` was performed with the following command.

```
$ dsrc c -d3 -q2 -b256 -l file.fastq file.fastq.dsrc
```

The argument `c` invokes the compression mode. The argument `-d3` denotes the usage of the best DNA compression mode (not in the scope of this manuscript). The argument `-q2` denotes the best quality compression mode. The argument `-b256` tells DSRC to use the (best) FASTQ input buffer size in MB. The argument `-l` finally invokes the lossy quality value compression mode (i.e., the Illumina binning scheme).

The decompression of the DSRC bitstream was performed with the following command.

```
$ dsrc d file.fastq.dsrc file.fastq.dsrc.fastq
```

Subsequently, the modified quality values were extracted from the file `file.fastq.dsrc.fastq` with the following command.

```
$ python xtract_part_fastq.py file.fastq.dsrc.fastq 3 1> file.fastq.dsrc.
   fastq.qual
```

Finally, the modified quality values were compressed using gzip, as recommended by the DSRC authors using the following command.

```
$ gzip -9 -c file.fastq.dsrc.fastq.qual > file.fastq.dsrc.fastq.qual.gz
```

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

```
$ wc -c file.fastq.dsrc.fastq.qual.gz
```

1.5 P-/R-Block

The P- and R-Block algorithms [CMT14] represent quality values by separating them into blocks of variable size, where all quality values contained in each block do not violate a given distortion constraint. For each block, its length and the representative value is losslessly compressed and stored.

Let $Q_{\max}$ and $Q_{\min}$ denote the largest and the smallest quality values within a block, respectively. In P-Block, all quality values in a block should then satisfy $Q_{\max} - Q_{\min} \leq 2p$, where p is a user-specified parameter. In R-Block, the quality values are subject to the constraint $Q_{\max}/Q_{\min} \leq r^2$, with r being a user-specified parameter.

Whereas in P-Block the maximum absolute distance allowed between a quality value and it representative value is constant for every quality value, this distance in general be smaller for low quality values in R-Block. Hence, R-Block preserves low quality values more precisely than high ones.

P- and R-Block where downloaded from https://github.com/rcanovas/libCSAM (Git commit hash: da36a12).

The following command was used to compress the quality values in a SAM file `file.sam` with P-Block.

```
$ CompressQual file.sam -q 1 -m 1 -l $p
```

The authors of P- and R-Block tested the P-Block algorithm with the user-defined parameter $p \in \{1, 2, 4, 8, 16, 32\}$. To limit the amount of simulations, we used P-Block with the parameter $p \in \{1, 4\}$.

The following command was used to compress the quality values in a SAM file `file.sam` with R-Block.

```
$ CompressQual file.sam -q 2 -m 1 -l $r
```

The authors of P- and R-Block tested the R-Block algorithm with the user defined parameter $r \in 1 + \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$ (which corresponds to the command line parameters `r={5,10,20,40,80,160,320,640}`). To limit the amount of simulations, we used R-Block with the parameter $r \in 1 + \{0.05, 0.2\}$.

Both commands above produce a bitstream file named `file.sam.cqual`. The decompression was performed with the following command.

```
$ DecompressQual file.sam.cqual
```

This produces the file `file.sam.cqual.qual` containing the reconstructed quality values.

1.6 Quartz

A necessary pre-processing step for the working of Quartz is the generation of a dictionary of common k-mers for each species. Then, for a given set of sequence reads, Quartz breaks these reads up into a set of overlapping so-called supporting k-mers.

Subsequently, every position in a supporting k-mer different from a dictionary k-mer is annotated as a possible variant. Quartz assumes that these divergent bases in supporting k-mers correspond to sequencing errors or single nucleotide polymorphisms (SNPs), respectively. The corresponding quality values are preserved by Quartz, whereas other quality values are set to a pre-defined default value.

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
```

1.7 QVZ 2

QVZ 2 [HOW16] models the quality value sequence from each read as a Markov chain of order one, based on the observation that quality values are highly correlated with their neighbours. The transition probabilities are derived from the entire dataset to be compressed. Subsequently, these transition probabilites are used to compute a set of Lloyd-Max quantizers, indexed by the position within a read and the previously quantized value. Finally, an adaptive arithmetic encoder is used to compress the quantized quality values.

QVZ 2 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 squared 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 data from the "Updated database for Evaluation of Genomic Information Compression and Storage" [Joi16b] compiled by the Joint Ad-hoc Group on Genomic Information Compression and Storage (JAhG) between ISO/IEC JTC 1/SC 29/WG 11, also known as Moving Picture Experts Group (MPEG), and ISO/TC 276/WG 5. The selected datasets are shown in Table 2. For the dataset H01 we selected the first pair of FASTQ files, namely `ERR174324_1.fastq` and `ERR174324_2.fastq`.

Table 2: Datasets selected for the evaluation.

ID	Name	Species	Sequencing technology	Coverage
H01	ERR174324	H. sapiens	Illumina HiSeq 2000	14×
H11	SRR1238539	H. sapiens	Ion Torrent	10×
H12	Garvan replicate	H. sapiens	Illumina HiSeq X	49×
B16	DH10B	E. coli	Illumina	422×
I19	dm3	D. melanogaster	PacBio	100×

The data were downloaded from the following locations.

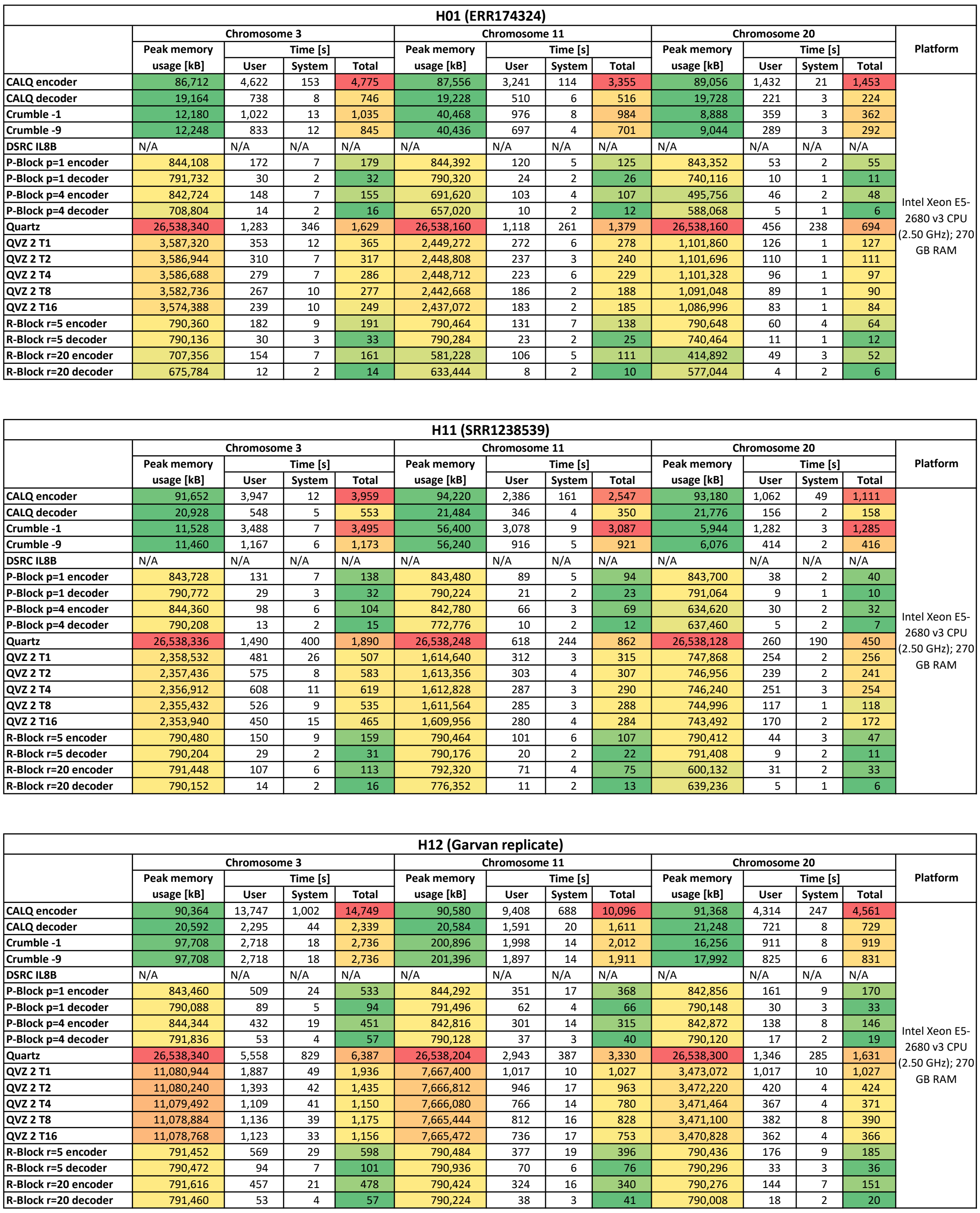
- H01: `http://www.ebi.ac.uk/ena/data/view/ERP001775/`
- H11: `ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA096/SRA096885/SRX517292/SRR1238539.fastq.bz2`
- H12, file 1: `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R1.fastq.gz`
- H12, file 2: `https://s3-ap-southeast-2.amazonaws.com/kccg-x10-truseq-nano-v2.5-na12878/NA12878_V2.5_Robot_2_R2.fastq.gz`
- B16: `ftp://webdata:webdata@ussd-ftp.illumina.com/Data/SequencingRuns/DH10B/MiSeq_Ecoli_DH10B_110721_PF.bam`
- I19: `http://bergmanlab.ls.manchester.ac.uk/data/tracks/dm3/dm3PacBio.bam`

3 Performance measurements

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

```
$ time -v -o file.sam.cq.time calq file.sam
```

The complete performance results for all tools and datasets are shown in Figure 3.



H01 (ERR174324)

	Chromosome 3				Chromosome 11				Chromosome 20				Platform
	Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			
		User	System	Total		User	System	Total		User	System	Total	
CALQ encoder	86,712	4,622	153	4,775	87,556	3,241	114	3,355	89,056	1,432	21	1,453	Intel Xeon E5-2680 v3 CPU (2.50 GHz); 270 GB RAM
CALQ decoder	19,164	738	8	746	19,228	510	6	516	19,728	221	3	224	
Crumble -1	12,180	1,022	13	1,035	40,468	976	8	984	8,888	359	3	362	
Crumble -9	12,248	833	12	845	40,436	697	4	701	9,044	289	3	292	
DSRC IL8B	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
P-Block p=1 encoder	844,108	172	7	179	844,392	120	5	125	843,352	53	2	55	
P-Block p=1 decoder	791,732	30	2	32	790,320	24	2	26	740,116	10	1	11	
P-Block p=4 encoder	842,724	148	7	155	691,620	103	4	107	495,756	46	2	48	
P-Block p=4 decoder	708,804	14	2	16	657,020	10	2	12	588,068	5	1	6	
Quartz	26,538,340	1,283	346	1,629	26,538,160	1,118	261	1,379	26,538,160	456	238	694	
QVZ 2 T1	3,587,320	353	12	365	2,449,272	272	6	278	1,101,860	126	1	127	
QVZ 2 T2	3,586,944	310	7	317	2,448,808	237	3	240	1,101,696	110	1	111	
QVZ 2 T4	3,586,688	279	7	286	2,448,712	223	6	229	1,101,328	96	1	97	
QVZ 2 T8	3,582,736	267	10	277	2,442,668	186	2	188	1,091,048	89	1	90	
QVZ 2 T16	3,574,388	239	10	249	2,437,072	183	2	185	1,086,996	83	1	84	
R-Block r=5 encoder	790,360	182	9	191	790,464	131	7	138	790,648	60	4	64	
R-Block r=5 decoder	790,136	30	3	33	790,284	23	2	25	740,464	11	1	12	
R-Block r=20 encoder	707,356	154	7	161	581,228	106	5	111	414,892	49	3	52	
R-Block r=20 decoder	675,784	12	2	14	633,444	8	2	10	577,044	4	2	6	

H11 (SRR1238539)

	Chromosome 3				Chromosome 11				Chromosome 20				Platform
	Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			
		User	System	Total		User	System	Total		User	System	Total	
CALQ encoder	91,652	3,947	12	3,959	94,220	2,386	161	2,547	93,180	1,062	49	1,111	Intel Xeon E5-2680 v3 CPU (2.50 GHz); 270 GB RAM
CALQ decoder	20,928	548	5	553	21,484	346	4	350	21,776	156	2	158	
Crumble -1	11,528	3,488	7	3,495	56,400	3,078	9	3,087	5,944	1,282	3	1,285	
Crumble -9	11,460	1,167	6	1,173	56,240	916	5	921	6,076	414	2	416	
DSRC IL8B	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
P-Block p=1 encoder	843,728	131	7	138	843,480	89	5	94	843,700	38	2	40	
P-Block p=1 decoder	790,772	29	3	32	790,224	21	2	23	791,064	9	1	10	
P-Block p=4 encoder	844,360	98	6	104	842,780	66	3	69	634,620	30	2	32	
P-Block p=4 decoder	790,208	13	2	15	772,776	10	2	12	637,460	5	2	7	
Quartz	26,538,336	1,490	400	1,890	26,538,248	618	244	862	26,538,128	260	190	450	
QVZ 2 T1	2,358,532	481	26	507	1,614,640	312	3	315	747,868	254	2	256	
QVZ 2 T2	2,357,436	575	8	583	1,613,356	303	4	307	746,956	239	2	241	
QVZ 2 T4	2,356,912	608	11	619	1,612,828	287	3	290	746,240	251	3	254	
QVZ 2 T8	2,355,432	526	9	535	1,611,564	285	3	288	744,996	117	1	118	
QVZ 2 T16	2,353,940	450	15	465	1,609,956	280	4	284	743,492	170	2	172	
R-Block r=5 encoder	790,480	150	9	159	790,464	101	6	107	790,412	44	3	47	
R-Block r=5 decoder	790,204	29	2	31	790,176	20	2	22	791,408	9	2	11	
R-Block r=20 encoder	791,448	107	6	113	792,320	71	4	75	600,132	31	2	33	
R-Block r=20 decoder	790,152	14	2	16	776,352	11	2	13	639,236	5	1	6	

H12 (Garvan replicate)

	Chromosome 3				Chromosome 11				Chromosome 20				Platform
	Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			Peak memory usage [kB]	Time [s]			
		User	System	Total		User	System	Total		User	System	Total	
CALQ encoder	90,364	13,747	1,002	14,749	90,580	9,408	688	10,096	91,368	4,314	247	4,561	Intel Xeon E5-2680 v3 CPU (2.50 GHz); 270 GB RAM
CALQ decoder	20,592	2,295	44	2,339	20,584	1,591	20	1,611	21,248	721	8	729	
Crumble -1	97,708	2,718	18	2,736	200,896	1,998	14	2,012	16,256	911	8	919	
Crumble -9	97,708	2,718	18	2,736	201,396	1,897	14	1,911	17,992	825	6	831	
DSRC IL8B	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
P-Block p=1 encoder	843,460	509	24	533	844,292	351	17	368	842,856	161	9	170	
P-Block p=1 decoder	790,088	89	5	94	791,496	62	4	66	790,148	30	3	33	
P-Block p=4 encoder	844,344	432	19	451	842,816	301	14	315	842,872	138	8	146	
P-Block p=4 decoder	791,836	53	4	57	790,128	37	3	40	790,120	17	2	19	
Quartz	26,538,340	5,558	829	6,387	26,538,204	2,943	387	3,330	26,538,300	1,346	285	1,631	
QVZ 2 T1	11,080,944	1,887	49	1,936	7,667,400	1,017	10	1,027	3,473,072	1,017	10	1,027	
QVZ 2 T2	11,080,240	1,393	42	1,435	7,666,812	946	17	963	3,472,220	420	4	424	
QVZ 2 T4	11,079,492	1,109	41	1,150	7,666,080	766	14	780	3,471,464	367	4	371	
QVZ 2 T8	11,078,884	1,136	39	1,175	7,665,444	812	16	828	3,471,100	382	8	390	
QVZ 2 T16	11,078,768	1,123	33	1,156	7,665,472	736	17	753	3,470,828	362	4	366	
R-Block r=5 encoder	791,452	569	29	598	790,484	377	19	396	790,436	176	9	185	
R-Block r=5 decoder	790,472	94	7	101	790,936	70	6	76	790,296	33	3	36	
R-Block r=20 encoder	791,616	457	21	478	790,424	324	16	340	790,276	144	7	151	
R-Block r=20 decoder	791,460	53	4	57	790,224	38	3	41	790,008	18	2	20	

Figure 3: Performance measurements results.

Figure 4 shows the peak memory usages of all tools.

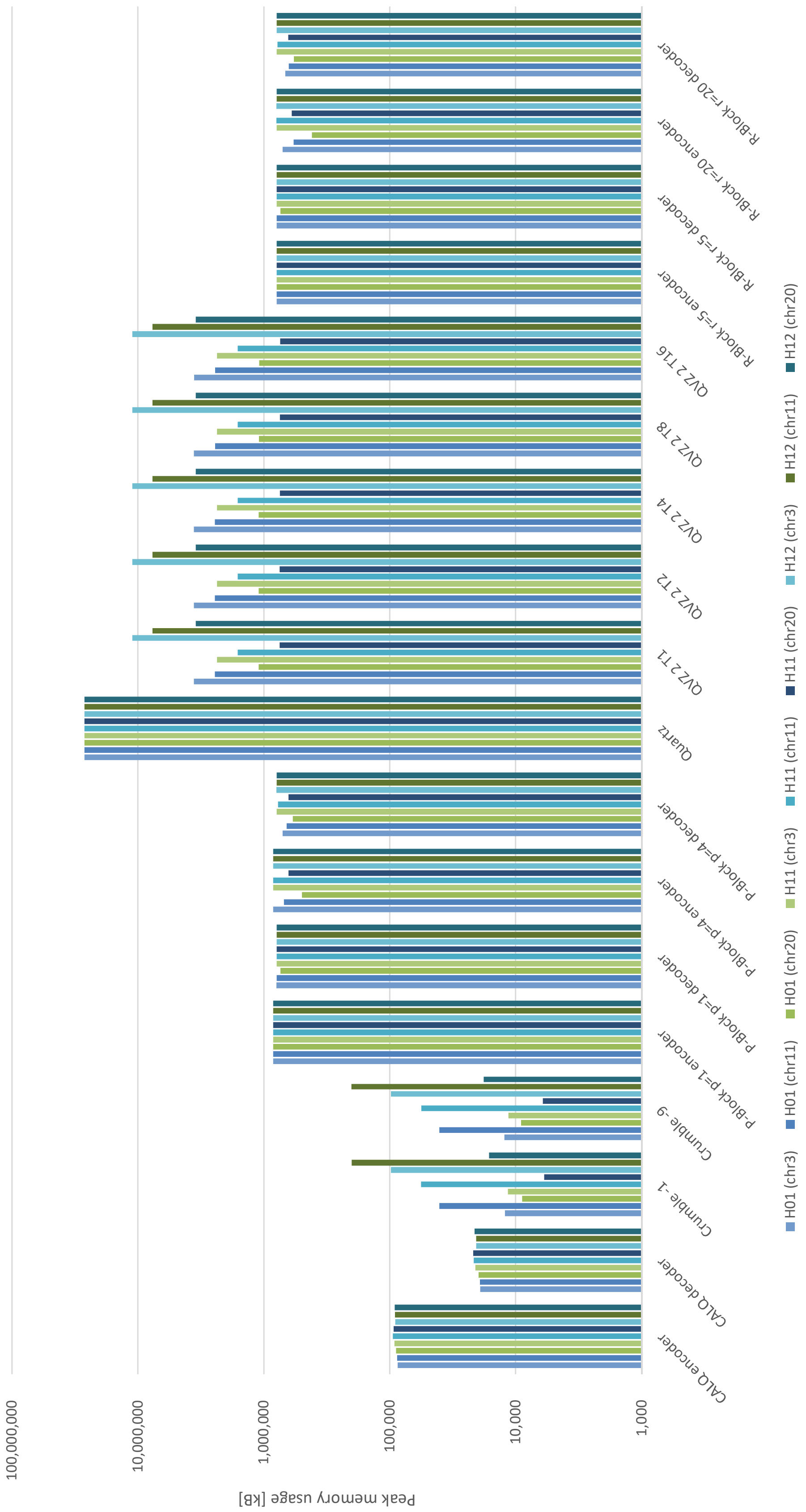


Figure 4: Peak memory usages.

Furthermore, the running times of the evaluated tools are shown in Figure 5.

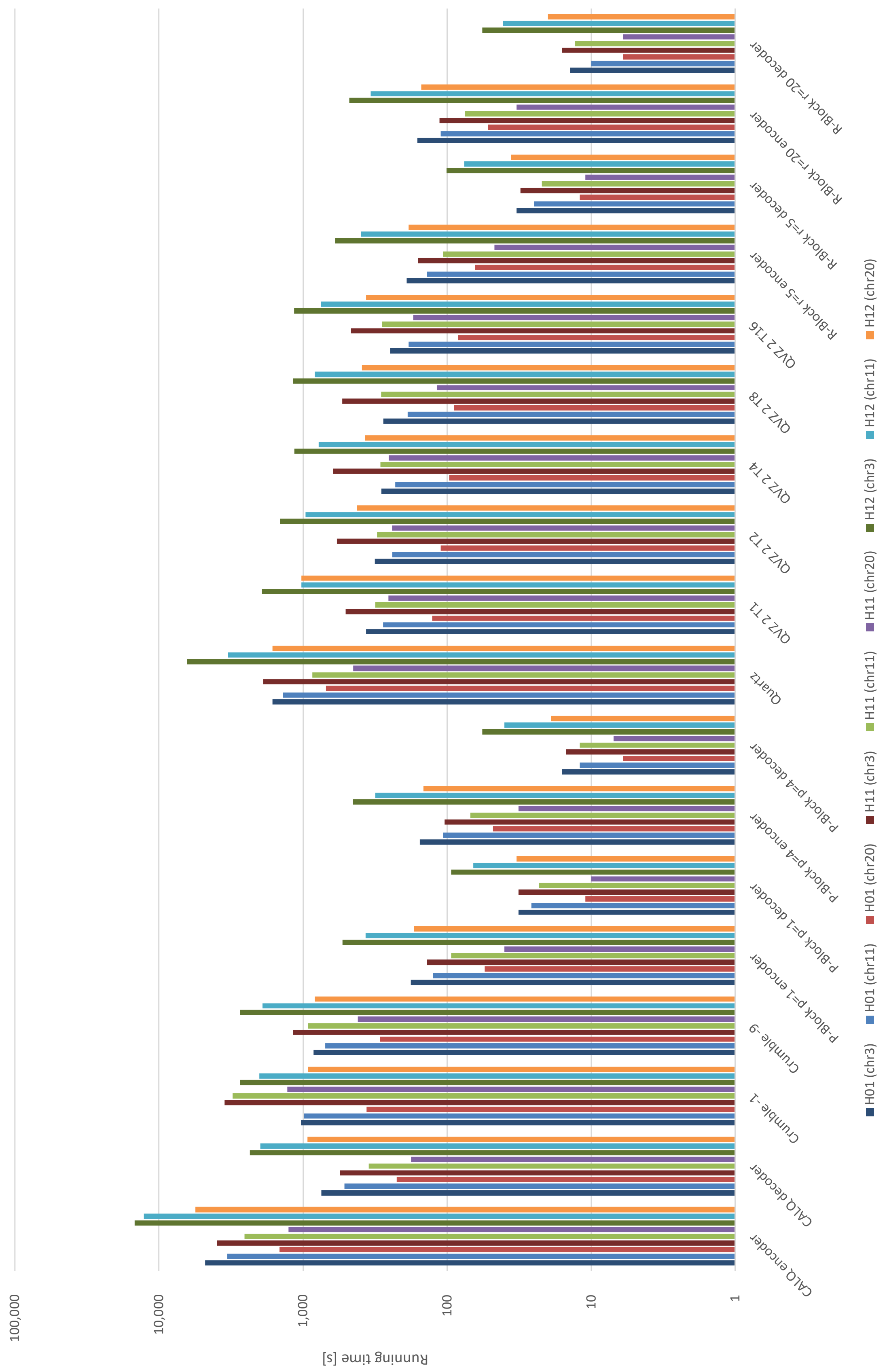


Figure 5: Running times.

4 Variant calling pipelines

This section provides information on the specific configurations of the variant calling pipelines used to assess the performance of CALQ 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. This specific pipeline was also proposed by the JAhG [Joi16a].
- 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.

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.

An overview of the used variant calling pipelines is given in Figure 6.

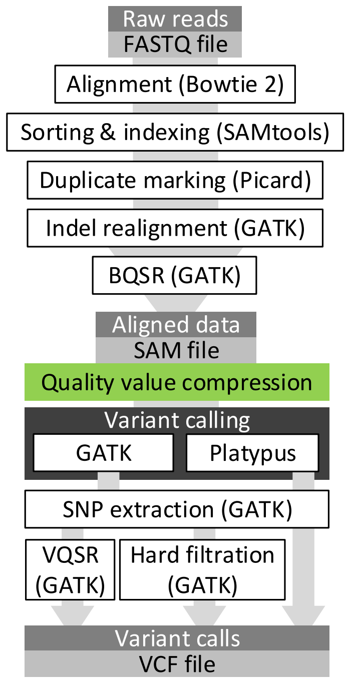


Figure 6: Variant calling pipelines.

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 on the datasets H01 and H11 from Table 2 with the GATK + VQSR pipeline. Here, in addition to the results for the datasets H01 and H11, we show the results for dataset H12. We averaged the Recall and Precision metrics over the chromosomes 3, 11 and 20 and over the four VQSR filter values ($\theta \in \{90, 99, 99.9, 100\}$) which in total yields 6 plots (i.e., 3 data sets $\times$ 2 metrics).

Also in this section, for the GATK + VQSR pipeline, we show a table containing the obtained Recall and Precision for the three chromosomes and the four VQSR filter values separately. In addition, the table shows the obtained F-scores. That is, we show 540 separate Recall, Precision, and F-score values (i.e., 3 data sets $\times$ 3 metrics $\times$ 4 filter values $\times$ (14+1) tools/configurations, including the results for the uncompressed data).

Furthermore, we show tables containing the Recall, Precision, and F-score results for the GATK + hard filtration pipeline and for the Platypus pipeline.

5.1 GATK + VQSR pipeline

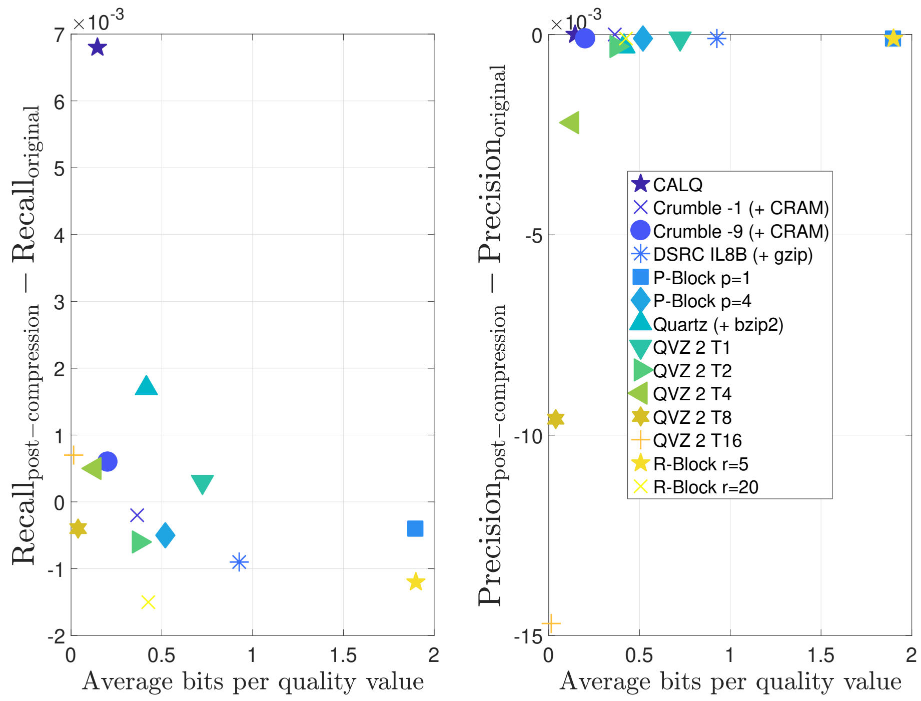


Figure 7: Recall and Precision results for the Illumina HiSeq 2000 data set ERR174324 with a coverage of 14×. The Recall and Precision metrics were averaged over the four VQSR filtering values as well as over all chromosomes 3, 11 and 20.

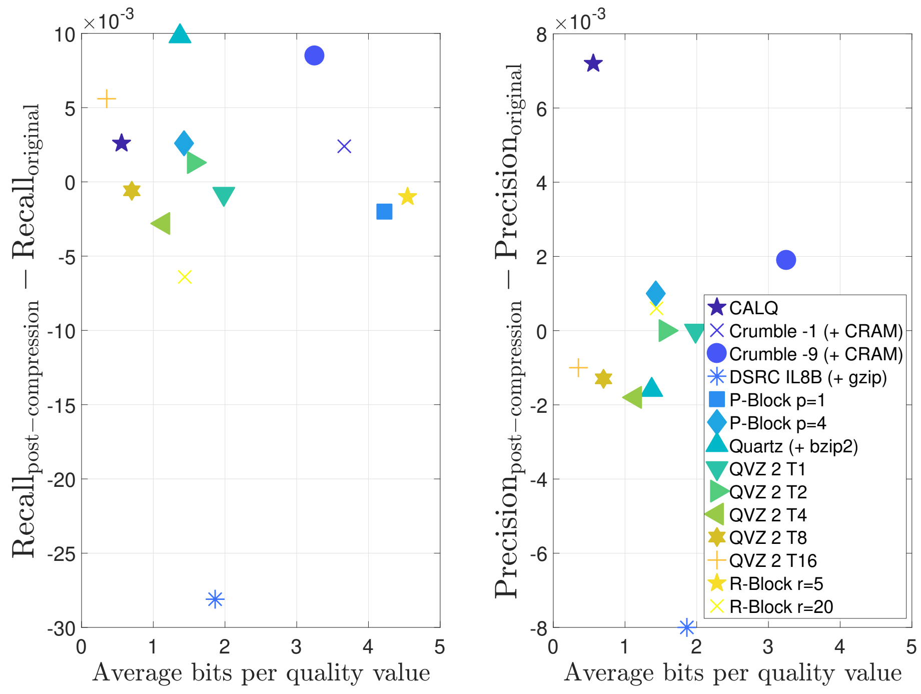


Figure 8: Recall and Precision results for the Ion Torrent data set SRR1238539 with a coverage of 10×. The Recall and Precision metrics were averaged over the four VQSR filtering values as well as over all chromosomes 3, 11 and 20.

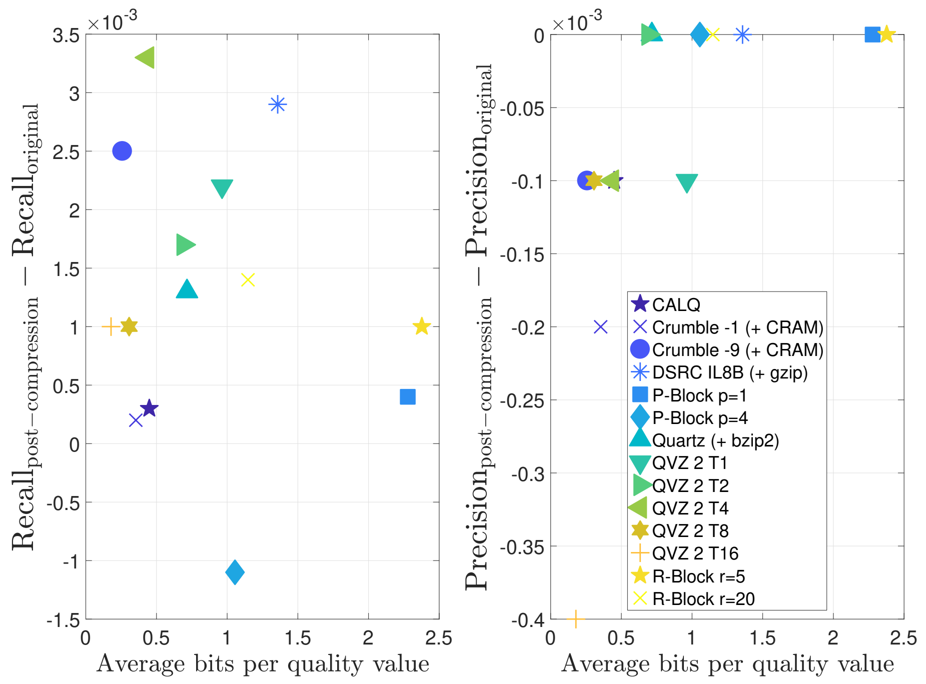
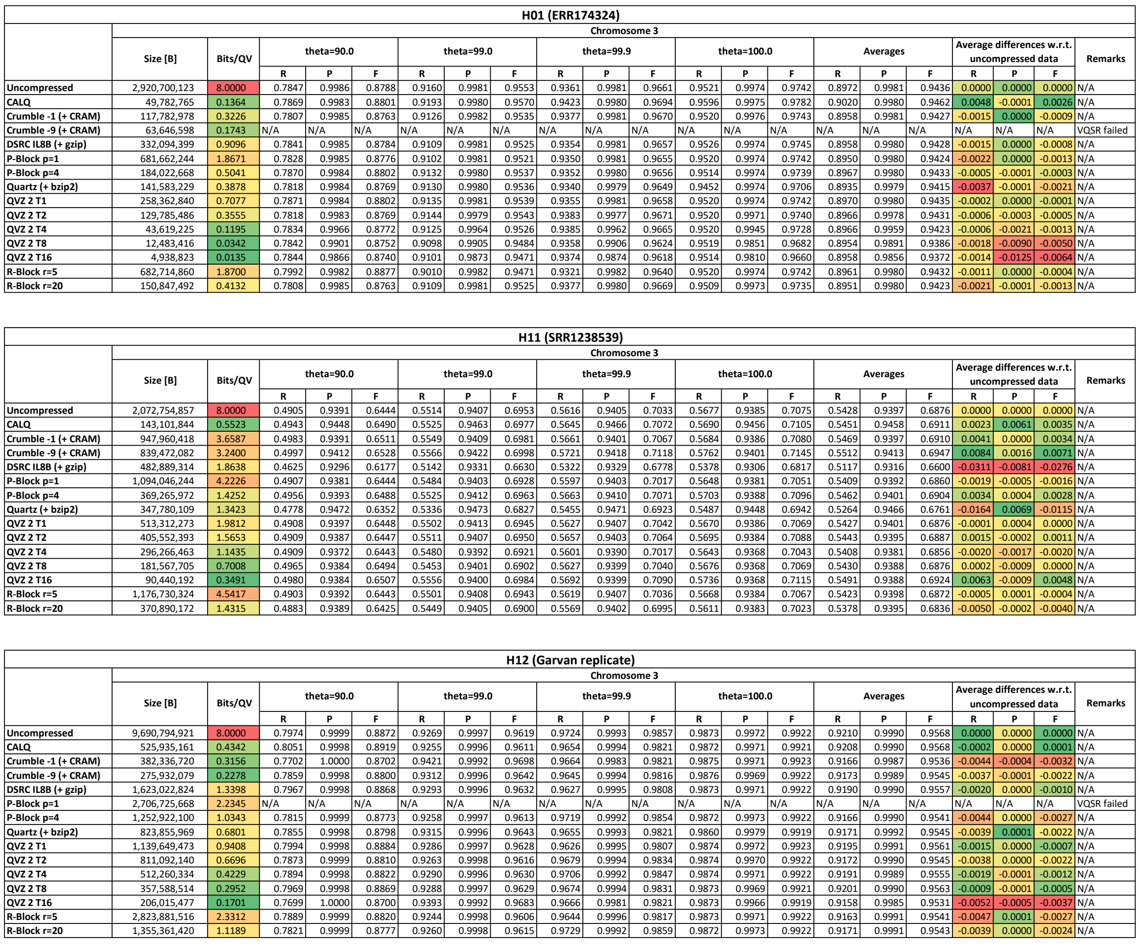


Figure 9: Recall and Precision results for the Illumina HiSeq X data set generated by the Garvan Institute with a coverage of 49×. The Recall and Precision metrics were averaged over the four VQSR filtering values as well as over all chromosomes 3, 11 and 20.



H01 (ERR174324) — Chromosome 3

	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	2,920,700,123	8.0000	0.7847	0.9986	0.8788	0.9160	0.9981	0.9553	0.9361	0.9981	0.9661	0.9521	0.9974	0.9742	0.8972	0.9981	0.9436	0.0000	0.0000	0.0000	N/A
CALQ	49,782,765	0.1364	0.7869	0.9983	0.8801	0.9193	0.9980	0.9570	0.9423	0.9980	0.9694	0.9596	0.9975	0.9782	0.9020	0.9980	0.9462	0.0048	-0.0001	0.0026	N/A
Crumble -1 (+ CRAM)	117,782,978	0.3226	0.7807	0.9985	0.8763	0.9126	0.9982	0.9535	0.9377	0.9981	0.9670	0.9520	0.9976	0.9743	0.8958	0.9981	0.9427	-0.0015	0.0000	-0.0009	N/A
Crumble -9 (+ CRAM)	63,646,598	0.1743	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
DSRC IL8B (+ gzip)	332,094,399	0.9096	0.7841	0.9985	0.8784	0.9109	0.9981	0.9525	0.9354	0.9981	0.9657	0.9526	0.9974	0.9745	0.8958	0.9980	0.9428	-0.0015	0.0000	-0.0008	N/A
P-Block p=1	681,662,244	1.8671	0.7828	0.9985	0.8776	0.9102	0.9981	0.9521	0.9350	0.9981	0.9655	0.9520	0.9974	0.9742	0.8950	0.9980	0.9424	-0.0022	0.0000	-0.0013	N/A
P-Block p=4	184,022,668	0.5041	0.7870	0.9984	0.8802	0.9132	0.9980	0.9537	0.9352	0.9980	0.9656	0.9514	0.9974	0.9739	0.8967	0.9980	0.9433	-0.0005	-0.0001	-0.0003	N/A
Quartz (+ bzip2)	141,583,229	0.3878	0.7818	0.9984	0.8769	0.9130	0.9980	0.9536	0.9340	0.9979	0.9649	0.9452	0.9974	0.9706	0.8935	0.9979	0.9415	-0.0037	-0.0001	-0.0021	N/A
QVZ 2 T1	258,362,840	0.7077	0.7871	0.9984	0.8802	0.9135	0.9981	0.9539	0.9355	0.9981	0.9658	0.9520	0.9974	0.9742	0.8970	0.9980	0.9435	-0.0002	0.0000	-0.0001	N/A
QVZ 2 T2	129,785,486	0.3555	0.7818	0.9983	0.8769	0.9144	0.9979	0.9543	0.9383	0.9977	0.9671	0.9520	0.9971	0.9740	0.8966	0.9978	0.9431	-0.0006	-0.0003	-0.0005	N/A
QVZ 2 T4	43,619,225	0.1195	0.7834	0.9966	0.8772	0.9125	0.9964	0.9526	0.9385	0.9962	0.9665	0.9520	0.9945	0.9728	0.8966	0.9959	0.9423	-0.0006	-0.0021	-0.0013	N/A
QVZ 2 T8	12,483,416	0.0342	0.7842	0.9901	0.8752	0.9098	0.9905	0.9484	0.9358	0.9906	0.9624	0.9519	0.9851	0.9682	0.8954	0.9891	0.9386	-0.0018	-0.0090	-0.0050	N/A
QVZ 2 T16	4,938,823	0.0135	0.7844	0.9866	0.8740	0.9101	0.9873	0.9471	0.9374	0.9874	0.9618	0.9514	0.9810	0.9660	0.8958	0.9856	0.9372	-0.0014	-0.0125	-0.0064	N/A
R-Block r=5	682,714,860	1.8700	0.7992	0.9982	0.8877	0.9010	0.9982	0.9471	0.9321	0.9982	0.9640	0.9520	0.9974	0.9742	0.8961	0.9980	0.9432	-0.0011	0.0000	-0.0004	N/A
R-Block r=20	150,847,492	0.4132	0.7808	0.9985	0.8763	0.9109	0.9981	0.9525	0.9377	0.9980	0.9669	0.9509	0.9973	0.9735	0.8951	0.9980	0.9423	-0.0021	-0.0001	-0.0013	N/A

H11 (SRR1238539) — Chromosome 3

	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	2,072,754,857	8.0000	0.4905	0.9391	0.6444	0.5514	0.9407	0.6953	0.5616	0.9405	0.7033	0.5677	0.9385	0.7075	0.5428	0.9397	0.6876	0.0000	0.0000	0.0000	N/A
CALQ	143,101,844	0.5523	0.4943	0.9448	0.6490	0.5525	0.9463	0.6977	0.5645	0.9466	0.7072	0.5690	0.9456	0.7105	0.5451	0.9458	0.6911	0.0023	0.0061	0.0035	N/A
Crumble -1 (+ CRAM)	947,960,418	3.6587	0.4983	0.9391	0.6511	0.5549	0.9409	0.6981	0.5661	0.9401	0.7067	0.5684	0.9386	0.7080	0.5469	0.9397	0.6910	0.0041	0.0000	0.0034	N/A
Crumble -9 (+ CRAM)	839,472,082	3.2400	0.4997	0.9412	0.6528	0.5566	0.9422	0.6998	0.5721	0.9418	0.7118	0.5762	0.9401	0.7145	0.5512	0.9413	0.6947	0.0084	0.0016	0.0071	N/A
DSRC IL8B (+ gzip)	482,889,314	1.8638	0.4625	0.9296	0.6177	0.5142	0.9331	0.6630	0.5322	0.9329	0.6778	0.5378	0.9306	0.6817	0.5117	0.9316	0.6600	-0.0311	-0.0081	-0.0276	N/A
P-Block p=1	1,094,046,244	4.2226	0.4907	0.9381	0.6444	0.5484	0.9403	0.6928	0.5597	0.9403	0.7017	0.5648	0.9381	0.7051	0.5409	0.9392	0.6860	-0.0019	-0.0005	-0.0016	N/A
P-Block p=4	369,265,972	1.4252	0.4956	0.9393	0.6488	0.5525	0.9412	0.6963	0.5663	0.9410	0.7071	0.5703	0.9388	0.7096	0.5462	0.9401	0.6904	0.0034	0.0004	0.0028	N/A
Quartz (+ bzip2)	347,780,109	1.3423	0.4778	0.9472	0.6352	0.5336	0.9473	0.6827	0.5455	0.9471	0.6923	0.5487	0.9448	0.6942	0.5264	0.9466	0.6761	-0.0164	0.0069	-0.0115	N/A
QVZ 2 T1	513,312,273	1.9812	0.4908	0.9397	0.6448	0.5502	0.9413	0.6945	0.5627	0.9407	0.7042	0.5670	0.9386	0.7069	0.5427	0.9401	0.6876	-0.0001	0.0004	0.0000	N/A
QVZ 2 T2	405,552,393	1.5653	0.4909	0.9387	0.6447	0.5511	0.9407	0.6950	0.5657	0.9403	0.7064	0.5695	0.9384	0.7088	0.5443	0.9395	0.6887	0.0015	-0.0002	0.0011	N/A
QVZ 2 T4	296,266,463	1.1435	0.4909	0.9372	0.6443	0.5480	0.9392	0.6921	0.5601	0.9390	0.7017	0.5643	0.9368	0.7043	0.5408	0.9381	0.6856	-0.0020	-0.0017	-0.0020	N/A
QVZ 2 T8	181,567,705	0.7008	0.4965	0.9384	0.6494	0.5453	0.9401	0.6902	0.5627	0.9399	0.7040	0.5676	0.9368	0.7069	0.5430	0.9388	0.6876	0.0002	-0.0009	0.0000	N/A
QVZ 2 T16	90,440,192	0.3491	0.4980	0.9384	0.6507	0.5556	0.9400	0.6984	0.5692	0.9399	0.7090	0.5736	0.9368	0.7115	0.5491	0.9388	0.6924	0.0063	-0.0009	0.0048	N/A
R-Block r=5	1,176,730,324	4.5417	0.4903	0.9392	0.6443	0.5501	0.9408	0.6943	0.5619	0.9407	0.7036	0.5668	0.9384	0.7067	0.5423	0.9398	0.6872	-0.0005	0.0001	-0.0004	N/A
R-Block r=20	370,890,172	1.4315	0.4883	0.9389	0.6425	0.5449	0.9405	0.6900	0.5569	0.9402	0.6995	0.5611	0.9383	0.7023	0.5378	0.9395	0.6836	-0.0050	-0.0002	-0.0040	N/A

H12 (Garvan replicate) — Chromosome 3

	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			Remarks
			R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	
Uncompressed	9,690,794,921	8.0000	0.7974	0.9999	0.8872	0.9269	0.9997	0.9619	0.9724	0.9993	0.9857	0.9873	0.9972	0.9922	0.9210	0.9990	0.9568	0.0000	0.0000	0.0000	N/A
CALQ	525,935,161	0.4342	0.8051	0.9998	0.8919	0.9255	0.9996	0.9611	0.9654	0.9994	0.9821	0.9872	0.9971	0.9921	0.9208	0.9990	0.9568	-0.0002	0.0000	0.0001	N/A
Crumble -1 (+ CRAM)	382,336,720	0.3156	0.7702	1.0000	0.8702	0.9421	0.9992	0.9698	0.9664	0.9983	0.9821	0.9875	0.9971	0.9923	0.9166	0.9987	0.9536	-0.0044	-0.0004	-0.0032	N/A
Crumble -9 (+ CRAM)	275,932,079	0.2278	0.7859	0.9998	0.8800	0.9312	0.9996	0.9642	0.9645	0.9994	0.9816	0.9876	0.9969	0.9922	0.9173	0.9989	0.9545	-0.0037	-0.0001	-0.0022	N/A
DSRC IL8B (+ gzip)	1,623,022,824	1.3398	0.7967	0.9998	0.8868	0.9293	0.9996	0.9632	0.9627	0.9995	0.9808	0.9873	0.9971	0.9922	0.9190	0.9990	0.9557	-0.0020	0.0000	-0.0010	N/A
P-Block p=1	2,706,725,668	2.2345	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
P-Block p=4	1,252,922,100	1.0343	0.7815	0.9999	0.8773	0.9258	0.9997	0.9613	0.9719	0.9992	0.9854	0.9872	0.9973	0.9922	0.9166	0.9990	0.9541	-0.0044	0.0000	-0.0027	N/A
Quartz (+ bzip2)	823,855,969	0.6801	0.7855	0.9998	0.8798	0.9315	0.9996	0.9643	0.9655	0.9993	0.9821	0.9860	0.9979	0.9919	0.9171	0.9992	0.9545	-0.0039	0.0001	-0.0022	N/A
QVZ 2 T1	1,139,649,473	0.9408	0.7994	0.9998	0.8884	0.9286	0.9997	0.9628	0.9626	0.9995	0.9807	0.9874	0.9972	0.9923	0.9195	0.9991	0.9561	-0.0015	0.0000	-0.0007	N/A
QVZ 2 T2	811,092,140	0.6696	0.7873	0.9999	0.8810	0.9263	0.9998	0.9616	0.9679	0.9994	0.9834	0.9874	0.9970	0.9922	0.9172	0.9990	0.9545	-0.0038	0.0000	-0.0022	N/A
QVZ 2 T4	512,260,334	0.4229	0.7894	0.9998	0.8822	0.9290	0.9996	0.9630	0.9706	0.9992	0.9847	0.9874	0.9971	0.9922	0.9191	0.9989	0.9555	-0.0019	-0.0001	-0.0012	N/A
QVZ 2 T8	357,588,514	0.2952	0.7969	0.9998	0.8869	0.9288	0.9997	0.9629	0.9674	0.9994	0.9831	0.9873	0.9969	0.9921	0.9201	0.9990	0.9563	-0.0009	-0.0001	-0.0005	N/A
QVZ 2 T16	206,015,477	0.1701	0.7699	1.0000	0.8700	0.9393	0.9992	0.9683	0.9666	0.9981	0.9821	0.9873	0.9966	0.9919	0.9158	0.9985	0.9531	-0.0052	-0.0005	-0.0037	N/A
R-Block r=5	2,823,881,516	2.3312	0.7889	0.9999	0.8820	0.9244	0.9998	0.9606	0.9644	0.9996	0.9817	0.9873	0.9971	0.9922	0.9163	0.9991	0.9541	-0.0047	0.0001	-0.0027	N/A
R-Block r=20	1,355,361,420	1.1189	0.7821	0.9999	0.8777	0.9260	0.9998	0.9615	0.9729	0.9992	0.9859	0.9872	0.9973	0.9922	0.9171	0.9991	0.9543	-0.0039	0.0000	-0.0024	N/A

Figure 10: Variant calling results for chromosome 3 for the GATK+VQSR pipeline.

H01 (ERR174324)																				
	Chromosome 11																			
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	37,826,773	0.1523	0.7925	0.9982	0.8835	0.9201	0.9978	0.9574	0.9404	0.9978	0.9683	0.9552	0.9975	0.9759	0.9021	0.9978	0.9463	0.0067	0.0001	0.0038	N/A
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
DSRC IL8B (+ gzip)	230,849,498	0.9296	0.7866	0.9982	0.8799	0.9124	0.9977	0.9531	0.9339	0.9977	0.9647	0.9475	0.9971	0.9717	0.8951	0.9977	0.9424	-0.0003	-0.0001	-0.0002	N/A
P-Block p=1	473,222,724	1.9056	0.7853	0.9982	0.8790	0.9141	0.9978	0.9541	0.9343	0.9977	0.9650	0.9471	0.9971	0.9715	0.8952	0.9977	0.9424	-0.0002	-0.0001	-0.0001	N/A
P-Block p=4	130,204,148	0.5243	0.7858	0.9982	0.8794	0.9123	0.9978	0.9531	0.9330	0.9977	0.9643	0.9463	0.9972	0.9711	0.8944	0.9977	0.9420	-0.0010	0.0000	-0.0005	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
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
R-Block	473,985,764	1.9087	0.7987	0.9980	0.8873	0.8991	0.9979	0.9459	0.9313	0.9978	0.9634	0.9471	0.9971	0.9715	0.8941	0.9977	0.9420	-0.0013	-0.0001	-0.0005	N/A
R-Block r=20	106,633,644	0.4294	0.7837	0.9982	0.8780	0.9104	0.9977	0.9521	0.9325	0.9977	0.9640	0.9457	0.9971	0.9707	0.8931	0.9977	0.9412	-0.0023	-0.0001	-0.0013	N/A

H11 (SRR1238539)																				
	Chromosome 11																			
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	98,589,364	0.5623	0.4917	0.9429	0.6463	0.5398	0.9456	0.6873	0.5529	0.9449	0.6976	0.5547	0.9445	0.6989	0.5348	0.9445	0.6825	0.0024	0.0059	0.0036	N/A
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
DSRC IL8B (+ gzip)	326,159,362	1.8602	0.4653	0.9294	0.6201	0.5150	0.9322	0.6635	0.5256	0.9318	0.6721	0.5277	0.9302	0.6734	0.5084	0.9309	0.6573	-0.0240	-0.0076	-0.0216	N/A
P-Block p=1	740,567,588	4.2236	0.4822	0.9376	0.6369	0.5369	0.9398	0.6834	0.5495	0.9390	0.6933	0.5528	0.9371	0.6954	0.5304	0.9384	0.6772	-0.0020	-0.0002	-0.0017	N/A
P-Block p=4	250,497,900	1.4286	0.4856	0.9390	0.6401	0.5420	0.9406	0.6877	0.5544	0.9399	0.6974	0.5578	0.9378	0.6995	0.5350	0.9393	0.6812	0.0026	0.0008	0.0023	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
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
R-Block r=5	797,235,396	4.5468	0.4779	0.9380	0.6332	0.5395	0.9397	0.6855	0.5512	0.9387	0.6946	0.5545	0.9371	0.6967	0.5308	0.9384	0.6775	-0.0016	-0.0002	-0.0014	N/A
R-Block r=20	252,440,580	1.4397	0.4762	0.9381	0.6317	0.5335	0.9399	0.6807	0.5458	0.9396	0.6905	0.5491	0.9373	0.6925	0.5262	0.9387	0.6738	-0.0062	0.0002	-0.0050	N/A

H12 (Garvan replicate)																				
	Chromosome 11																			
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	386,319,365	0.4615	0.7916	0.9997	0.8836	0.9305	0.9994	0.9637	0.9663	0.9991	0.9824	0.9858	0.9970	0.9914	0.9186	0.9988	0.9553	-0.0015	-0.0002	-0.0007	N/A
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
DSRC IL8B (+ gzip)	1,140,807,424	1.3629	0.7993	0.9998	0.8884	0.9381	0.9995	0.9678	0.9763	0.9988	0.9874	0.9860	0.9972	0.9916	0.9249	0.9988	0.9588	0.0049	-0.0002	0.0028	N/A
P-Block p=1	1,920,062,916	2.2938	0.7879	0.9999	0.8813	0.9337	0.9997	0.9656	0.9719	0.9990	0.9853	0.9861	0.9973	0.9917	0.9199	0.9990	0.9560	-0.0001	0.0000	0.0000	N/A
P-Block p=4	890,220,628	1.0635	0.7899	0.9999	0.8826	0.9303	0.9997	0.9638	0.9743	0.9991	0.9865	0.9860	0.9972	0.9916	0.9201	0.9990	0.9561	0.0001	0.0000	0.0001	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
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
R-Block r=5	2,004,430,612	2.3946	0.7890	0.9999	0.8820	0.9309	0.9997	0.9641	0.9749	0.9990	0.9868	0.9861	0.9973	0.9917	0.9202	0.9990	0.9561	0.0002	0.0000	0.0002	N/A
R-Block r=20	970,006,292	1.1588	0.8072	0.9998	0.8932	0.9309	0.9997	0.9641	0.9705	0.9993	0.9847	0.9860	0.9972	0.9916	0.9237	0.9990	0.9584	0.0036	0.0000	0.0024	N/A

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

H01 (ERR174324)																					
	Chromosome 20																				
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	17,545,800	0.1594	0.8085	0.9978	0.8932	0.9331	0.9977	0.9643	0.9534	0.9977	0.9750	0.9643	0.9972	0.9805	0.9148	0.9976	0.9533	0.0088	-0.0001	0.0049	N/A
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
DSRC IL8B (+ gzip)	107,172,747	0.9734	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
P-Block p=1	217,100,876	1.9719	0.8016	0.9979	0.8890	0.9265	0.9978	0.9608	0.9453	0.9977	0.9708	0.9551	0.9970	0.9756	0.9071	0.9976	0.9491	0.0011	-0.0001	0.0007	N/A
P-Block p=4	61,445,676	0.5581	0.8067	0.9978	0.8921	0.9200	0.9977	0.9573	0.9430	0.9976	0.9695	0.9543	0.9969	0.9751	0.9060	0.9975	0.9485	0.0000	-0.0002	0.0001	N/A
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
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
R-Block r=5	217,474,276	1.9753	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
R-Block r=20	50,270,852	0.4566	0.8061	0.9979	0.8918	0.9222	0.9978	0.9585	0.9414	0.9978	0.9688	0.9537	0.9970	0.9749	0.9059	0.9976	0.9485	-0.0002	-0.0001	0.0001	N/A

H11 (SRR1238539)																					
	Chromosome 20																				
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	44,145,089	0.5671	0.4728	0.9356	0.6282	0.5219	0.9393	0.6710	0.5340	0.9386	0.6807	0.5357	0.9374	0.6818	0.5161	0.9377	0.6654	0.0031	0.0095	0.0051	N/A
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
DSRC IL8B (+ gzip)	145,109,631	1.8641	0.4413	0.9187	0.5962	0.4908	0.9210	0.6404	0.5004	0.9212	0.6485	0.5026	0.9184	0.6497	0.4838	0.9198	0.6337	-0.0292	-0.0083	-0.0267	N/A
P-Block p=1	328,643,924	4.2218	0.4673	0.9284	0.6217	0.5183	0.9300	0.6656	0.5281	0.9295	0.6735	0.5301	0.9269	0.6745	0.5110	0.9287	0.6588	-0.0020	0.0005	-0.0015	N/A
P-Block p=4	112,184,956	1.4412	0.4698	0.9300	0.6243	0.5217	0.9317	0.6689	0.5330	0.9309	0.6779	0.5351	0.9280	0.6788	0.5149	0.9302	0.6624	0.0020	0.0020	0.0021	N/A
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
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
R-Block r=5	354,488,988	4.5538	0.4667	0.9283	0.6211	0.5193	0.9300	0.6665	0.5301	0.9291	0.6750	0.5322	0.9265	0.6761	0.5121	0.9285	0.6597	-0.0009	0.0003	-0.0007	N/A
R-Block r=20	113,848,684	1.4625	0.4590	0.9295	0.6145	0.5115	0.9313	0.6603	0.5238	0.9308	0.6704	0.5260	0.9279	0.6714	0.5051	0.9299	0.6542	-0.0079	0.0017	-0.0062	N/A

H12 (Garvan replicate)																					
	Chromosome 20																				
	Size [B]	Bits/QV	theta=90.0			theta=99.0			theta=99.9			theta=100.0			Averages			Average differences w.r.t. uncompressed data			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
CALQ	177,252,135	0.4700	0.8077	0.9996	0.8935	0.9474	0.9994	0.9727	0.9779	0.9992	0.9884	0.9918	0.9974	0.9946	0.9312	0.9989	0.9623	0.0025	0.0000	0.0014	N/A
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
DSRC IL8B (+ gzip)	528,911,763	1.4024	0.8307	0.9998	0.9074	0.9364	0.9996	0.9670	0.9785	0.9994	0.9888	0.9920	0.9974	0.9947	0.9344	0.9991	0.9645	0.0057	0.0001	0.0036	N/A
P-Block p=1	897,910,236	2.3808	0.8065	0.9999	0.8928	0.9434	0.9997	0.9707	0.9769	0.9991	0.9879	0.9919	0.9972	0.9945	0.9297	0.9990	0.9615	0.0010	0.0000	0.0006	N/A
P-Block p=4	416,754,260	1.1050	0.7991	0.9999	0.8883	0.9472	0.9997	0.9727	0.9803	0.9990	0.9896	0.9920	0.9975	0.9947	0.9297	0.9990	0.9613	0.0010	0.0001	0.0005	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
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
R-Block r=5	938,065,172	2.4873	0.8263	0.9998	0.9048	0.9432	0.9996	0.9706	0.9830	0.9993	0.9911	0.9920	0.9974	0.9947	0.9361	0.9990	0.9653	0.0074	0.0001	0.0044	N/A
R-Block r=20	458,073,844	1.2146	0.8114	0.9998	0.8958	0.9497	0.9997	0.9741	0.9799	0.9993	0.9895	0.9918	0.9973	0.9945	0.9332	0.9990	0.9635	0.0045	0.0001	0.0026	N/A

Figure 12: Variant calling results for chromosome 20 for the GATK + VQSR pipeline.

H01 (ERR174324)

	Chromosomes 3, 11 and 20					Remarks
	Size [B]	Bits/QV	Average differences w.r.t. uncompressed data			
			R	P	F	
Uncompressed	5,788,156,177	8.0000	0.0000	0.0000	0.0000	N/A
CALQ 0b74e3d	105,155,338	0.1453	0.0068	0.0000	0.0037	N/A
Crumble -1 (+ CRAM)	263,454,340	0.3641	-0.0002	0.0000	-0.0002	N/A
Crumble -9 (+ CRAM)	145,696,428	0.2014	0.0006	-0.0001	0.0003	VQSR failed for chromosomes 3 and 20
DSRC IL8B (+ gzip)	670,116,644	0.9262	-0.0009	-0.0001	-0.0005	VQSR failed for chromosome 20
P-Block p=1	1,371,985,844	1.8963	-0.0004	-0.0001	-0.0002	N/A
P-Block p=4	375,672,492	0.5192	-0.0005	-0.0001	-0.0002	N/A
Quartz (+ bzip2)	300,097,704	0.4148	0.0017	-0.0003	0.0009	N/A
QVZ 2 T1	523,587,013	0.7237	0.0003	-0.0001	0.0002	N/A
QVZ 2 T2	269,875,362	0.3730	-0.0006	-0.0003	-0.0004	N/A
QVZ 2 T4	93,742,221	0.1296	0.0005	-0.0022	-0.0006	N/A
QVZ 2 T8	28,039,245	0.0388	-0.0004	-0.0096	-0.0045	N/A
QVZ 2 T16	10,299,097	0.0142	0.0007	-0.0147	-0.0063	N/A
R-Block r=5	1,374,174,900	1.8993	-0.0012	-0.0001	-0.0004	VQSR failed for chromosome 20
R-Block r=20	307,751,988	0.4254	-0.0015	-0.0001	-0.0008	N/A

H11 (SRR1238539)

	Chromosomes 3, 11 and 20					Remarks
	Size [B]	Bits/QV	Average differences w.r.t. uncompressed data			
			R	P	F	
Uncompressed	4,098,220,033	8.0000	0.0000	0.0000	0.0000	N/A
CALQ 0b74e3d	285,836,297	0.5580	0.0026	0.0072	0.0041	N/A
Crumble -1 (+ CRAM)	1,876,649,453	3.6633	0.0024	0.0003	0.0021	N/A
Crumble -9 (+ CRAM)	1,664,575,991	3.2494	0.0085	0.0019	0.0073	VQSR failed for chromosome 20
DSRC IL8B (+ gzip)	954,158,307	1.8626	-0.0281	-0.0080	-0.0253	N/A
P-Block p=1	2,163,257,756	4.2228	-0.0020	0.0000	-0.0016	N/A
P-Block p=4	731,948,828	1.4288	0.0026	0.0010	0.0024	N/A
Quartz (+ bzip2)	703,019,460	1.3723	0.0098	-0.0016	0.0074	N/A
QVZ 2 T1	1,016,017,763	1.9833	-0.0008	0.0000	-0.0006	N/A
QVZ 2 T2	801,652,334	1.5649	0.0013	0.0000	0.0011	N/A
QVZ 2 T4	584,378,772	1.1407	-0.0028	-0.0018	-0.0027	N/A
QVZ 2 T8	358,730,654	0.7003	-0.0006	-0.0013	-0.0008	N/A
QVZ 2 T16	179,463,451	0.3503	0.0056	-0.0010	0.0043	N/A
R-Block r=5	2,328,454,708	4.5453	-0.0010	0.0001	-0.0008	N/A
R-Block r=20	737,179,436	1.4390	-0.0064	0.0006	-0.0051	N/A

H12 (Garvan replicate)

	Chromosomes 3, 11 and 20					Remarks
	Size [B]	Bits/QV	Average differences w.r.t. uncompressed data			
			R	P	F	
Uncompressed	19,404,571,275	8.0000	0.0000	0.0000	0.0000	N/A
CALQ 0b74e3d	1,089,506,661	0.4492	0.0003	-0.0001	0.0003	N/A
Crumble -1 (+ CRAM)	858,452,616	0.3539	0.0002	-0.0002	-0.0001	N/A
Crumble -9 (+ CRAM)	626,479,443	0.2583	0.0025	-0.0001	0.0015	N/A
DSRC IL8B (+ gzip)	3,292,742,011	1.3575	0.0029	0.0000	0.0018	N/A
P-Block p=1	5,524,698,820	2.2777	0.0004	0.0000	0.0003	VQSR failed for chromosome 3
P-Block p=4	2,559,896,988	1.0554	-0.0011	0.0000	-0.0007	N/A
Quartz (+ bzip2)	1,736,545,040	0.7159	0.0013	0.0000	0.0008	N/A
QVZ 2 T1	2,337,887,888	0.9639	0.0022	-0.0001	0.0013	N/A
QVZ 2 T2	1,671,939,362	0.6893	0.0017	0.0000	0.0010	N/A
QVZ 2 T4	1,057,508,124	0.4360	0.0033	-0.0001	0.0019	N/A
QVZ 2 T8	743,537,475	0.3065	0.0010	-0.0001	0.0006	N/A
QVZ 2 T16	433,127,845	0.1786	0.0010	-0.0004	0.0003	N/A
R-Block r=5	5,766,377,300	2.3773	0.0010	0.0000	0.0007	N/A
R-Block r=20	2,783,441,556	1.1475	0.0014	0.0000	0.0009	N/A

Figure 13: Variant calling results averaged over all three chromosomes as well as over all four VQSR filter values for the GATK + VQSR pipeline.

5.2 GATK + hard filtration pipeline

H01 (ERR174324)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	2,920,700,123	8.0000	0.8778	0.9982	0.9341	0.0000	0.0000	0.0000	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
CALQ	49,782,765	0.1364	0.8851	0.9982	0.9383	0.0073	0.0000	0.0041	37,826,773	0.1523	0.8740	0.9981	0.9319	0.0078	0.0002	0.0045	17,545,800	0.1594	0.9029	0.9976	0.9479	0.0087	-0.0001	0.0048
Crumble -1 (+ CRAM)	117,782,978	0.3226	0.8783	0.9983	0.9345	0.0005	0.0001	0.0003	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
Crumble -9 (+ CRAM)	63,646,598	0.1743	0.8793	0.9979	0.9349	0.0015	-0.0003	0.0007	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
DSRC IL8B (+ gzip)	332,094,399	0.9096	0.8783	0.9982	0.9344	0.0005	0.0000	0.0003	230,849,498	0.9296	0.8664	0.9979	0.9275	0.0002	0.0000	0.0001	107,172,747	0.9734	0.8947	0.9977	0.9434	0.0005	0.0000	0.0003
P-Block p=1	681,662,244	1.8671	0.8778	0.9982	0.9341	0.0000	0.0000	0.0000	473,222,724	1.9056	0.8661	0.9979	0.9273	-0.0001	0.0000	-0.0001	217,100,876	1.9719	0.8941	0.9976	0.9430	-0.0001	-0.0001	-0.0001
P-Block p=4	184,022,668	0.5041	0.8772	0.9981	0.9338	-0.0006	-0.0001	-0.0004	130,204,148	0.5243	0.8655	0.9979	0.9270	-0.0007	0.0000	-0.0004	61,445,676	0.5581	0.8933	0.9976	0.9426	-0.0009	-0.0001	-0.0005
Quartz (+ bzip2)	141,583,229	0.3878	0.8760	0.9981	0.9331	-0.0018	-0.0001	-0.0011	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
QVZ 2 T1	258,362,840	0.7077	0.8778	0.9982	0.9341	0.0000	0.0000	0.0000	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
QVZ 2 T2	129,785,486	0.3555	0.8779	0.9979	0.9341	0.0001	-0.0003	-0.0001	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
QVZ 2 T4	43,619,225	0.1195	0.8780	0.9957	0.9332	0.0002	-0.0025	-0.0010	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
QVZ 2 T8	12,483,416	0.0342	0.8780	0.9868	0.9292	0.0002	-0.0114	-0.0049	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
QVZ 2 T16	4,938,823	0.0135	0.8783	0.9825	0.9275	0.0005	-0.0157	-0.0067	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
R-Block r=5	682,714,860	1.8700	0.8778	0.9982	0.9341	0.0000	0.0000	0.0000	473,985,764	1.9087	0.8661	0.9979	0.9273	-0.0001	0.0000	-0.0001	217,474,276	1.9753	0.8941	0.9977	0.9431	-0.0001	0.0000	-0.0001
R-Block r=20	150,847,492	0.4132	0.8769	0.9981	0.9336	-0.0009	-0.0001	-0.0006	106,633,644	0.4294	0.8650	0.9979	0.9267	-0.0012	0.0000	-0.0007	50,270,852	0.4566	0.8929	0.9976	0.9424	-0.0013	-0.0001	-0.0008

H11 (SRR1238539)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	2,072,754,857	8.0000	0.3565	0.9398	0.5169	0.0000	0.0000	0.0000	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
CALQ	143,101,844	0.5523	0.3265	0.9430	0.4851	-0.0300	0.0032	-0.0319	98,589,364	0.5623	0.3108	0.9416	0.4673	-0.0293	0.0036	-0.0319	44,145,089	0.5671	0.3002	0.9373	0.4548	-0.0294	0.0062	-0.0321
Crumble -1 (+ CRAM)	947,960,418	3.6587	0.3567	0.9399	0.5171	0.0002	0.0001	0.0002	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
Crumble -9 (+ CRAM)	839,472,082	3.2400	0.3586	0.9416	0.5194	0.0021	0.0018	0.0025	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
DSRC IL8B (+ gzip)	482,889,314	1.8638	0.3496	0.9312	0.5084	-0.0069	-0.0086	-0.0086	326,159,362	1.8602	0.3347	0.9299	0.4922	-0.0054	-0.0081	-0.0070	145,109,631	1.8641	0.3200	0.9209	0.4750	-0.0096	-0.0102	-0.0119
P-Block p=1	1,094,046,244	4.2226	0.3553	0.9394	0.5156	-0.0012	-0.0004	-0.0013	740,567,588	4.2236	0.3391	0.9373	0.4980	-0.0010	-0.0007	-0.0012	328,643,924	4.2218	0.3281	0.9307	0.4852	-0.0015	-0.0004	-0.0017
P-Block p=4	369,265,972	1.4252	0.3527	0.9402	0.5130	-0.0038	0.0004	-0.0039	250,497,900	1.4286	0.3375	0.9385	0.4965	-0.0026	0.0005	-0.0027	112,184,956	1.4412	0.3258	0.9328	0.4829	-0.0038	0.0017	-0.0039
Quartz (+ bzip2)	347,780,109	1.3423	0.3384	0.9471	0.4986	-0.0181	0.0073	-0.0183	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
QVZ 2 T1	513,312,273	1.9812	0.3558	0.9401	0.5162	-0.0007	0.0003	-0.0007	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
QVZ 2 T2	405,552,393	1.5653	0.3592	0.9403	0.5198	0.0027	0.0005	0.0029	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
QVZ 2 T4	296,266,463	1.1435	0.3600	0.9384	0.5204	0.0035	-0.0014	0.0035	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
QVZ 2 T8	181,567,705	0.7008	0.3615	0.9389	0.5220	0.0050	-0.0009	0.0051	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
QVZ 2 T16	90,440,192	0.3491	0.3605	0.9385	0.5209	0.0040	-0.0013	0.0040	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
R-Block r=5	1,176,730,324	4.5417	0.3558	0.9397	0.5162	-0.0007	-0.0001	-0.0008	797,235,396	4.5468	0.3390	0.9378	0.4980	-0.0011	-0.0002	-0.0012	354,488,988	4.5538	0.3288	0.9308	0.4859	-0.0008	-0.0003	-0.0009
R-Block r=20	370,890,172	1.4315	0.3498	0.9396	0.5098	-0.0067	-0.0002	-0.0071	252,440,580	1.4397	0.3342	0.9380	0.4928	-0.0059	0.0000	-0.0064	113,848,684	1.4625	0.3220	0.9315	0.4786	-0.0076	0.0004	-0.0083

H12 (Garvan replicate)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	9,690,794,921	8.0000	0.8823	0.9996	0.9373	0.0000	0.0000	0.0000	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
CALQ	525,935,161	0.4342	0.8818	0.9993	0.9369	-0.0005	-0.0003	-0.0004	386,319,365	0.4615	0.8669	0.9993	0.9284	-0.0007	-0.0002	-0.0005	177,252,135	0.4700	0.8901	0.9993	0.9415	-0.0012	-0.0001	-0.0007
Crumble -1 (+ CRAM)	382,336,720	0.3156	0.8803	0.9995	0.9361	-0.0020	-0.0001	-0.0012	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
Crumble -9 (+ CRAM)	275,932,079	0.2278	0.8801	0.9994	0.9360	-0.0022	-0.0002	-0.0013	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
DSRC IL8B (+ gzip)	1,623,022,824	1.3398	0.8825	0.9996	0.9374	0.0002	0.0000	0.0001	1,140,807,424	1.3629	0.8677	0.9995	0.9289	0.0001	0.0000	0.0001	528,911,763	1.4024	0.8914	0.9995	0.9424	0.0001	0.0001	0.0001
P-Block p=1	2,706,725,668	2.2345	0.8823	0.9996	0.9373	0.0000	0.0000	0.0000	1,920,062,916	2.2938	0.8675	0.9996	0.9289	-0.0001	0.0001	0.0000	897,910,236	2.3808	0.8914	0.9994	0.9423	0.0001	0.0000	0.0001
P-Block p=4	1,252,922,100	1.0343	0.8821	0.9996	0.9372	-0.0002	0.0000	-0.0001	890,220,628	1.0635	0.8673	0.9995	0.9287	-0.0003	0.0000	-0.0002	416,754,260	1.1050	0.8910	0.9995	0.9421	-0.0003	0.0001	-0.0001
Quartz (+ bzip2)	823,855,969	0.6801	0.8910	0.9996	0.9422	0.0087	0.0000	0.0049	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
QVZ 2 T1	1,139,649,473	0.9408	0.8825	0.9996	0.9374	0.0002	0.0000	0.0001	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
QVZ 2 T2	811,092,140	0.6696	0.8825	0.9996	0.9374	0.0002	0.0000	0.0001	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
QVZ 2 T4	512,260,334	0.4229	0.8825	0.9996	0.9374	0.0002	0.0000	0.0001	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
QVZ 2 T8	357,588,514	0.2952	0.8825	0.9995	0.9374	0.0002	-0.0001	0.0001	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
QVZ 2 T16	206,015,477	0.1701	0.8830	0.9993	0.9376	0.0007	-0.0003	0.0003	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
R-Block r=5	2,823,881,516	2.3312	0.8823	0.9996	0.9373	0.0000	0.0000	0.0000	2,004,430,612	2.3946	0.8676	0.9996	0.9289	0.0000	0.0001	0.0000	938,065,172	2.4873	0.8913	0.9994	0.9423	0.0000	0.0000	0.0000
R-Block r=20	1,355,361,420	1.1189	0.8823	0.9997	0.9373	0.0000	0.0001	0.0000	970,006,292	1.1588	0.8675	0.9996	0.9289	-0.0001	0.0001	0.0000	458,073,844	1.2146	0.8912	0.9995	0.9422	-0.0001	0.0001	0.0000

Figure 14: Variant calling results for the GATK + hard filtration pipeline.

5.3 Platypus pipeline

H01 (ERR174324)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	2,920,700,123	8.0000	0.7924	0.9986	0.8836	0.0000	0.0000	0.0000	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
CALQ	49,782,765	0.1364	0.7998	0.9986	0.8882	0.0074	0.0000	0.0046	37,826,773	0.1523	0.7828	0.9984	0.8776	0.0069	0.0000	0.0044	17,545,800	0.1594	0.8026	0.9982	0.8898	0.0079	0.0001	0.0049
Crumble -1 (+ CRAM)	117,782,978	0.3226	0.7926	0.9988	0.8838	0.0002	0.0002	0.0002	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
Crumble -9 (+ CRAM)	63,646,598	0.1743	0.7934	0.9989	0.8844	0.0010	0.0003	0.0007	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
DSRC IL8B (+ gzip)	332,094,399	0.9096	0.7939	0.9987	0.8846	0.0015	0.0001	0.0010	230,849,498	0.9296	0.7774	0.9983	0.8741	0.0015	-0.0001	0.0009	107,172,747	0.9734	0.7966	0.9981	0.8860	0.0019	0.0000	0.0012
P-Block p=1	681,662,244	1.8671	0.7923	0.9986	0.8836	-0.0001	0.0000	-0.0001	473,222,724	1.9056	0.7758	0.9984	0.8731	-0.0001	0.0000	-0.0001	217,100,876	1.9719	0.7946	0.9981	0.8848	-0.0001	0.0000	-0.0001
P-Block p=4	184,022,668	0.5041	0.7906	0.9986	0.8825	-0.0018	0.0000	-0.0011	130,204,148	0.5243	0.7742	0.9984	0.8721	-0.0017	0.0000	-0.0011	61,445,676	0.5581	0.7931	0.9982	0.8839	-0.0016	0.0001	-0.0010
Quartz (+ bzip2)	141,583,229	0.3878	0.7875	0.9985	0.8805	-0.0049	-0.0001	-0.0031	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
QVZ 2 T1	258,362,840	0.7077	0.7921	0.9986	0.8839	0.0010	-0.0009	0.0003	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
QVZ 2 T2	129,785,486	0.3555	0.7927	0.9986	0.8838	0.0003	0.0000	0.0002	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
QVZ 2 T4	43,619,225	0.1195	0.7934	0.9977	0.8839	0.0010	-0.0009	0.0003	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
QVZ 2 T8	12,483,416	0.0342	0.7941	0.9925	0.8823	0.0017	-0.0061	-0.0013	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
QVZ 2 T16	4,938,823	0.0135	0.7943	0.9847	0.8793	0.0019	-0.0139	-0.0043	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
R-Block r=5	682,714,860	1.8700	0.7923	0.9986	0.8836	-0.0001	0.0000	-0.0001	473,985,764	1.9087	0.7758	0.9984	0.8731	-0.0001	0.0000	-0.0001	217,474,276	1.9753	0.7946	0.9981	0.8848	-0.0001	0.0000	-0.0001
R-Block r=20	150,847,492	0.4132	0.7887	0.9985	0.8813	-0.0037	-0.0001	-0.0023	106,633,644	0.4294	0.7725	0.9983	0.8710	-0.0034	-0.0001	-0.0022	50,270,852	0.4566	0.7908	0.9981	0.8824	-0.0039	0.0000	-0.0024

H11 (SRR1238539)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	2,072,754,857	8.0000	0.0711	0.9597	0.1324	0.0000	0.0000	0.0000	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
CALQ	143,101,844	0.5523	0.0674	0.9556	0.1259	-0.0037	-0.0041	-0.0065	98,589,364	0.5623	0.0615	0.9540	0.1156	-0.0045	-0.0076	-0.0080	44,145,089	0.5671	0.0576	0.9536	0.1086	-0.0029	-0.0063	-0.0052
Crumble -1 (+ CRAM)	947,960,418	3.6587	0.0711	0.9596	0.1324	0.0000	-0.0001	0.0000	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
Crumble -9 (+ CRAM)	839,472,082	3.2400	0.0712	0.9594	0.1326	0.0001	-0.0003	0.0002	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
DSRC IL8B (+ gzip)	482,889,314	1.8638	0.0719	0.9590	0.1338	0.0008	-0.0007	0.0014	326,159,362	1.8602	0.0667	0.9610	0.1247	0.0007	-0.0006	0.0012	145,109,631	1.8641	0.0610	0.9579	0.1147	0.0005	-0.0020	0.0009
P-Block p=1	1,094,046,244	4.2226	0.0711	0.9594	0.1324	0.0000	-0.0003	0.0000	740,567,588	4.2236	0.0659	0.9618	0.1233	-0.0001	0.0002	-0.0002	328,643,924	4.2218	0.0605	0.9599	0.1138	0.0000	0.0000	0.0000
P-Block p=4	369,265,972	1.4252	0.0715	0.9594	0.1331	0.0004	-0.0003	0.0007	250,497,900	1.4286	0.0660	0.9612	0.1235	0.0000	-0.0004	0.0000	112,184,956	1.4412	0.0605	0.9597	0.1138	0.0000	-0.0002	0.0000
Quartz (+ bzip2)	347,780,109	1.3423	0.0580	0.9602	0.1094	-0.0131	0.0005	-0.0230	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
QVZ 2 T1	513,312,273	1.9812	0.0709	0.9597	0.1320	-0.0002	0.0000	-0.0003	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
QVZ 2 T2	405,552,393	1.5653	0.0714	0.9599	0.1329	0.0003	0.0002	0.0005	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
QVZ 2 T4	296,266,463	1.1435	0.0717	0.9600	0.1334	0.0006	0.0003	0.0010	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
QVZ 2 T8	181,567,705	0.7008	0.0720	0.9605	0.1340	0.0009	0.0008	0.0016	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
QVZ 2 T16	90,440,192	0.3491	0.0727	0.9593	0.1352	0.0016	-0.0004	0.0028	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
R-Block r=5	1,176,730,324	4.5417	0.0711	0.9598	0.1324	0.0000	0.0001	0.0000	797,235,396	4.5468	0.0659	0.9617	0.1233	-0.0001	0.0001	-0.0002	354,488,988	4.5538	0.0605	0.9603	0.1138	0.0000	0.0004	0.0000
R-Block r=20	370,890,172	1.4315	0.0705	0.9593	0.1313	-0.0006	-0.0004	-0.0010	252,440,580	1.4397	0.0652	0.9617	0.1221	-0.0008	0.0001	-0.0014	113,848,684	1.4625	0.0597	0.9594	0.1124	-0.0008	-0.0005	-0.0014

H12 (Garvan replicate)																								
	Chromosome 3								Chromosome 11								Chromosome 20							
	Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data			Size [B]	Bits/QV	R	P	F	Differences			Size [B]	Bits/QV	R	P	F	Differences w.r.t. uncompressed data		
						R	P	F						R	P	F						R	P	F
Uncompressed	9,690,794,921	8.0000	0.7521	0.9996	0.8584	0.0000	0.0000	0.0000	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
CALQ	525,935,161	0.4342	0.7418	0.9994	0.8515	-0.0103	-0.0002	-0.0068	386,319,365	0.4615	0.7134	0.9993	0.8325	-0.0114	-0.0003	-0.0078	177,252,135	0.4700	0.7155	0.9996	0.8340	-0.0142	-0.0001	-0.0096
Crumble -1 (+ CRAM)	382,336,720	0.3156	0.7554	0.9996	0.8605	0.0033	0.0000	0.0021	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
Crumble -9 (+ CRAM)	275,932,079	0.2278	0.7561	0.9996	0.8610	0.0040	0.0000	0.0026	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
DSRC IL8B (+ gzip)	1,623,022,824	1.3398	0.7556	0.9996	0.8606	0.0035	0.0000	0.0023	1,140,807,424	1.3629	0.7292	0.9995	0.8432	0.0044	-0.0001	0.0029	528,911,763	1.4024	0.7353	0.9996	0.8473	0.0056	-0.0001	0.0037
P-Block p=1	2,706,725,668	2.2345	0.7521	0.9996	0.8584	0.0000	0.0000	0.0000	1,920,062,916	2.2938	0.7248	0.9996	0.8403	0.0000	0.0000	0.0000	897,910,236	2.3808	0.7297	0.9997	0.8436	0.0000	0.0000	0.0000
P-Block p=4	1,252,922,100	1.0343	0.7525	0.9997	0.8587	0.0004	0.0001	0.0003	890,220,628	1.0635	0.7254	0.9996	0.8407	0.0006	0.0000	0.0004	416,754,260	1.1050	0.7306	0.9998	0.8443	0.0009	0.0001	0.0006
Quartz (+ bzip2)	823,855,969	0.6801	0.7865	0.9997	0.8804	0.0344	0.0001	0.0220	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
QVZ 2 T1	1,139,649,473	0.9408	0.7515	0.9996	0.8580	-0.0006	0.0000	-0.0004	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
QVZ 2 T2	811,092,140	0.6696	0.7510	0.9996	0.8576	-0.0011	0.0000	-0.0007	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
QVZ 2 T4	512,260,334	0.4229	0.7480	0.9997	0.8557	-0.0041	0.0001	-0.0026	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
QVZ 2 T8	357,588,514	0.2952	0.7508	0.9996	0.8575	-0.0013	0.0000	-0.0008	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
QVZ 2 T16	206,015,477	0.1701	0.7545	0.9995	0.8599	0.0024	-0.0001	0.0015	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
R-Block r=5	2,823,881,516	2.3312	0.7521	0.9996	0.8584	0.0000	0.0000	0.0000	2,004,430,612	2.3946	0.7248	0.9996	0.8403	0.0000	0.0000	0.0000	938,065,172	2.4873	0.7297	0.9997	0.8436	0.0000	0.0000	0.0000
R-Block r=20	1,355,361,420	1.1189	0.7521	0.9997	0.8584	0.0000	0.0001	0.0000	970,006,292	1.1588	0.7248	0.9996	0.8403	0.0000	0.0000	0.0000	458,073,844	1.2146	0.7299	0.9998	0.8438	0.0002	0.0001	0.0002

Figure 15: Variant calling results for the Platypus pipeline.

References

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

[CMT14] Rodrigo Cánovas, Alistair Moffat, and Andrew Turpin. Lossy compression of quality scores in genomic data. *Bioinformatics*, 30(15):2130–2136, 2014.

[DG11] Sebastian Deorowicz and Szymon Grabowski. Compression of DNA sequence reads in FASTQ format. *Bioinformatics*, 27(6):860–862, 2011.

[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.

[Joi16a] Joint AhG on Genomic Information Compression And Storage. Benchmark framework for lossy compression of sequencing quality values. Technical report, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) and ISO/TC 276/WG 5, Document Number N16525/N119, Chengdu (CN), 2016.

[Joi16b] Joint AhG on Genomic Information Compression And Storage. Updated database for Evaluation of Genomic Information Compression and Storage. Technical report, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) and ISO/TC 276/WG 5, Document Number N16324/N100, Chengdu (CN), 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.

[RD14] Lukasz Roguski and Sebastian Deorowicz. DSRC 2 - Industry-oriented compression of FASTQ files. *Bioinformatics*, 30(15):2213–2215, 2014.

[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.