Additional file 2.1

Instruction to compile a Graphtyper version modified for the cattle chromosome complement

Modified Graphtyper for variant discovery and genotyping in cattle

The most convenient way to run a *Graphtyper* version compiled for the bovine chromosome complement is to use *Docker* (which deals with all required dependencies). The command below starts to download modified *Graphtyper* software hosted at the Dockerhub:

```
docker run --rm cdanang/graphtyper_cattle graphtyper
```

We built the docker images using *Ubuntu* 18.04 as a base image. If you are working on a Linux 64-bit machine you could also get a static executable with command below. We placed the *Graphtyper* binary in /usr/local/bin) and executing command below will copy the *Graphtyper* binary from docker images to the current working directory:

```
docker run --rm -v ${PWD}:/io cdanang/graphtyper_cattle \
cp /usr/local/bin/graphtyper /io

### And then run the software as a standard binary
./graphtyper
```

If you prefer to modify and build a modified version of Graphtyper for the bovine chromosome complement directly from the source, please follow the instructions below:

1. Clone the *Graphtyper Github*

```
git clone --recursive https://github.com/DecodeGenetics/graphtyper.git
```

2. Create a new *branch* at this specific commit tag. We built graphtyper at this specific commit hash (04ab5ee460fa36129fb0d8ea5d4b72adc3836f52), to compile at the same software version that we use in the paper, please use this commit tag. We named the branch as *cattle modification*

3. Change directory into *graphtyper* and modify the chromosomal specifications in the files include $graphtyper/graph/absolute_position.hpp$ and *src/typer/vcf.cpp* using UMD 3.1 cattle chromosomal names and lengths.

```
git checkout -b cattle_modification \
04ab5ee460fa36129fb0d8ea5d4b72adc3836f52
```

The first modification enables all cattle chromosomes (esp. for chromosome number > 23) as the current software release set the maximum allowed length for each chromosomes according to the human GRChb37 and GRCh38. The second modifications are required that the respective chromosomal information is written to the *vcf header*.

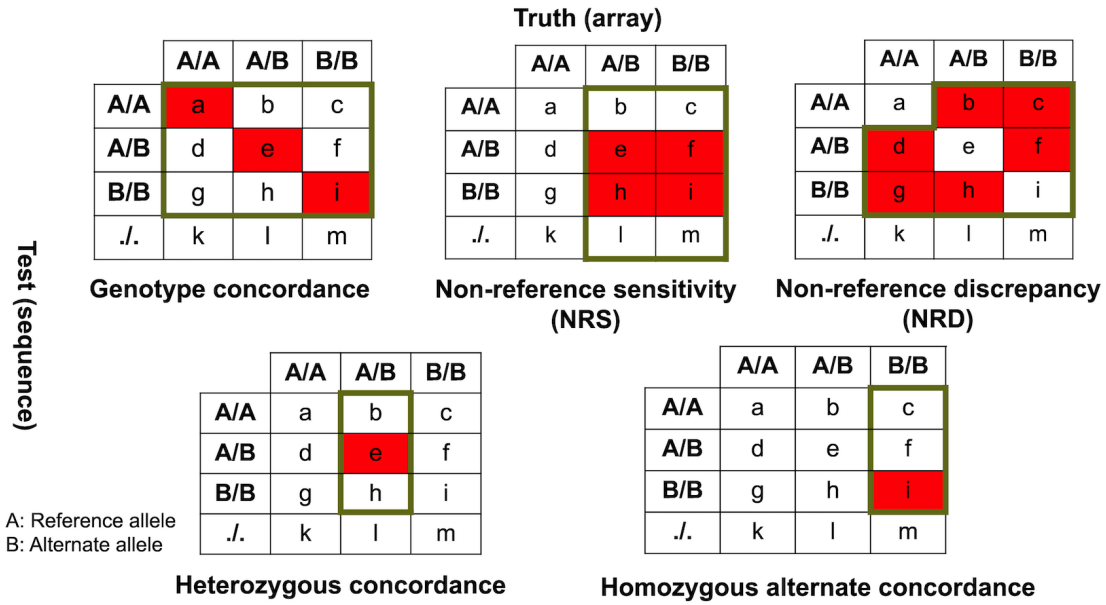
4. Make sure that these dependencies are installed:
    - C++ compiler with C++11 supported (we tested gcc 4.8.5 or gcc 6.3.0
    - Boost≥1.57.0
    - zlib≥1.2.8
    - libbz2
    - liblzma
    - Autotools, Automake, libtool, Make, and CMake¿=2.8.8
5. Follow installation procedures as below. This will put the software in *releasebuild/bin/graphtyper*

```
mkdir -p release-build && cd release-build
cmake ..
make -j4 graphtyper
bin/graphtyper # Run Graphtyper with modified cattle chromosome
specifications
```

Additional file 2.2

Properties of the different metrics used for the evaluation of sequence variant genotyping accuracy.

The metrics were calculated using the sum of the red cells as numerator and the cells within the green frame as denominator.

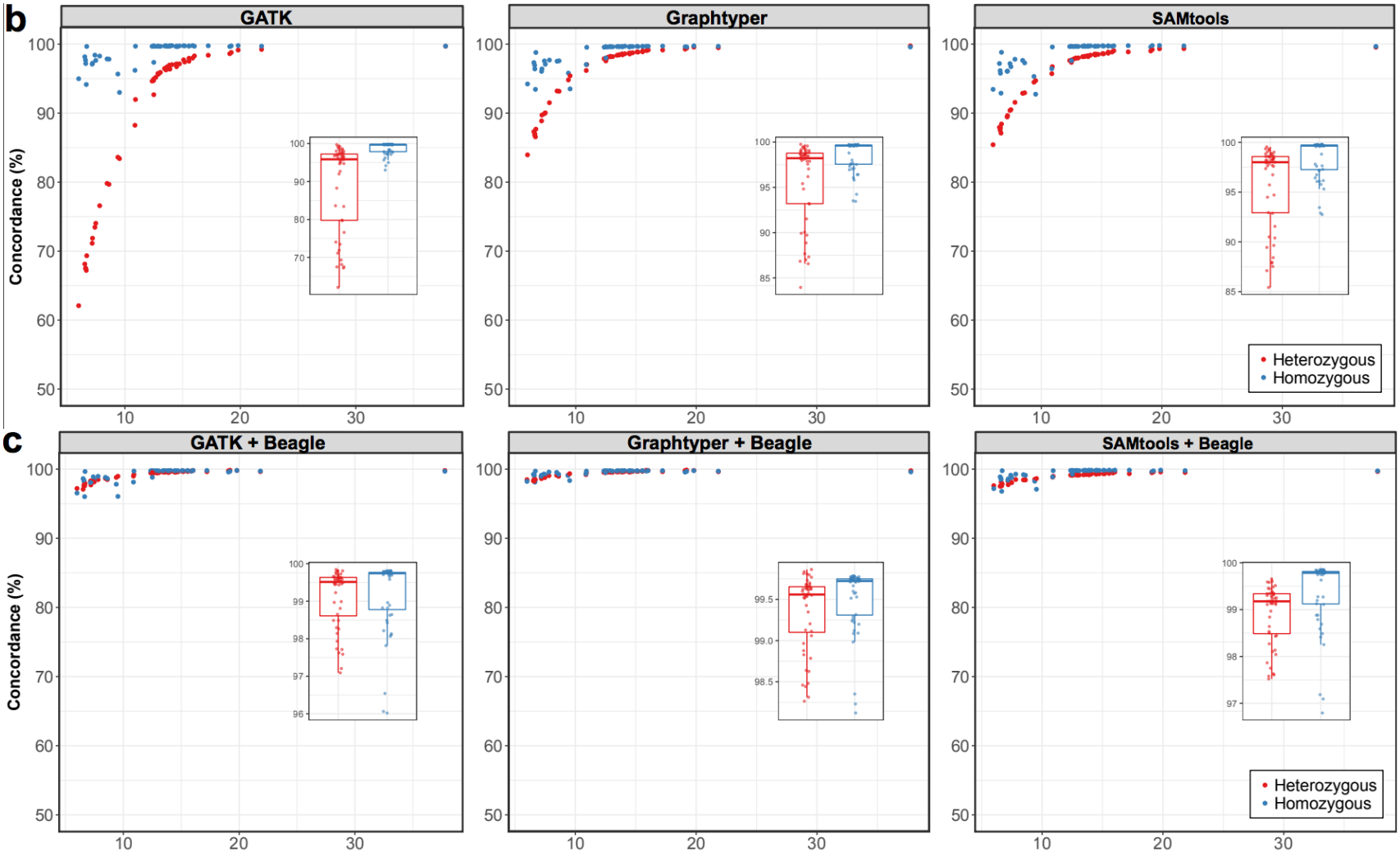


Additional file 2.3

Concordance statistics

The concordance of heterozygous and alternate homozygous genotypes in 49 Original Braunvieh cattle (**a**) and the concordance at the different sequencing depth for the (**b**) raw and (**c**) imputed datasets.

	Heterozygous concordance				Homozygous concordance			
	full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp
GATK	89.17	99.11	89.24	99.21	*98.74*	99.18	*98.75*	99.27
Graphtyper	*95.79*	*99.36*	*95.82*	*99.44*	98.55	*99.51*	98.59	*99.57*
SAMtools	95.73	98.91	95.77	98.99	98.46	99.37	98.48	99.41


b
GATK
Graphtyper
SAMtools
c
GATK + Beagle
Graphtyper + Beagle
SAMtools + Beagle
Concordance (%)
Heterozygous
Homozygous

Additional file 2.4

Sequence variant genotyping quality for 18 and 31 animals that were sequenced at a lower and higher than 12-fold sequencing coverage, respectively.

Asterisks denote significant differences with the best value (italic) for a respective parameter.

Coverage less than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	90.99***	98.7***	91.02***	98.82***	85.63***	98.91	85.51***	98.73	14.64***	2.09***	14.59***	1.91***
Graphtyper	94.89	99.07	94.91	99.17	96.44	99	96.13	98.71	8.04	1.49	8	1.31
SAMtools	94.87	98.61***	94.89	98.67***	96.24***	98.94	95.75***	98.45***	8.11	2.24***	8.09	2.11***

Coverage more than 12

	Genotype concordance				Non-reference sensitivity				Non-reference discrepancy			
	full		filtered		full		filtered		full		filtered	
	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp	raw	imp
GATK	98.73***	99.66	98.76***	99.71	98.3***	99.61	98.14***	99.39	1.8***	0.48*	1.76***	0.42
Graphtyper	99.26	99.67	99.3	99.72	99.25	99.54***	98.88	99.16***	1.04	0.45	0.99	0.4
SAMtools	99.21***	99.59***	99.24***	99.62***	99.21**	99.58***	98.51***	98.79***	1.12***	0.58***	1.08***	0.54***

Additional file 2.5

Twelve 1-Mb regions for which *Graphtyper* initially failed to genotype sequence variants

The algorithm either ran out of memory or exceeded the allocated runtime (12 h). Graphtyper eventually produced genotypes for the sequence variants when these regions were re-run in 10-kb segments.

No	Chromosome	Region (Mb)
1	1	0-1
2	1	145-146
3	3	69-70
4	7	58-57
5	8	110-111
6	12	76-77
7	23	26-27
8	23	29-30
9	26	50-51
10	27	37-38
11	28	39-40
12	29	30-31

Additional file 2.6

Variant filtration using *GATK*

The best practice guidelines for variant discovery using *GATK* recommend sequence variants to be filtered using Variant Quality Score Recalibration (VQSR) because it implements advanced machine learning-based methods to differentiate between true and false-positive variants. However, VQSR relies on sets of high confidence truth/training variants, which are currently not (publicly) available in cattle. Thus, we ran *GATK* with best practice recommendations for variant filtering when applying VQSR is not possible, i.e., we used a generic baseline hard-filtering threshold for each variant annotation (see https://gatkforums.broadinstitute.org/GATK/discussion/2806/howto-apply-hard-filters-to-acall-set). This threshold-based filtering is commonly applied the cattle genomics community (Koufariotis et al., 2018; Chen et al., 2018)

To facilitate running the VQSR module in sheep and goat, i.e., species where sets of truth/training variants are not (publicly) available, (Alberto et al., 2018) used an intersection of high confidence variants that had been discovered from multiple variant callers as truth/training sets, i.e., they derived truth/training sets directly from the analyzed data. We implemented their approach to apply *GATK* VQSR to our variant dataset. Training and truth sets were constructed using the overlap of the filtered variants from the GATK, *Graphtyper* and *SAMtools* pipelines (truth=false, training=true, known=false, prior= 10) and markers from the BovineHD BeadChip (truth=true, training=true, known=false, prior= 15), respectively. Moreover, we used variants listed in dbSNP (version 150) as known variants (truth=false, training=false, known=true, prior=3.0). Following *GATK* VQSR, we retained variants in the 99.9% tranche sensitivity threshold (best practice).

Variant filtration using *GATK* VQSR removed more variants from the raw data than *GATK* hard filtering (Table 1). However, VQSR retained more HD SNPs than *GATK* hard filtering, possibly reflecting bias that results from the use of HD SNPs as training/truth sets. The values of the concordance statistics (genotype concordance, non-reference sensitivity, nonreference discrepancy) were almost identical between *GATK* VQSR and *GATK* hard filtration (Table 2) indicating that the choice of either filtration option does not notably affect the concordance between sequence-derived and BovineHD SNP array-derived genotypes. These findings are in line with (Vander Jagt et al., 2018) who showed that the concordance between microarray-called and sequence-derived genotypes is almost identical using either *GATK* VQSR or the *GATK* 1000 bull genomes project hard filters, even though they used stringently filtered truth/training sets based on a more comprehensive catalogue of variants than in our study. Interestingly, in agreement with (Vander Jagt et al., 2018), the proportion of opposing homozygous genotypes in sire/son-pairs (which does not suffer from ascertainment bias because

it is calculated using sequence-derived SNPs) is less using *GATK* hard filter than *GATK* VQSR.

The performance of *GATK* VQSR may be assessed using the novel variant sensitivity tranche plot (Figure 2). In the lowest 90% tranches (highest specificity) the filtering model still retained many false positive variants (orange box and low Ti/Tv ratio). However, when the 99.9% tranche sensitivity is used as filtration criterion as recommended by the *GATK* best practice guidelines, a high proportion of true positive variants is removed from the data.
Overall, our findings suggest that

(i) *GATK* VQSR removes more variants from the data than *GATK* hard filtering,

(ii) *GATK* VQSR does not notably improve the concordance between sequencederived and microarray-called genotypes compared to *GATK* hard filtering,

(iii) the proportion of opposing homozygous genotypes in sire/son-pairs is higher using *GATK* VQSR than *GATK* hard filtering, and

(iv) improving VQSR may be possible by providing more sophisticated truth/training variant datasets produced by orthogonal sequencing technology other than the ones used for training, e.g. (Li et al., 2018)

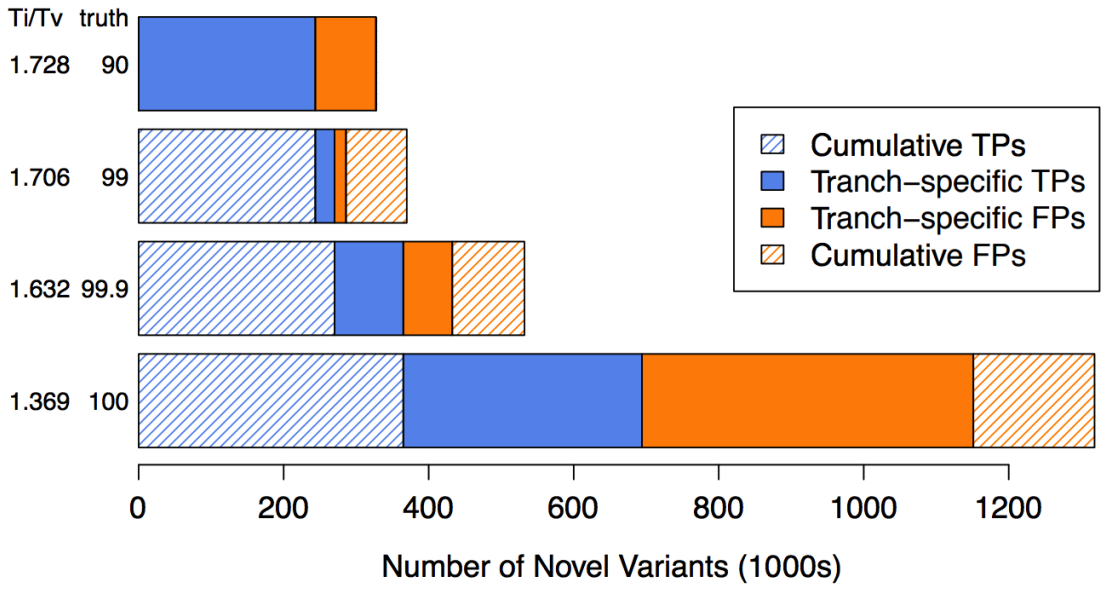
Table 1 Comparison of variants statistics between unfiltered and filtered datasets using either hard-filtering or VQSR.

	GATK full	*GATK* hard-filter	*GATK* VQSR
Total SNPs	18,594,182	17,248,593	16,537,577
Biallelic	18,347,962	17,111,806	16,430,734
Multi-allelic	246,220	136,787	106,843
Ti/Tv ratio	2.09	2.17	2.16
BovineHD	99.46	99.21	99.38
BovineSNP50	99.14	98.91	98.98

Table 2 The concordance statistics between hard-filtered and VQSR

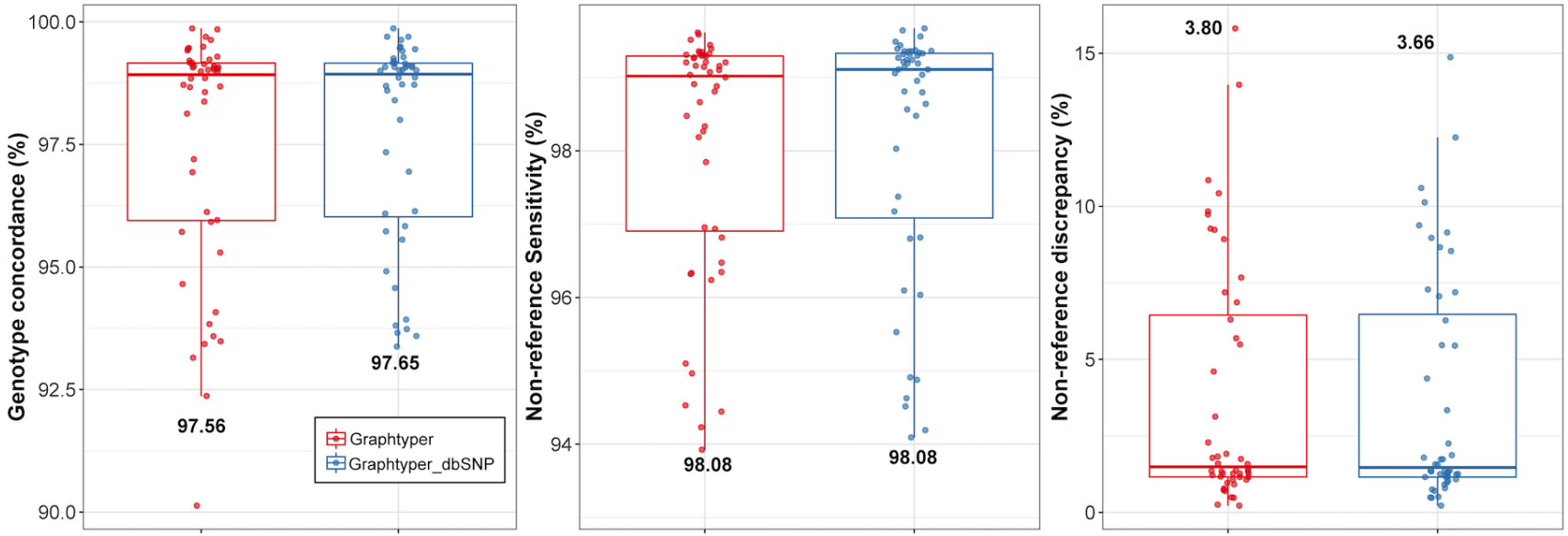
	Genotype concordance	Non-reference sensitivity	Non-reference discrepancy	Opposing Homozygous
GATK hard-filter	96.02	93.67	6.3	0.72
GATK VQSR	96.01	93.77	6.32	0.75

Figure 1 Tranche sensitivity plot of novel variants as reported by the VQSR model fitting



Additional file 2.7

Accuracy and sensitivity of sequence variant genotyping on bovine chromosome 25 from a variation-aware genome graph that incorporated 2,143,417 dbSNP variants as prior known variants.



Supplementary References

F. J. Alberto, F. Boyer, P. Orozco-terWengel, I. Streeter, B. Servin, P. De Villemereuil, B. Benjelloun, P. Librado, F. Biscarini, L. Colli, et al. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1):1–9, 2018.

N. Chen, Y. Cai, Q. Chen, R. Li, K. Wang, Y. Huang, S. Hu, S. Huang, H. Zhang, Z. Zheng, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in east asia. *Nature Communications*, 9(1):1–13, 2018.

L. Koufariotis, B. Hayes, M. Kelly, B. Burns, R. Lyons, P. Stothard, A. Chamberlain, and S. Moore. Sequencing the mosaic genome of brahman cattle identifies historic and recent introgression including polled. *Scientific reports*, 8(1):1–12, 2018.

H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, and D. MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8):595–597, 2018.

C. Vander Jagt, A. Chamberlain, R. Schnabel, B. Hayes, and H. Daetwyler. Which is the best variant caller for large whole-genome sequencing datasets. In *Proceedings of the 11th world congress on genetics applied to livestock production*, pages 11–16, 2018.