NGC: Lossless and Lossy Compression of Aligned High-throughput Sequencing Data
Supplementary Materials

Niko Popitsch¹, Arndt von Haeseler

Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Dr. Bohr Gasse 9, A-1030, Vienna, Austria

Contents
I. Supplementary Description of Variant Calling Pipelines ..................................................................................................................... 2
   GATK pipeline ................................................................................................................................................................................................................. 3
   SAMTOOLS pipeline ......................................................................................................................................................................................................... 3
II. Definitions ........................................................................................................................................................................................................ 4
   Counting horizontal and vertical run-lengths ........................................................................................................................................................................... 4
III. Compression parameters ....................................................................................................................................................................................................... 5
   NGC v0.0.1 ........................................................................................................................................................................................................ 5
   Cram v0.85 ........................................................................................................................................................................................................ 5
   Goby v2.1 ........................................................................................................................................................................................................ 5
IV. Computing Environment ................................................................................................................................................................................................... 6
V. Supplementary Data Tables .................................................................................................................................................................................................... 6

¹ Corresponding author
I. Supplementary Description of Variant Calling Pipelines

The Reseq/chr20 (human) data set corresponds to the file NA12878.HiSeq.WGS.bwa.cleaned.recal.hg19.20.bam that was downloaded from the GATK resource bundle v1.5 b37 (available at ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/1.5/b37, be sure to login as user gsapubftp-anonymous/<blank>). The file can be found in the subfolder bundle/1.5/b37). This file contains mapped reads of the NA12878 individual’s chromosome 20. Note that this data set contains all original q-values of the reads (probably the ones before q-value recalibration) stored in OQ tags which makes this file less compressible in comparison to the others.

All other data sets were downloaded from the SRA respectively given to us as FASTQ files and were mapped using bwa v 0.6.1-r104 (http://bio-bwa.sourceforge.net/) with standard parameters for single respectively paired end data. The following reference sequences were used for mapping:

- *Homo sapiens* genome human g1k v37, available at ftp://ftp.sanger.ac.uk/pub/1000genomes/tk2/main_project_reference/human_g1k_v37.fasta.gz
- *Mus musculus* genome mm9, NCBI Build 37, available at http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/mm9.2bit
- *Escherichia coli* str. K-12 substr. MG1655 genome, available at ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__MG1655_uid57779/NC_000913.fna
- *Arabidopsis thaliana* TAIR10 genome, available at ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/

Note that we pruned unmapped reads from the data sets using SAMTOOLS v0.1.18 (http://samtools.sourceforge.net/) before our tests and that our algorithm expects the input SAM/BAM file to be sorted by mapping coordinates. Note also that our solution preserves also hard- and soft-clipped bases that actually break the order in such a coordinate-sorted file by storing those in some extra data streams (cf. main paper, Figure 1).
**GATK Pipeline**
Our pipeline for GATK (17) variant calling consisted of the following steps:

- Addition of read groups using Picard v1.56 ([http://picard.sourceforge.net](http://picard.sourceforge.net)), AddOrReplaceReadGroups
- INDEL realignment with GATK v1.4 RealignerTargetCreator and IndelRealigner
- Duplicate removal with Picard MarkDuplicates
  - *Duplicates were removed from the alignment*
  - *Note that this step was omitted for the RNA-seq (e.coli) data set!*
- Variant calling with GATK UnifiedGenotyper
- Variant filtration with GATK VariantFiltration
  - For SNPs, we filtered for low coverage (<5), low quality (<50), variant confidence (<1.5), strand bias and mapping quality
  - For INDELs, we filtered for variant confidence, read position bias and strand bias
- Both variant sets were then combined using GATK CombineVariants
- *Note that the first three pipeline stages were omitted for the Reseq/chr20 (human) data set as comparable preprocessing was already done for the respective BAM file*

**SAMTOOLS Pipeline**
Our pipeline for variant calling with SAMTOOLS consisted of the following steps:

- Addition of read groups using Picard AddOrReplaceReadGroups
- INDEL realignment with GATK RealignerTargetCreator and IndelRealigner
- Duplicate removal with Picard MarkDuplicates
  - *This step was omitted for the RNAseq data set.*
- Variant calling with SAMTOOLS v0.1.18 mpileup/bcftools/vcfutils
  - Variants were filtered by minimum (5) and maximum (2000) read depth
- *Note that the first three pipeline stages were omitted for the Reseq/chr20 (human) data set as comparable preprocessing was already done for the respective BAM file*

Note that we did not do any quality score recalibration in this evaluation pipeline as such a step would affect all quality scores in the data sets which would blur the effects of the different quantization scenarios.

Further note, that the Picard algorithm for the removal of redundant sequences (stemming mainly from PCR amplification) also incorporates read bases qualities and will therefore also lead to different read-sets depending on the used q-value quantization strategy.
II. Definitions
True positives (tp), false positives (fp) and false negatives (fn) were determined by comparing the sets of called variants from the BAM files created by compressing and uncompressing the original alignment using (1) the lossless mode (unquantized) and (2) the respective quantization mode.

Compression ratio is defined as $cr = \frac{\text{compressed size}}{\text{uncompressed size}}$.

Compression rates are, in contrast, usually measured in bits per sample.

Space savings is defined as $1 - cr$.

Precision is defined as $\text{prec} = \frac{tp}{tp + fp}$.

Sensitivity (aka recall rate) is defined as $\text{sens} = \frac{tp}{tp + fn}$.

Genotype preservation percentage was calculated as $\text{gpp} = 1 - \frac{\text{cgt}}{tp}$ where cgt is the number of variants from the set of true positives that changed their genotype classification from homozygous to heterozygous or vice versa.

Coverage was calculated as $\text{cov} = \frac{sb}{\text{ref}}$ where sb is the number of sequenced bases and ref is the length of the reference sequence.

Counting Horizontal and Vertical Run-lengths
For counting the run-lengths (RLs) presented in Figure 2 of the main paper we considered two streams of read bases. The “horizontal” one was created by simply iterating over the reads in their given order in the alignment file. The “vertical” one was produced as described in the paper (cf. Figure 2). Note that clipped bases are not counted and that RLs are also counted within insertions. In the vertical mode, we consider each inserted “column” as follows: for each read spanning the insertion, we consider either its inserted base character (A,C,T,G,N and all other IUPAC/FASTA characters) in this column or, if none, (e.g., in padded regions) a special character “X”. These base characters are then “diffed” against the reference character “X” (so that a padded position finally results in an “E” character). Note that this method slightly penalizes the vertical mode, i.e., a more optimized method would count even less required RLs in the vertical mode.

Further note that the given numbers are the “theoretical” numbers of required run-lengths as described in the paper. In practice, however, not only these numbers but also the amount of bits needed to represent this information is relevant. For representing the length-values of the RLs one could, for example, use fixed-sized 8-bit code words. This would mean that RLs with a length value greater than 256 have to be split-up, which again increases the number of required code words. Reserving more bits would again reduce the number of code words but would require more disc space per code word. With our data we actually found that using one byte per RL length was the best tradeoff in this regard (data not shown).
III. Compression Parameters

We used the following commandline parameters for the evaluation experiments:

**NGC v0.0.1**

*Lossless:* -best --truncateNames

- **m1:** -best -q1levels standard -qvalRleEncoding -truncateNames
- **m2:** -best -q1levels standard -q2levels standard -qvalRleEncoding -truncateNames
- **m3:** -best -q1levels standard -qvalRleEncoding -truncateNames
- **m4:** -best -q1levels 30 -q2levels 30 -qvalRleEncoding -truncateNames
- **m5:** -best -q1levels 30 -q2levels 30 -q4levels standard -qvalRleEncoding -mmq 20 -q5levels 30 -truncateNames

*recovery:* -best -q1levels 30 -q2levels standard -qvalRleEncoding -truncateNames -variantList <variants.vcf>

*drop all:* -best -q1levels 30 -q2levels 30 -q3levels 30 -q4levels 30 -qvalRleEncoding -truncateNames

**Cram v0.85**

*lossless compression:* --capture-all-quality-scores --capture-all-tags

*lossy compression:* --capture-insertion-quality-scores --capture-piled-quality-scores --capture-substitution-quality-scores

--capture-unmapped-quality-scores --capture-all-tags

*decompression:* --calculate-md-tag --calculate-nm-tag

**Goby v2.1**

*lossless compression:* --sorted --preserve-all-tags --preserve-soft-clips --preserve-all-mapped-qualities --ambiguity-threshold 1000000

*decompression:* <no options>

NOTE: Goby could compress/decompress the Reseq/hm (human) data set, however, the resulting BAM file contained less reads than the original. Goby could not compress the Reseq/chr20 (human) with 16GByte of dedicated RAM within two days after which we stopped the job.
IV. Computing Environment

The evaluation experiments were conducted on a server equipped with 2 Xeon E5520 processors and a total of 32GBytes of RAM. Both, cram and NGC, were executed using Java 1.7.0_02 64Bit with 4Gbytes of maximum heap size (-Xmx switch). Goby was executed using Java 1.6.0_27 64Bit as the latest version we used (v 2.1) did not run with Java 1.7. We also used 4Gbytes of maximum heap size, except for compressing the Reseq (E. coli, PE) data set where we used 16GByte as we ran into OutOfMemory errors.

V. Supplementary Data Tables

This section contains the absolute data values from our evaluation experiments.

| id  | Sample                  | Evaluation mode | mapped reads | Bases and qvalues | BAM [byte] | CRAM [byte] | Comp. ratio | Comp [s] | Decomp [s] | comp [min] | decomp [min] |
|-----|-------------------------|-----------------|--------------|------------------|------------|-------------|-------------|----------|------------|------------|--------------|
| 1   | Exome-seq (human)      | lossless        | 3,239,217    | 323,921,700      | 199,019,319| 129,302,897 | 0.65        | 67.52    | 166.60     | 1.1        | 2.8          |
|     |                         | lossy           | 3,239,217    | 323,921,700      | 199,019,319| 8,993,932   | 0.05        | 44.44    | 91.57      | 0.7        | 1.5          |
| 2   | ChiP-seq (mouse)       | lossless        | 13,824,441   | 497,679,876      | 637,462,912| 282,756,168 | 0.44        | 200.71   | 468.84     | 3.3        | 7.8          |
|     |                         | lossy           | 13,824,441   | 497,679,876      | 637,462,912| 55,625,624  | 0.09        | 180.33   | 360.41     | 3.0        | 6.0          |
| 3   | Reseq/hapma (human)    | lossless        | 487,201      | 62,435,260       | 53,036,863 | 24,054,223  | 0.45        | 27.64    | 58.70      | 0.5        | 1.0          |
|     |                         | lossy           | 487,201      | 62,435,260       | 53,036,863 | 4,342,572   | 0.08        | 27.45    | 50.81      | 0.5        | 0.8          |
| 4   | RNA-seq (e.coli)       | lossless        | 6,927,728    | 249,398,208      | 177,499,665| 105,288,858 | 0.59        | 78.80    | 183.76     | 1.3        | 3.1          |
|     |                         | lossy           | 6,927,728    | 249,398,208      | 177,499,665| 4,979,867   | 0.03        | 61.80    | 124.19     | 1.0        | 2.1          |
| 5   | Reseq (e.coli,PE)      | lossless        | 44,938,891   | 4,538,790,538    | 2,945,453,583| 2,308,990,877| 0.78        | 1,212.96 | 2,976.40   | 20.2       | 49.6         |
|     |                         | lossy           | 44,938,891   | 4,538,790,538    | 2,945,453,583| 647,050,983  | 0.22        | 975.03   | 2,036.86   | 16.3       | 33.9         |
| 6   | Reseq (e. coli)        | lossless        | 19,379,287   | 697,654,332      | 654,501,526 | 339,334,229 | 0.52        | 218.61   | 521.37     | 3.6        | 8.7          |
|     |                         | lossy           | 19,379,287   | 697,654,332      | 654,501,526 | 19,259,853  | 0.03        | 167.68   | 349.53     | 2.8        | 5.8          |
| 7   | Reseq (a. thaliana)    | lossless        | 11,651,942   | 419,469,912      | 503,542,318 | 235,324,311 | 0.47        | 159.18   | 376.83     | 2.7        | 6.3          |
|     |                         | lossy           | 11,651,942   | 419,469,912      | 503,542,318 | 48,464,278  | 0.10        | 128.03   | 267.89     | 2.1        | 4.5          |
| 8   | Reseq/chr20 (human)    | lossless        | 50,663,069   | 5,116,969,969    | 5,867,598,471| 4,055,011,808| 0.69        | 1,572.47 | 3,894.58   | 26.2       | 64.9         |
|     |                         | lossy           | 50,663,069   | 5,116,969,969    | 5,867,598,471| 2,068,227,640| 0.35        | 1,624.37 | 3,339.40   | 27.1       | 55.7         |

Table 1s: Cram evaluation data. The options for these modes are set as written above in Section III.
| data set | compression | decompression | BIM [byte] | NGC [byte] | Comp. ratio | Space saving | NGC bases | % of total | decompression | BIM [byte] | Comp. ratio | Space saving | NGC bases | % of total | decompression | BIM [byte] |
|----------|-------------|---------------|-------------|-------------|-------------|--------------|-------------|-----------|--------------|-------------|-------------|--------------|-------------|-----------|--------------|-------------|
| (human)  | requant (f)  | 1,319,417     | 557,158,510 | 194,019,319 | 0.56        | 0.13         | 43,462      | 0.00      | 557,158,510 | 194,019,319 | 0.56        | 0.13         | 43,462      | 0.00      | 557,158,510 | 194,019,319 | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | (m3)        | 4,898,519     | 557,158,510 | 194,019,319 | 0.56        | 0.13         | 43,462      | 0.00      | 91,932,668  | 37,128,727  | 0.56        | 0.13         | 43,462      | 0.00      | 91,932,668  | 37,128,727  | 0.56        |
| (human)  | requant (f)  | 17,126,557    | 557,158,510 | 194,019,319 | 0.56        | 0.13         | 43,462      | 0.00      | 106,571,578 | 42,620,203  | 0.56        | 0.13         | 43,462      | 0.00      | 106,571,578 | 42,620,203  | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | requant (f)  | 14,284,941    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 171,503,159 | 68,713,794  | 0.56        | 0.13         | 43,462      | 0.00      | 171,503,159 | 68,713,794  | 0.56        |
| (human)  | bestacc3    | 112,505,193   | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 1,436,323   | 542,688      | 0.56        | 0.13         | 43,462      | 0.00      | 1,436,323   | 542,688      | 0.56        |
| (human)  | requant (f)  | 17,126,557    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 5,749,243   | 2,219,524    | 0.56        | 0.13         | 43,462      | 0.00      | 5,749,243   | 2,219,524    | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | requant (f)  | 6,751,733     | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 1,967,637   | 722,224      | 0.56        | 0.13         | 43,462      | 0.00      | 1,967,637   | 722,224      | 0.56        |
| (human)  | quant | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | dropall     | 1,432,519     | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 1,432,519   | 542,688      | 0.56        | 0.13         | 43,462      | 0.00      | 1,432,519   | 542,688      | 0.56        |
| (human)  | requant (f)  | 72,766,925    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 72,766,925  | 2,973,219    | 0.56        | 0.13         | 43,462      | 0.00      | 72,766,925  | 2,973,219    | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | requant (f)  | 60,955,660    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 60,955,660  | 2,496,888    | 0.56        | 0.13         | 43,462      | 0.00      | 60,955,660  | 2,496,888    | 0.56        |
| (human)  | requant (f)  | 10,061,098    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 10,061,098  | 4,207,201    | 0.56        | 0.13         | 43,462      | 0.00      | 10,061,098  | 4,207,201    | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |
| (human)  | requant (f)  | 25,902,383    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 25,902,383  | 9,743,105    | 0.56        | 0.13         | 43,462      | 0.00      | 25,902,383  | 9,743,105    | 0.56        |
| (human)  | requant (f)  | 32,814,519    | 407,478,876 | 462,462,812 | 0.56        | 0.13         | 43,462      | 0.00      | 32,814,519  | 12,397,899   | 0.56        | 0.13         | 43,462      | 0.00      | 32,814,519  | 12,397,899   | 0.56        |
| (human)  | requant (f)  | -             | -            | -           | -            | -            | -           | -         | -             | -           | -            | -            | -           | -         | -             | -           |

Table 2s: NGC compression ratios/times
| Dataset | Q-value quantization strategy | Ty | Lp | Lr | Lc | Met | Nom | Metnome 融合 | Nomnome 融合 | Recommend | Recomen | Recommend | Recommen | Recommen |
|--------|-----------------------------|----|----|----|----|-----|-----|---------------|---------------|------------|---------|------------|---------|---------|
| 1 | noquant (human) | 3238 | 0 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 2 | noquant (human) | 3387 | 651 | 17 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 3 | noquant (human) | 3402 | 24 | 18 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 4 | noquant (human) | 3797 | 99 | 19 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 5 | noquant (human) | 3926 | 86 | 65 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 6 | noquant (human) | 31108 | 426 | 42 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 7 | noquant (human) | 31084 | 91 | 16 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 8 | noquant (human) | 31056 | 105 | 17 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |

| Dataset | Q-value quantization strategy | Ty | Lp | Lr | Lc | Met | Nom | Metnome 融合 | Nomnome 融合 | Recommend | Recomen | Recommend | Recommen | Recommen |
|--------|-----------------------------|----|----|----|----|-----|-----|---------------|---------------|------------|---------|------------|---------|---------|
| 1 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 2 | quant (human) | 1397 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 3 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 4 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 5 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 6 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 7 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |
| 8 | quant (human) | 1287 | 100 | 0 | 0 | 0 | 100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 100.0% | 100.0% |

**Table 4s:** Variant recovery (GATK)

**Table 3s:** Variant recovery (SAMTOOLS)
| Data set       | Q-value quantization strategy | prec  | sens  | %     | Sorted coverage (%) |
|---------------|------------------------------|-------|-------|-------|---------------------|
| eXome-seq     | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (human)       | dropout (drop-call)          | 0.539 | 0.966 | 0.998 | 96.6%               |
|               | quant-bestcall 1 (m2)        | 0.984 | 0.994 | 0.998 | 99.6%               |
|               | quant-bestcall 2 (m3)        | 0.995 | 0.996 | 0.999 | 99.6%               |
|               | quant-bestcall 3 (m3)        | 0.978 | 0.993 | 0.997 | 99.3%               |
|               | quant-recover (recovery)     | 1.000 | 0.993 | 0.998 | 99.3%               |
|               | quant-bestcall (m4)          | 0.978 | 0.983 | 0.994 | 98.3%               |
|               | bestcomp (m5)                | 0.974 | 0.983 | 0.994 | 98.3%               |
| ChIP-seq      | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (mouse)       | dropout (drop-call)          | 0.773 | 0.838 | 0.984 | 85.8%               |
|               | quant-bestcall 1 (m2)        | 0.988 | 0.987 | 0.993 | 98.7%               |
|               | quant-bestcall 2 (m3)        | 0.995 | 0.995 | 0.999 | 99.5%               |
|               | quant-bestcall 3 (m3)        | 0.979 | 0.983 | 0.993 | 98.3%               |
|               | quant-recover (recovery)     | 0.992 | 0.994 | 0.997 | 99.4%               |
|               | quant-bestcall (m4)          | 0.980 | 0.969 | 0.984 | 96.9%               |
|               | bestcomp (m5)                | 0.957 | 0.972 | 0.984 | 96.7%               |
| Reseq/hapmap  | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (human)       | dropout (drop-call)          | 0.838 | 0.908 | 0.986 | 90.8%               |
|               | quant-bestcall 1 (m2)        | 0.998 | 0.994 | 0.999 | 99.9%               |
|               | quant-bestcall 2 (m3)        | 0.998 | 0.997 | 1.000 | 99.7%               |
|               | quant-bestcall 3 (m3)        | 0.997 | 0.991 | 0.999 | 99.9%               |
|               | quant-recover (recovery)     | 0.998 | 0.997 | 0.999 | 99.9%               |
|               | quant-bestcall (m4)          | 0.995 | 0.987 | 0.985 | 98.7%               |
|               | bestcomp (m5)                | 0.965 | 0.976 | 0.988 | 97.6%               |
| RNA-seq       | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (e. coli)     | dropout (drop-call)          | 0.326 | 0.784 | 0.996 | 78.4%               |
|               | quant-bestcall 1 (m2)        | 0.993 | 0.948 | 1.000 | 94.8%               |
|               | quant-bestcall 2 (m3)        | 0.997 | 0.980 | 1.000 | 99.0%               |
|               | quant-bestcall 3 (m3)        | 0.910 | 0.948 | 1.000 | 94.8%               |
|               | quant-recover (recovery)     | 0.997 | 0.997 | 1.000 | 99.7%               |
|               | quant-bestcall (m4)          | 0.846 | 0.886 | 1.000 | 88.6%               |
|               | bestcomp (m5)                | 0.843 | 0.879 | 1.000 | 87.9%               |
| Reseq/hapmap  | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (mouse)       | dropout (drop-call)          | 0.153 | 0.984 | 1.000 | 98.4%               |
|               | quant-bestcall 1 (m2)        | 0.970 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall 2 (m3)        | 1.000 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall 3 (m3)        | 0.824 | 1.000 | 1.000 | 100.0%              |
|               | quant-recover (recovery)     | 0.980 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall (m4)          | 0.869 | 0.904 | 1.000 | 90.4%               |
|               | bestcomp (m5)                | 0.869 | 0.904 | 1.000 | 90.4%               |
| Reseq         | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (e. coli)     | dropout (drop-call)          | 0.467 | 0.986 | 1.000 | 98.6%               |
|               | quant-bestcall 1 (m2)        | 0.987 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall 2 (m3)        | 0.994 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall 3 (m3)        | 0.955 | 0.994 | 1.000 | 99.6%               |
|               | quant-recover (recovery)     | 0.980 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall (m4)          | 0.948 | 0.973 | 1.000 | 97.3%               |
|               | bestcomp (m5)                | 0.941 | 0.973 | 1.000 | 97.3%               |
| Reseq/hapmap  | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (a. thaliana) | dropout (drop-call)          | 0.921 | 0.958 | 0.991 | 95.8%               |
|               | quant-bestcall 1 (m2)        | 0.997 | 0.994 | 0.998 | 99.4%               |
|               | quant-bestcall 2 (m3)        | 0.999 | 0.988 | 1.000 | 99.8%               |
|               | quant-bestcall 3 (m3)        | 0.994 | 0.992 | 0.998 | 99.2%               |
|               | quant-recover (recovery)     | 0.998 | 0.997 | 1.000 | 99.7%               |
|               | quant-bestcall (m4)          | 0.994 | 0.985 | 0.999 | 98.9%               |
|               | bestcomp (m5)                | 0.984 | 0.987 | 0.999 | 98.7%               |
| Reseq/hapmap  | no quant (lossless)          | 1.000 | 1.000 | 1.000 | 100.0%              |
| (human)       | dropout (drop-call)          | 0.911 | 0.972 | 0.997 | 97.4%               |
|               | quant-bestcall 1 (m2)        | 0.998 | 0.999 | 0.999 | 99.9%               |
|               | quant-bestcall 2 (m3)        | 1.000 | 1.000 | 1.000 | 100.0%              |
|               | quant-bestcall 3 (m3)        | 0.996 | 0.998 | 0.999 | 99.8%               |
|               | quant-recover (recovery)     | 0.999 | 0.999 | 1.000 | 99.9%               |
|               | quant-bestcall (m4)          | 0.998 | 0.994 | 0.999 | 99.4%               |
|               | bestcomp (m5)                | 0.997 | 0.994 | 0.999 | 99.4%               |
| Table 5: Variant recovery (averaged) | | | | |