INVESTIGATIONS

Improving imputation quality in BEAGLE for crop and livestock data

T. Pook[*†,1], **M. Mayer**[‡], **J. Geibel**[*†], **S. Weigend**[†§], **D. Cavero**[**], **C.C. Schoen**[‡] **and H. Simianer**[*†]

[*]University of Goettingen, Department of Animal Sciences, Animal Breeding and Genetics Group, 37075 Goettingen, Germany, [†]Center for Integrated Breeding Research, University of Goettingen, 37075 Goettingen, Germany, [‡]Technical University of Munich, Plant Breeding, TUM School of Life Sciences Weihenstephan, 85354 Freising, Germany, [§]Friedrich-Loeffler-Institut, Institute of Farm Animal Genetics, 31353 Neustadt-Mariensee, Germany, [**]H&N International, 27472 Cuxhaven, Germany

ABSTRACT Imputation is one of the key steps in the preprocessing and quality control protocol of any genetic study. Most imputation algorithms were originally developed for the use in human genetics and thus are optimized for a high level of genetic diversity. As the software BEAGLE offers the user considerable flexibility to tune the algorithm to the specific genetic structure of the respective dataset. Different versions of BEAGLE were evaluated on genetic datasets of doubled haploids of two European landraces in maize, a commercial breeding line and a diversity panel in chicken, respectively, with different levels of genetic diversity and structure. BEAGLE 5.0 showed the best performance and was less dependent on adapted parameter settings than the earlier versions. For all versions, the parameter of the effective population size had a major effects on the error rate for imputation of ungenotyped markers, reducing error rates by up to 98.5%. For BEAGLE 4.0 and 4.1 imputation accuracies were further improved by tuning parameters like modelscale, buildwindow and nsamples. The number of markers with extremely high error rates for the maize datasets were more than halved by the usage of a flint reference genome (F7, PE0075 etc.) instead of the commonly used B73. On average, error rates for imputation of ungenotyped markers were reduced by 8.5% by excluding genetically distant individuals from the reference panel. Strategies to find a balance between representing as much of the genetic diversity as possible while avoiding the introduction of noise by including genetically distant individuals are discussed.

KEYWORDS
imputation
BEAGLE
reference panel
reference genome

INTRODUCTION

Imputation is one of the key steps in preprocessing genetic data generated by SNP-chips or DNA sequencing, as later applications like genomic prediction (Meuwissen *et al.* 2001) often do not allow for missing values. In some applications the usage of a higher marker density can lead to better results even though individuals were not genotyped for most markers (e.g. in genome-wide association studies previously not identified regions can be detected (Yan *et al.* 2017)). Over the years a wide variety of methods and corresponding programs like BEAGLE (Browning *et al.* 2018), MiniMac (Das *et al.* 2016) and Impute (Howie *et al.* 2009) have been developed and improved to account for the increasing number of individuals and marker densities in genetic studies. All these methods are based on Hidden Markov Models (HMM) (Baum and Petrie 1966; Rabiner 1989) which were introduced to genetic imputation by Li and Stephens (2003). To account for the specific structure of livestock and crop datasets, special tools for both cases have been developed. As fully homozygous lines are especially relevant present in crops, the software TASSEL (Bradbury *et al.* 2007) was constructed to work well on this data structure (Swarts *et al.* 2014). An example from livestock is FImpute (Sargolzaei *et al.* 2014), that focuses on using pedigree information in the imputing process and is able to process a high number of individuals, as present in modern cattle breeding programs, with linear increase in computation time. In the imputation process all those methods use the fact that physically close markers are likely inherited together, resulting in non-random associations of alleles. These methods thereby rely on the knowledge of position or at least

Manuscript compiled: Saturday 16th March, 2019
[1]T. Pook; University of Goettingen, Department of Animal Sciences, Center for Integrated Breeding Research, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany
Email: torsten.pook@uni-goettingen.de

the physical order of markers for modeling linkage and thus the resulting linkage disequilibrium (LD). In contrast, the software LinkImpute (Money *et al.* 2015) accounts for LD between pairs of markers and not their physical positions. This can be especially relevant for species in which no reference sequence is available or whose genomes are known for a high amount of translocations and inversions (e.g. maize).

In contrast to other methods using a HMM, the markov chain in BEAGLE is not initialized by the genotypes or haplotypes themselves, but instead the genetic dataset is used to initialize a haplotype cluster (Browning and Browning 2007), which subsequently initializes the HMM. Imputation is then performed by basically identifying the most likely path through the haplotype cluster based on the non-missing genotypes.

As BEAGLE is originally developed for application in human genetics, default settings are chosen to work well for imputation in outbred human populations. Nevertheless the user still has considerable flexibility to tune the algorithm to the specific genetic structure of the respective dataset. As imputation oftentimes is just one step in the preprocessing and quality control protocol, authors tend to use the default settings of a recent version of some imputation software.

To increase the operational marker density via imputation an additional dataset (reference panel) genotyped under a higher density can be used. With increasing computational power and more efficient methods available the common advice here is to use as many individuals as possible to get a good representation of the population (Zhang *et al.* 2013; Browning *et al.* 2018).

In this paper, we compare different BEAGLE versions and analyze the influence of different parameter settings in BEAGLE on imputation quality for a variety of livestock and crop datasets. We further evaluate which individuals to include in a reference panel when aiming at increasing the marker density of a dataset.

Since imputation algorithms like BEAGLE rely on the assumed physical order of markers, the used reference genome influences the imputation quality. Recently, a variety of new reference genomes have been made public (Unterseer *et al.* 2017). We here compare the imputation performance of the commonly used B73v4 (Schnable *et al.* 2009; Jiao *et al.* 2017) and new reference genomes from flint lines in maize that should be genetically closer to our used material. All reference genomes derived in chicken were generated based on an inbreed Red Jungle Fowl (*Gallus gallus gallus*) that was used in all tests (International Chicken Genome Sequencing Consortium 2004; Bellott *et al.* 2010).

For all tests we considered BEAGLE 4.0, 4.1, and 5.0 (Browning *et al.* 2018).

MATERIALS AND METHODS

Genotype data used

In the following, we will consider genotypic data of 910 doubled haploid (DH) lines of two European maize (*Zea mays*) landraces ($n = 501$ Kemater Landmais Gelb (KE) and $n = 409$ Petkuser Ferdinand Rot (PE)) genotyped using the 600k Affymetrix® Axiom® Maize Array (Unterseer *et al.* 2014). Markers were filtered for being assigned to the highest quality class (Poly High Resolution (Pirani *et al.* 2013)), having a callrate >90%, and for having <5% heterozygous calls, as no heterozygous calls are expected for DH lines. The remaining heterozygous calls were set to NA and subsequently imputed using BEAGLE 4.0 with nsamples=50, resulting in a dataset of 501'124 markers with known haplotype phases.

We further considered two chicken (*Gallus gallus*) datasets genotyped with the 580k SNP Affymetrix® Axiom® Genome-Wide Chicken Genotyping Array (Kranis *et al.* 2013). Firstly, a chicken diversity panel containing 1'810 chicken of 82 breeds including Asian, European and wild types, but also commercial broilers and layers (Weigend *et al.* 2014). Secondly, a dataset containing 888 chicken of a commercial breeding program from Lohmann Tierzucht GmbH. For quality control SNPs/animals with less than 99%/95% callrate were removed. We will here focus on chromosome 1, 7 and 20 with 56'773/65'177, 12'585/13'533 and 5'539/5'940 SNPs representing cases for large, medium and small size chromosomes in the diversity/breeding panel. Both chicken panels were imputed using BEAGLE 4.1 default.

For tests regarding imputation of ungenotyped markers in maize we used the overlapping markers (45'655 SNPs) of the Illumina® MaizeSNP50 BeadChip chip (Ganal *et al.* 2011) as a smaller SNP array. As there is no similar smaller array with a majority of overlapping markers for the chicken panels, we simply used a subset of every tenth marker. All tests regarding imputation quality were performed on imputed datasets. This should favor the respective method used for the imputation. As the missingness in the maize data (1.20%), diversity panel (0.27%) and commercial chicken breeding line (0.32%) were low in the raw data, this effect should only be minor and is neglected here.

To assess the genetic diversity of the three datasets, we derived the LD decay (Figure 1) resulting in the highest rates of association for the European maize landraces, followed by the commercial chicken dataset and the chicken diversity panel. All used datasets show far smaller effective population sizes than an outbred human population. It should be noted that this comparison does not account for possible differences in ascertainment bias (Albrechtsen *et al.* 2010) between the arrays or genetic diversity of species and their genomes. Since BEAGLE (and other HMM based imputation methods) are relying on local associations between markers this should still be a good indication for potential imputation performance.

Evaluation Pipeline

The imputation process itself can be split up into three internally linked steps which can be of different importance based on the data at hand and, in the following, will be analyzed separately:

1. Inference: All partly or fully missing individual genotypes in the actual dataset are completed, but no additional markers are added.

2. Imputation of ungenotyped markers (UM imputation): Additional markers are added to the genetic data based on information provided by a second dataset (reference panel) with higher marker density.

3. Phasing: The two haplotypes of diploid individuals, i.e. their gametic phases, are estimated from genotype data.

To assess the quality of inference and UM imputation we used the following testing pipeline and repeated the procedure 50 times for each test. We start from a completed dataset in which missing genotypes have been imputed, and consider this as the "true" genotype dataset:

1. Randomly generate missing values (NAs) in the "true" genotype dataset.

    - In case of inference set randomly chosen alleles of all genotypes to NA (in our case: 1% of all alleles with no partly missing genotypes).

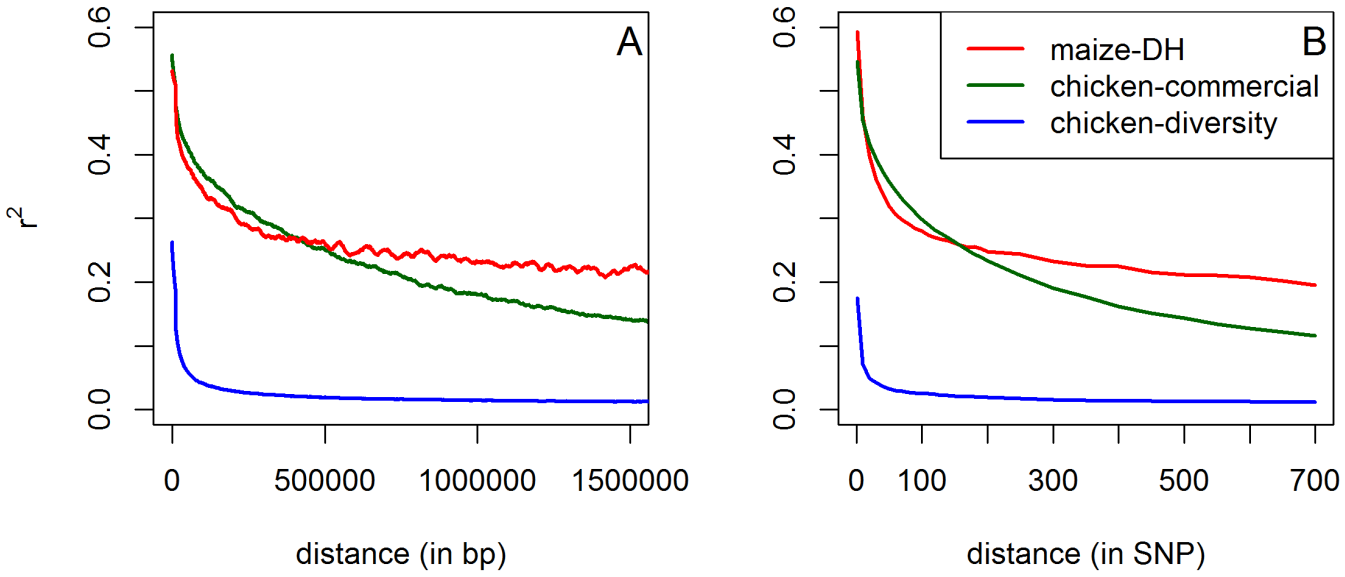

Figure 1 LD decay based on physical length (A) and marker distance (B) for chromosome 1 for all considered datasets. Outliers in (A) are corrected for by using a Nadaraya-Watson-estimator (Nadaraya 1964), using a Gaussian kernel and a bandwidth of 50 kb. (B) is using average values for each SNP distance.

- In case of UM imputation additionally set all entries in a particular marker to NA (maize: according to existing low density array (Ganal *et al.* 2011); chicken: 90% of all markers).

2. Perform the imputation procedure under a chosen parameter setting, software and potential use of a reference panel.

3. Evaluation of performance by comparison to the "true" dataset (for more on this we refer to the following subsections).

Evaluation of imputation quality

To evaluate the quality of inference and UM imputation we count the total number of entries in the genotype matrix different to the "true" dataset. In this procedure, markers with a low minor allele frequency have a lower influence on the overall quality than in the commonly used practice of calculating the correlation between imputed and "true" dataset. To account for this, we will provide error rates depending on the allele frequency as well. A disadvantage of using a correlation is that it does not account for fixed markers (correlation not defined) and those markers thus have to be excluded from the analysis. As rare variants tend to be more difficult to impute and those variants tend to be fixed at a higher rate, this leads to lower average correlations for methods imputing a rare allele. For a fair comparison only those markers that are not fixed over all settings/software should be used. Especially for the imputation of ungenotyped markers this would lead to a much smaller set of markers to be considered for a fair comparison.
To evaluate phasing quality we use the switch error rate as defined in (Lin *et al.* 2002), which evaluates the number of switches between neighboring heterozygous sites to recover the true haplotype phase compared to the total number of heterozygous markers and thereby chances for switch errors to occur.

Evaluation of phasing quality

The evaluation of phasing quality is more complex since the true haplotype phase is usually not known and there is a potential bias towards the method that was used to derive the haplotype phase. Since we are working with doubled haploid lines in the maize dataset, the true gametic phase is known and a "true" dataset for testing was generated by randomly combining two doubled haploid lines to a Pseudo S_0. The rest of the pipeline can be performed in the same way as the inference testing. Additionally, we considered datasets with no missing genotypes to remove any possible noise caused by inference errors.

Choice of reference panel in UM imputation

A common first question in generating genetic data is how many individuals need to be genotyped with high density to obtain sufficient imputation quality for individuals genotyped with lower marker density. To evaluate this, we performed imputation on datasets containing 50 individuals as the "true" dataset in our pipeline and generated reference panels containing 25, 50, 100, 150, 200, 250, 300, or 350 individuals, respectively.
In a second step, we ask the question which individuals to include in a reference panel. This is especially relevant if possible candidates for the reference panel vary in their relationship to the dataset. For this, we split the chicken diversity panel into 10 subpopulations by iteratively minimizing the total sum of squared genetic distances between breeds within the subpopulations. Distances between the breeds were calculated as Nei standard genetic distances (Nei 1972). In a first step, the custom made algorithm randomly assigned the breeds to 10 equal sized subpopulations. The contribution of each breed to the sum of squared distances was calculated and the algorithm started iteratively exchanging the most noisy breeds to other subpopulations. If there was a reduction of the total sum of squared distances within the subpopulations, the exchange was accepted and the contributions were calculated again. The process was repeated until no exchange could improve the fit. To overcome results depending on specific starting positions, the process was repeated for 60 random starting points. Nei standard genetic distances for evaluation of UM imputation quality of BEAGLE were calculated based on the subpopulation assignment of individuals and UM imputation was

performed using the following reference panels:

(A) All other individuals of the same subpopulation

(B) All individuals of one other subpopulation

(C) All individuals of all other subpopulations

(D) All individuals of subpopulations with less than average Nei standard genetic distance to the dataset

(E) All individuals of those subpopulations with reduced error rates when testing A + B compared to A as the reference panel

Additionally combinations of panels A + B, A + C, A + D and A + E were tested. Tests were repeated 20 times for each subpopulation with datasets containing 50 randomly sampled individuals. For each dataset, all different reference panels were tested.

Data Availability

Genetic data for chromosome 1 for all three panels used are available at https://github.com/tpook92/HaploBlocker. Supplemental files are available at FigShare.

RESULTS AND DISCUSSION

In the following, we will use BEAGLE 4.1 on default settings as the standard and compare all results to it. We will here focus on showing the obtained error rates under each parameter setting and not extensively discuss their influence on the imputation algorithm itself. For details on that we refer to the BEAGLE publications (Browning 2006; Browning and Browning 2007, 2013a,b, 2016; Browning *et al.* 2018).
Unless otherwise mentioned, we will report the error rates in the landrace KE averaged over all chromosomes as the maize data. Results for PE were similar with on average slightly increased error rates. For details on that we refer to Supplementary Figures S3, S4, S5, S6 and Table S1.

Inference quality

When comparing error rates for inference under multiple settings in BEAGLE, one can observe major differences. On default settings in BEAGLE 4.1, we obtained an average error rate of 0.255% for the maize data. Error rates are significantly higher for alleles with low frequency (Figure 2). In regard to the location of inference errors one can observe a high volatility with a tendency to have massively increased error rates in telomeric regions (Figure 3). Additionally error rates in regions of high LD tend to be lower (Supplementary Figure S8).

BEAGLE 4.0 leads to similar error rates (0.201%) whereas BEAGLE 5.0 clearly outperforms previous versions (0.014%) on default settings. By tuning parameter settings, especially results in older version can be improved, leading to error rates of 0.031% in BEAGLE 4.0 (buildwindow = 25 & nsamples = 25 & burnin-its) and 0.043% in BEAGLE 4.1 (modelscale = 1.5), whereas no significant improvements can be made in BEAGLE 5.0. Improvements in overall inference quality can be observed for all allele frequency classes and regions in the genome (Figures 2 & 3). When using slightly lower values for buildwindow (e.g. 10) in BEAGLE 4.0, one can observe a further reduction of the average error rate, but also a massive increase of the error rate in some single markers (Supplementary Figure S1).
Especially the parameters buildwindow (default: 1'200) and modelscale (default: 0.8) have a major impact on the inference quality (Figure 4) as both parameters implicitly control how different haplotypes in the haplotype cluster can be while still being considered jointly. Both a lower value for buildwindow and a higher value of modelscale will lead to less similar haplotypes to be clustered jointly. The parameter singlescale in BEAGLE 4.0 has the same effect on the haplotype cluster as modelscale in BEAGLE 4.1 but performed slightly worse in terms of inference quality than buildwindow. Tuning of both parameters jointly did not further improve performance.

When working with Pseudo S_0 instead of DH-lines, an additional source of noise is introduced as haplotype phase is not known. Nevertheless error rates are decreasing on default settings for both BEAGLE 4.0 (0.125%) and BEAGLE 4.1 (0.040%) compared to inference for DH-lines. This is consistent with what is reported in (Swarts *et al.* 2014). A possible explanation for this is that in contrast to a DH dataset, imputation of the value 1 instead of only 0 and 2 is possible in case the algorithm is indifferent as to which allele to impute. BEAGLE 5.0 (0.0168%) seems to fix those issues and inference quality is on a similar level for DH-lines and S_0.
The distribution of errors and the ideal parametrization stays similar (buildwindow slightly higher, modelscale slightly lower) with the exception of an added benefit of increasing the number of iterations used to generate the haplotype cluster. As the algorithm starts with randomly phased genotypes and improves the phase in each iteration, this should not be surprising. Overall error rates on tuned parameter settings are 0.026% for BEAGLE 4.0 (using buildwindow = 50 & nsamples = 25 & phase-its = 25 & burnin-its = 25), 0.018% in BEAGLE 4.1 (using modelscale = 1.0 & niterations = 25) and 0.0166% in BEAGLE 5.0 (using iterations = 25). For a detailed comparison and the share of improvement each parameter contributes we refer to Figure 5.

The error rates for the considered chicken diversity panel are higher (1.13%) than for the maize data using BEAGLE 4.1 default and possible improvements in all three considered versions are relatively small (BEAGLE 4.0 - 1.01% using buildwindow = 25, nsamples = 25, burnin-its = 25; BEAGLE 4.1 - 0.80% using modelscale = 1.25, niterations = 25; BEAGLE 5.0 - 0.82% default & 0.81% using burnin = 25). Overall improvements are obtained by changing parameters in the same direction as in the maize dataset but effects are much smaller. As the chicken diversity panel contains much more variation and is structurally more similar to outbred human data than the European landraces in maize, this should not be that surprising. The dataset from the commercial chicken breeding program showed error rates between 0.200% and 0.230% for basically all tested settings, leading us to conclude that inference on this dataset there is not much potential to decrease error rates. A potential reason for this is that other error source like SNP calling errors may already be higher than error rates on default.

UM Imputation quality

When performing UM imputation, error rates were much higher than in the inference case. Overall error rates decreased with the size of the reference panel (Figure 6). Especially for high error rates the relative gain of a larger reference panel in BEAGLE 5 is higher than in BEAGLE 4.1. When using 350 DH-lines of KE as a reference panel we obtained an average error rate of 6.59% on default settings and 3.09% in BEAGLE 5. In all our testing the parameter effective population size ne with default 1'000'000 was found to have a major impact on the UM imputation error rates (Figure 7). Tuning the effective population size leads to error rates of 0.096% in BEAGLE 4.1 (ne = 300) and 0.088% in BEAGLE 5.0 (ne = 1'000). In BEAGLE 4.0, there is no parameter to control the

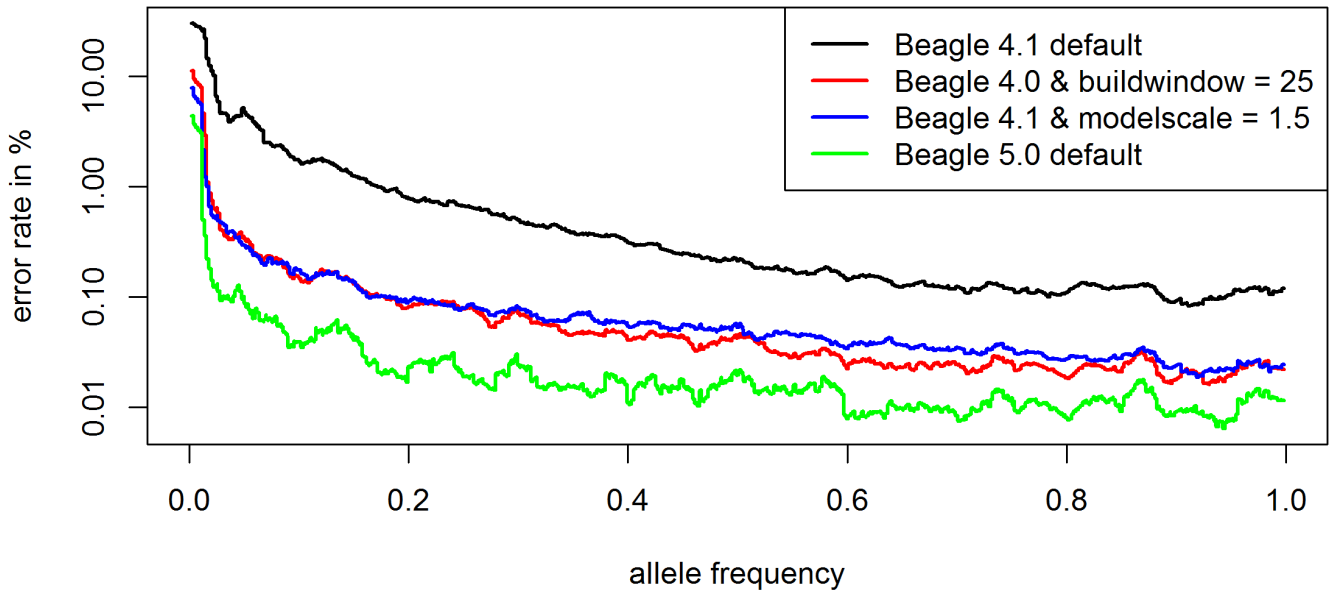

Figure 2 Error rate depending on the allele frequency under different BEAGLE settings for the maize data. Y-axis is log-scaled.

effective population size but default settings work slightly better (5.15%) than in BEAGLE 4.1. The effect of other parameters is rather small and relative differences of tuning can only be observed after adaptation of ne. Error rates were minimized for BEAGLE 4.0 (0.80%) by using buildwindow = 100 & nsamples = 25. Since there are less informative markers in a window of 100 SNPs than in the case of inference, an increase of buildwindow also makes sense from a modeling perspective. For BEAGLE 4.1 we used modelscale = 1.5 & ne = 300 leading to an error rate of 0.088%. In BEAGLE 5.0, there was only a minor improvement (0.087%) by use of imp-states = 500 & ne = 1'000. Overall the relative effect of the size of the reference panel became stronger after tuning the parameter settings (Figure 6).

Similarly, error rates for the diversity chicken panel (BEAGLE 4.1: 3.69% / BEAGLE 5.0: 3.31%) and the commercial breeding line (0.95%/0.77%) on default are higher than inference error rates. Error rates for the diversity panel were reduced to 2.73%/2.48% by using ne = 3'000. For the breeding line ne = 1'000 worked best, leading to error rates of 0.28%/0.28%. Overall error rates of UM imputation for BEAGLE 4.1 were reduced by 98.5% in the maize landrace, 70.5% for the commercial chicken line and 26.0% for the chicken diversity panel. With this, the bigger gains by tuning the effective population size nicely support our expectation of the effective population sizes of the underlying populations. Additionally, BEAGLE 5.0 was more robust to changes in the effective population size than BEAGLE 4.1 (Figure 7) and overall error rates differ only by 0.013% for an effective population size between ne = 1 and ne = 10'000 for the maize dataset, indicating that the usage of any reasonable value should work here in a robust way. As the default of 1'000'000 is not realistic for most livestock and crop datasets, adaptation is necessary and critical when performing UM imputation.

Phasing quality

When using a dataset with no missing values, we observed a switch error rate of 0.0170% for the maize data using default settings in BEAGLE 4.1. This is equivalent to one switch error per 5'876 heterozygous markers. Error rates for BEAGLE 5.0 are similar (0.0163%), while BEAGLE 4.0 is clearly outperformed (0.0405%). By tuning parameter settings this can even be improved to an error rate of 0.0136% by usage of burnin = 2 & phase-segments = 10 & phase-states = 500 & iterations = 40 in BEAGLE 5 with phase-segments having the biggest impact. Overall, the relative effect is small and these differences should not have a major impact for most applications that require genotype phase. When working with datasets containing 1% missing values, error rates overall are similar.

Comparison of reference genomes

The most commonly used reference genome in maize genetics is the dent line B73 (Schnable *et al.* 2009; Jiao *et al.* 2017). The used European landraces are considered as flint germplasm with possible major differences in the physical map (Unterseer *et al.* 2016). After reducing error rates of inference by choosing appropriate parameter settings (here: BEAGLE 4.0 with buildwindow = 50), markers with high error rates tend to be clustered (Figure 8). Markers and regions with high inference error rate can be considered as candidates for misalignment in the genetic map. To compare our results obtained with B73v4 (Jiao *et al.* 2017), we additionally used reference genomes of the flint lines F7, EP1, DK105 and PE0075 (Unterseer *et al.* 2017).

Since the array itself was constructed using B73 as a reference (Unterseer *et al.* 2014) more markers can be mapped to the B73 reference than to the other reference genomes. For those markers mapped to both B73 and one of the flint reference genomes average error rates for inference are reduced by 3-5% (Table 1). The main factor for this is a significantly lower number of markers with extremely high error rates. The overall number of markers with error rates above 10% (here referred to as: "critical" markers) on average is reduced by 57%. For a detailed list of the "critical" markers for all reference genomes mapped on the 600k array (Unterseer *et al.* 2014), we refer to Supplementary Table S2 & S3. We found no notable difference between the inference quality for PE when using PE0075 as the reference genome compared to other

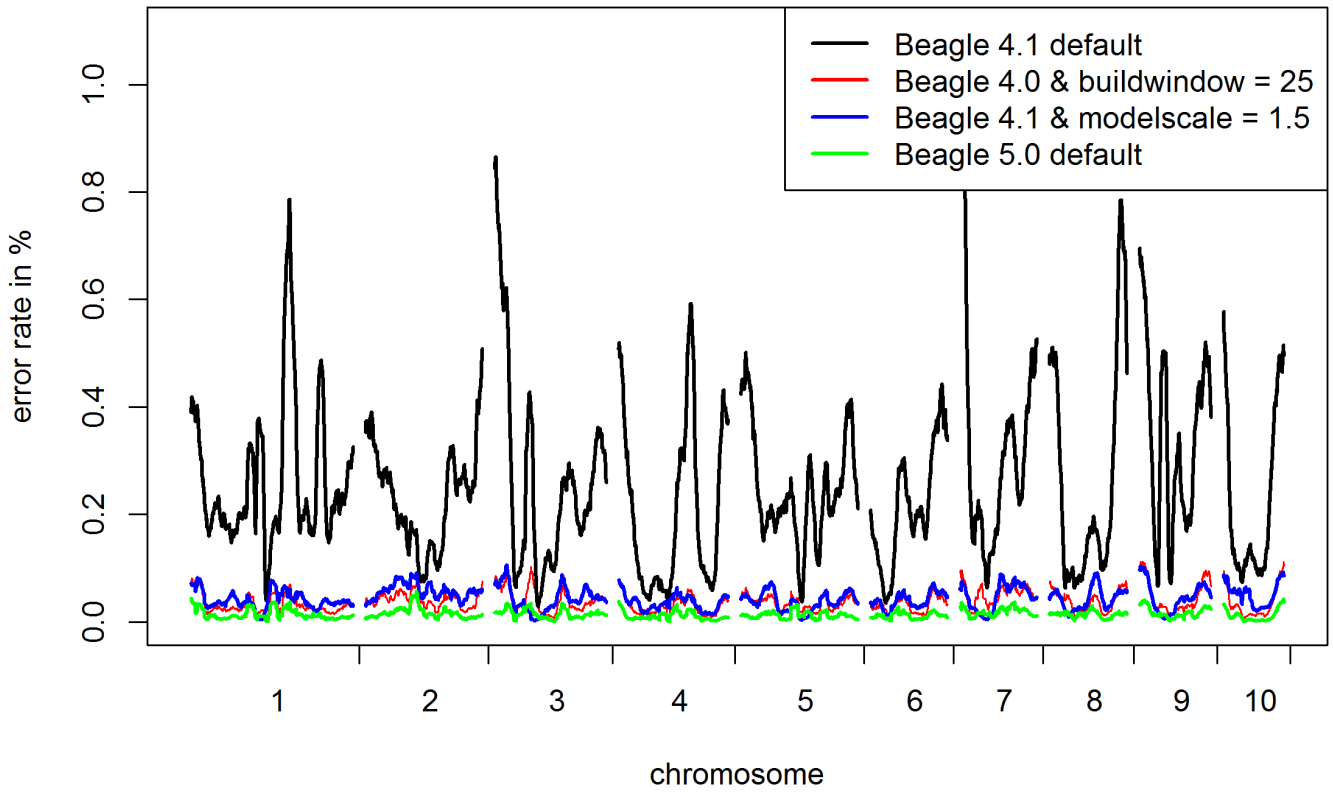

Figure 3 Inference error rate based on the location of the genome. Outliers are corrected for by using a Nadaraya-Watson-estimator (Nadaraya 1964), using a Gaussian kernel and a bandwidth of 3′000 markers for the maize data.

flint references (Supplementary Table S1).

Size of the reference panel

For all tests in this subsection, we used BEAGLE 5.0 with an effective population size of ne = 10′000. As already shown before, the error rate of UM imputation is decreasing when increasing the number of individuals in the reference panel (Figure 6). The relative effect of this improvement is highest when using appropriate parameter settings, whereas there is only a minor change in UM imputation quality with default settings. It should be noted that the minimum size of the reference to get adequate results is highly dependent on the dataset. As a rule of thumb, one can say that datasets containing more diversity in general need more individuals in the reference panel for similar UM imputation quality.

Choice of the reference panel

In case the reference population has a lot of stratification, the design of a good reference population becomes more difficult as genetically distant individuals may introduce more noise than relevant information. When comparing results for all considered reference datasets for UM imputation of a single subpopulation it becomes apparent that UM imputation without other individuals from the same subpopulation leads to extremely high error rates (>15%) and thus should in practice only be performed with extreme caution. In terms of including other subpopulations in the reference panel, the answer becomes less clear. When including single other subpopulations in the reference panel we observe significant effects on the overall error rate of UM imputation. Absolute differences of UM imputation error rates are between -0.307% and +0.604% with overall error rates between 1% and 4%. For a detailed list containing all changes in error rates when including a single other subpopulation in the reference panel, we refer to Supplementary Table S1. It should be noted that subpopulations with lower genetic distance to the dataset tend to reduce the error rate and less related subpopulations lead to increased error rates (Figure 9).

For all ten subpopulations the slope of the error rate in regard to distance to the subgroup is statistically significantly positive with the main difference between the subpopulations being the intercept. The most extreme case for this is subpopulation 6 (turquoise △ in Figure 9; including wild types). For this group the inclusion of any other subpopulation in the reference panel decreases the imputation quality and is ignored for all averages and statistics in this subsection. Even though SNP based genetic distance to other subgroups is relatively low, one can assume a long time since the last common ancestor to any other subpopulation and thus a lack of conserved haplotypes. Overall imputation quality when using a reference panel containing all subpopulations is worse than using a reference panel with only those subpopulation with below average genetic distance (Nei 1972) to the dataset (2.25% vs. 2.18% - Figure 10). Even though results are statistically significant (two-sample t-test: p-value: 0.0117), differences are minor and probably of limited practial relevance for most applications. In our analysis a reference panel containing only the individuals of the same subpopulation on average lead to an UM imputation error of 2.26% with no statistically significant difference to reference

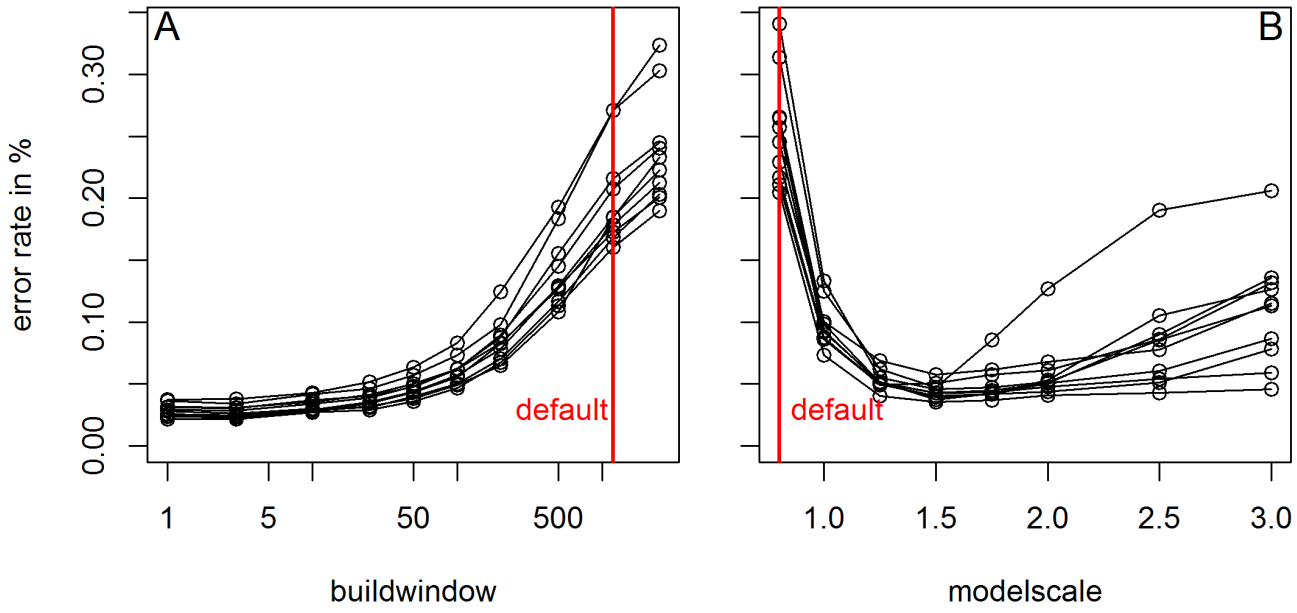

Figure 4 A: Error rate depending on the parameter buildwindow in BEAGLE 4.0 in the maize data. B: Error rate depending on the parameter modelscale in BEAGLE 4.1. Error rates are given for all ten chromosomes separately.

Table 1 Inference error rates using different reference genomes compared to B73 for KE DH-lines. Only markers mapped on both the flint reference genome & B73v4 (Jiao *et al.* 2017) are considered for "critical" markers (error rate > 10%).

Reference genome	F7	EP1	DK105	PE0075
Overlapping markers to B73v4	352'326	342'037	338'882	338'244
"Critical" markers when using this map	109	113	115	114
"Critical" markers when using B73v4	271	264	262	262
Relative change in error rate	- 5.11%	-3.87%	-4.68%	-3.32%

panels containing all subpopulations. When performing in-depth analysis for which regions of the dataset UM imputation quality is improved, we observed that especially those individuals with rare variants and overall higher error rates benefited from including more samples in the reference. On the other hand, already well imputed individuals usually had similar or slightly increased error rates. When using a reference panel containing all those subpopulations that individually lead to reduced error rates, average error rates are reduced to 2.06%. It should be noted that, in practice, a selection based on error rates in UM imputation is usually not possible. Nevertheless the result demonstrates that there is some potential in using more sophisticated approaches than just selecting all subpopulation with below average Nei distance (Nei 1972) as the reference panel. For a detailed list containing error rate for all 4 different structures of reference panels, we refer to Supplementary Table S2.

To optimize UM imputation quality, one has to find a balance between representing as much of the genetic diversity of the dataset as possible while avoiding the introduction of noise by including genetically too distant individuals in the panel. Overall, the potential improvement of excluding distantly related individuals is relatively small compared to effects on the error rate when excluding highly related individuals. Especially when computational time is an issue or there is a subset of individuals in the reference panel that is close to a full representation of the genetic diversity of the dataset, it is still possible to improve UM imputation quality and computational feasibility by excluding less related individuals.

Computation time

Computation time in BEAGLE scales linear in the number of markers and slightly less than quadratic in the number of haplotypes in the sample Browning *et al.* (2018). In each iteration of the algorithm a similar computation time is needed, leading to a linear increase in the number of iterations for each sub-step. BEAGLE 5.0 needs far less computation time than all previous versions (Figure 11). E.g. running time for inference of chromosome 1 containing 501 DH-lines with 64'080 biallelic markers using 4 cores (Intel E5-2670 v2 2.5 GHz) needed 63 minutes on BEAGLE 4.1

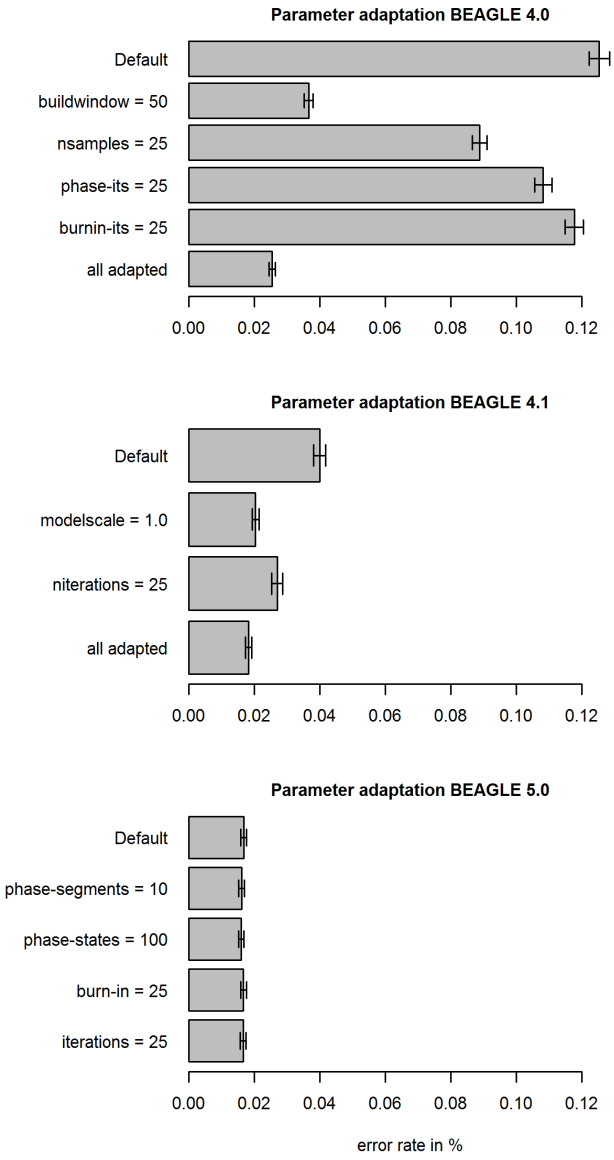

Figure 5 Error rates under different parameter settings and versions for Pseudo S_0 based on the maize data.

default whereas BEAGLE 4.0/5.0 only needed 13.9/4.6 minutes respectively. Older versions of BEAGLE run slightly faster when reducing buildwindow (BEAGLE 4.0: 11.5 minutes) and modelscale (BEAGLE 4.1: 35.8 minutes) but still do not compare favorably to BEAGLE 5.0. Depending on the imputation problem, one should consider modifying the number of iterations used in the algorithm. Especially when computation time is not an issue, we recommend increasing the number of iterations (BEAGLE 4.0: burnin-its, phase-its, impute-its; BEAGLE 4.1: niterations; BEAGLE 5.0: iterations, burnin) since basically all iterations just use the previous step as a starting value and try to improve that solution without much of a downside in our tests. As the main benefit of additional iterations is a more accurate phase, the number of iterations can be reduced when working with DH-lines.

When performing UM imputation on the maize dataset with a study sample of 50 and a reference panel of 350 computation times in BEAGLE 5.0 (24 seconds) were significantly lower than BEAGLE 4.1 (43 seconds) and BEAGLE 4.0 (108.6 seconds). When increasing the size of the reference panel the gains are even higher. For the chicken diversity panel with a study panel of 100 and reference panel of 1'710 BEAGLE 5 (1.45 minutes) was over ten times faster than the respective default settings of BEAGLE 4.1 (21 minutes) or BEAGLE 4.0 (64 minutes) for chromosome 1. The gains in computation time should only be increasing when further increasing the size of the reference panel (Browning *et al.* 2018). Additionally, the needed memory in BEAGLE 5.0 is massively reduced, especially when using binary reference (bref) format (Browning and Browning 2016), and thus enabling the use of BEAGLE for routine application in large size cattle breeding programs.

Significance of improvement

When comparing error rates under different settings one has to keep in mind the relevance of that optimization. A difference in error rates of 1% in a dataset containing 1% missing genotypes will only result in an improved overall data quality of 0.01% and thus might be negligible compared to other error sources like calling errors (Unterseer *et al.* 2014). If those improvements would mainly occur in the markers of interest (e.g. markers with low minor allele frequency) or the overall share of missing positions is high (as in UM imputation), this improvement could still be significant for later steps of the analysis.

It should be noted that positions set to NA in this study are chosen at random whereas in a real dataset there might be causal reasons like deletions, leading to some markers with much higher missing rates. When performing imputation on the actual NAs, we observed a higher variance in the imputed allele under different random seeds. As all considered methods always input one of the two allelic variants, this is ignored here but it should be noted that actual error rates are probably a bit higher than reported in this study.

Conclusion

Overall we can conclude that the quality for inference, UM imputation and phasing in BEAGLE 5.0 was usually at least as good has previous versions and less tuning of parameters is necessary to obtain good performance for livestock and crop datasets. Even in BEAGLE 5.0 an adaptation of parameters is especially necessary for the effective population size (ne) when performing UM imputation and working with genetic dataset with less diversity than a human outbred population. Especially when no parameter tuning in BEAGLE 4.0/4.1 was done, one should consider re-running previous preprocessing and quality control protocols.

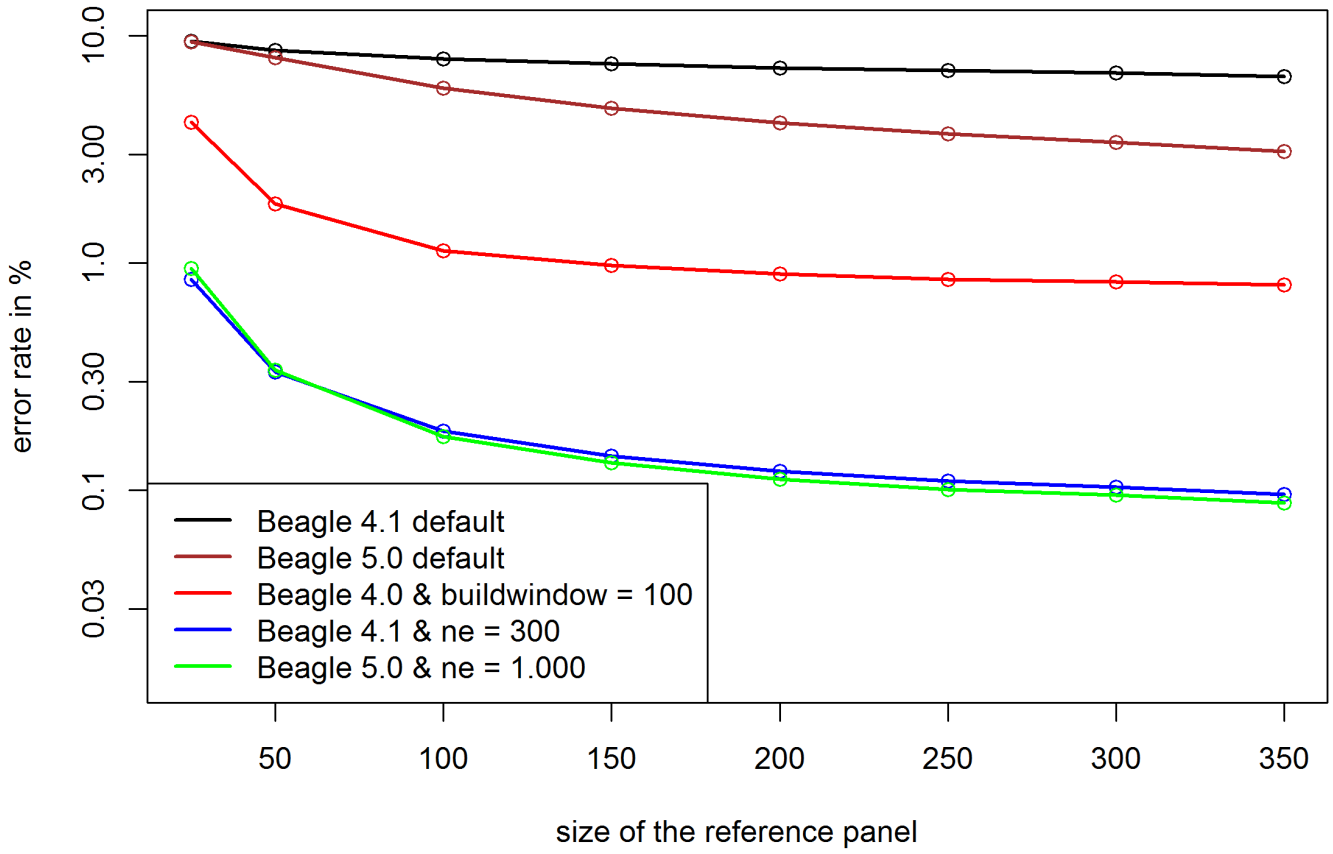

Figure 6 Error rates for UM imputation depending on the size of the reference panel in the maize data. Y-axis is log-scaled.

When considering increasing the marker density for later analysis like a genome-wide association study, one has to weight the potential gain of information of a larger marker panel against potential false positive results caused by imputation errors.

Improvements for inference and phasing quality are relatively small, when comparing BEAGLE 5.0 to previous versions with tuned parameter settings. In case default setting in BEAGLE 4.0/4.1 were used, error rates can differ quite substantially. When working with inbreds (like DH-lines) or default parameter settings imputation quality in older versions was significantly worse in all tests. Additional benefits of the use of BEAGLE 5.0 are massively reduced computation times and memory requirements. This is especially true for UM imputation when processing large reference panels and can enable the usage of BEAGLE 5.0 for datasets with a high number individuals even though increase in computation time is still close to quadratic in the number of individuals in the study sample. In case computation time is of no concern we additionally recommend an increase of the number of iterations (BEAGLE 5: burnin & iterations).

The used reference genome only mildly affected overall error rates in maize. Main benefit of the usage of the genetically more related flint reference genomes was a lower number of markers with extremely high error rates, whereas overall error rates were similar. With an increasing number of new reference genomes we recommend the use of a reference genome of similar genetic origin.

In terms of the design of an ideal reference panel we conclude that UM imputation without any individuals from similar genetic origin (in our case the same subpopulation) will lead to extremely high error rates and should only be done with caution. The needed size of the reference panel is highly dependent on the genetic diversity of the dataset. Without further information on genetic origin and sufficient computational power, we recommend to use a large reference panel since error rates are usually only mildly increasing, indicating that the algorithm underlying BEAGLE is quite good at filtering out irrelevant information. In case most of the genetic diversity of the study sample can be represented in a subset of the individuals in reference panel (e.g. a reference panel containing all founder individuals), significant improvements to UM imputation performance can be made by excluding genetically distant individuals.

ACKNOWLEDGEMENTS

The authors thank the German Federal Ministry of Education and Research (BMBF) for the funding of our project (MAZE - “Accessing the genomic and functional diversity of maize to improve quantitative traits”; Funding ID: 031B0195). The "Synbreed - Synergistic Plant and Animal Breeding" project was funded by the German Federal Ministry of Education and Research (FKZ 0315528E).

We also thank Brian Browning for providing quick and thoughtful replies to all our questions regarding insides on BEAGLE.

LITERATURE CITED

Albrechtsen, A., F. C. Nielsen, and R. Nielsen, 2010 Ascertainment biases in snp chips affect measures of population divergence. Molecular biology and evolution **27**: 2534–2547.

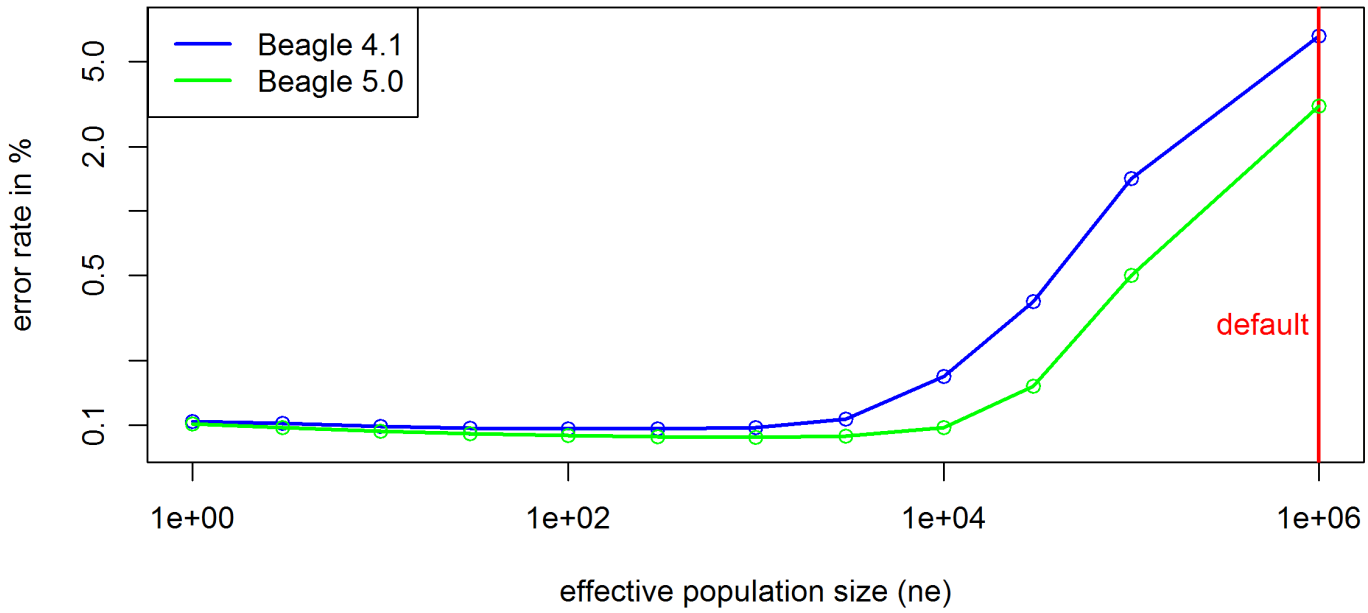

Figure 7 Error rates for UM imputation depending on the parameter ne in the maize data. Y-axis is log-scaled.

Baum, L. E. and T. Petrie, 1966 Statistical inference for probabilistic functions of finite state markov chains. The Annals of Mathematical Statistics **37**: 1554–1563.

Bellott, D. W., H. Skaletsky, T. Pyntikova, E. R. Mardis, T. Graves, *et al.*, 2010 Convergent evolution of chicken z and human x chromosomes by expansion and gene acquisition. Nature **466**: 612.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, *et al.*, 2007 Tassel: Software for association mapping of complex traits in diverse samples. Bioinformatics **23**: 2633–2635.

Browning, B. L. and S. R. Browning, 2007 Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genetic Epidemiology **31**: 365–375.

Browning, B. L. and S. R. Browning, 2013a Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194**: 459–471.

Browning, B. L. and S. R. Browning, 2013b Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194**: 459–471.

Browning, B. L. and S. R. Browning, 2016 Genotype imputation with millions of reference samples. The American Journal of Human Genetics **98**: 116–126.

Browning, B. L., Y. Zhou, and S. R. Browning, 2018 A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics **103**: 338–348.

Browning, S. R., 2006 Multilocus association mapping using variable-length markov chains. The American Journal of Human Genetics **78**: 903–913.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, *et al.*, 2016 Next-generation genotype imputation service and methods. Nature Genetics **48**: 1284.

Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, *et al.*, 2011 A large maize (zea mays l.) snp genotyping array: development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. PLOS ONE **6**: e28334.

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLOS Genetics **5**: e1000529.

International Chicken Genome Sequencing Consortium, 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**: 695.

Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, *et al.*, 2017 Improved maize reference genome with single-molecule technologies. Nature **546**: 524.

Kranis, A., A. A. Gheyas, C. Boschiero, F. Turner, Le Yu, *et al.*, 2013 Development of a high density 600k snp genotyping array for chicken. BMC Genomics **14**: 59.

Li, N. and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**: 2213–2233.

Lin, S., D. J. Cutler, M. E. Zwick, and A. Chakravarti, 2002 Haplotype inference in random population samples. The American Journal of Human Genetics **71**: 1129–1137.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong, *et al.*, 2015 Linkimpute: fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genetics **5**: 2383–2390.

Nadaraya, E. A., 1964 On estimating regression. Theory of Probability & Its Applications **9**: 141–142.

Nei, M., 1972 Genetic distance between populations. The American Naturalist **106**: 283–292.

Pirani, A., H. Gao, L. Bellon, and T. A. Webster, 2013 Best practices for genotyping analysis of plant and animal genomes with affymetrix® axiom® arrays: 2013:p0997.

Rabiner, L. R., 1989 A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the

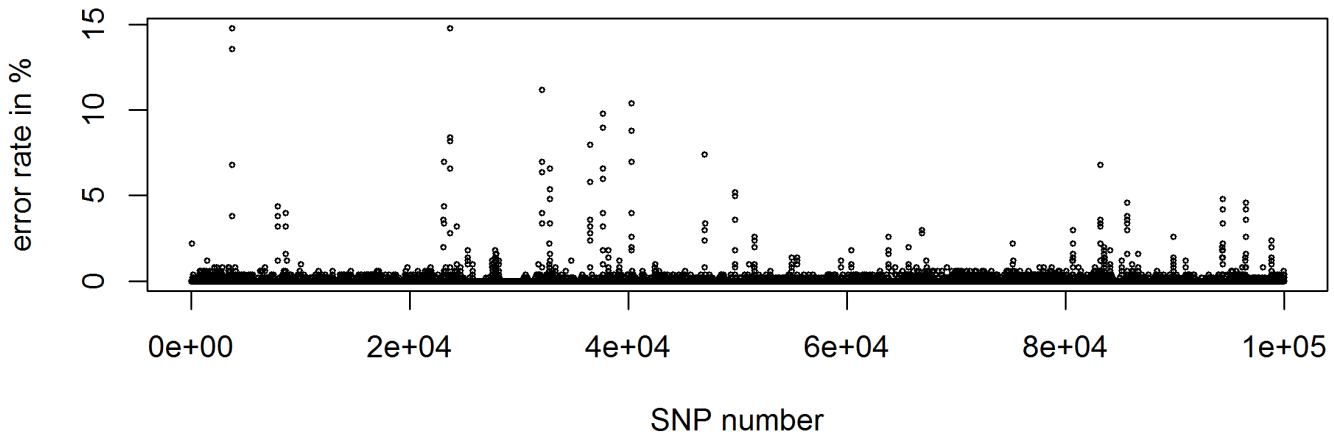

Figure 8 Error rate per marker for the first 100′000 SNPs according to physical position (starting with chromosome 1) using BEAGLE 5.0 default with B73v4 (Jiao *et al.* 2017) as a reference genome.

IEEE **77**: 257–286.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel, 2014 A new approach for efficient genotype imputation using information from relatives. BMC Genomics **15**: 478.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The b73 maize genome: complexity, diversity, and dynamics. Science **326**: 1112–1115.

Swarts, K., H. Li, J. A. Romero Navarro, D. An, M. C. Romay, *et al.*, 2014 Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. The Plant Genome **7**.

Unterseer, S., E. Bauer, G. Haberer, M. Seidel, C. Knaak, *et al.*, 2014 A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k snp genotyping array. BMC Genomics **15**: 823.

Unterseer, S., S. D. Pophaly, R. Peis, P. Westermeier, M. Mayer, *et al.*, 2016 A comprehensive study of the genomic differentiation between temperate dent and flint maize. Genome Biology **17**: 137.

Unterseer, S., M. A. Seidel, E. Bauer, G. Haberer, F. Hochholdinger, *et al.*, 2017 European flint reference sequences complement the maize pan-genome. bioRxiv p. 103747.

Weigend, S., U. Janßen-Tapken, M. Erbe, U. Ober, A. Weigend, *et al.*, 2014 Biodiversität beim huhn–potenziale für die praxis. Züchtungskunde **86**: 25–41.

Yan, G., R. Qiao, F. Zhang, W. Xin, S. Xiao, *et al.*, 2017 Imputation-based whole-genome sequence association study rediscovered the missing qtl for lumbar number in sutai pigs. Scientific Reports **7**: 615.

Zhang, P., X. Zhan, N. A. Rosenberg, and S. Zöllner, 2013 Genotype imputation reference panel selection using maximal phylogenetic diversity. Genetics pp. 319–330.

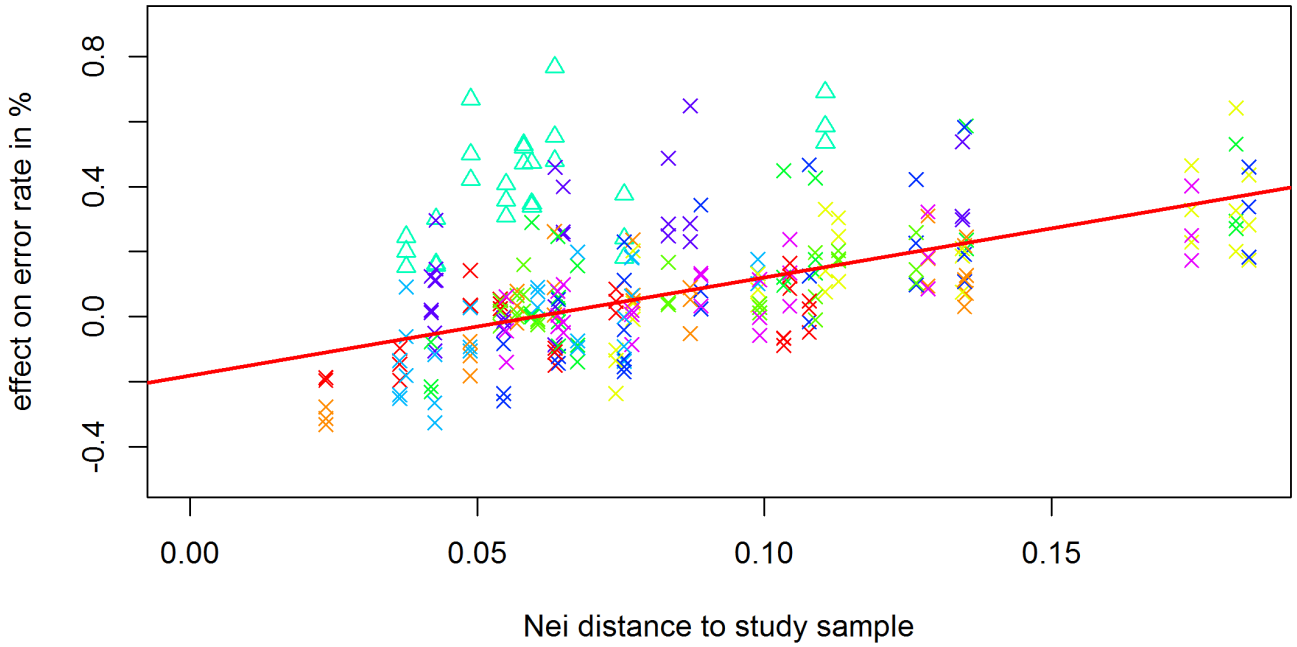

Figure 9 Effect of the inclusion of a single subpopulation in the reference panel based on their genetic distance to the dataset for the chicken diversity panel. Colors according to the subpopulation used as the real dataset in Supplementary Figure S7. Subpopulation 6 (including wild types - turquoise △) is ignored in the regression.

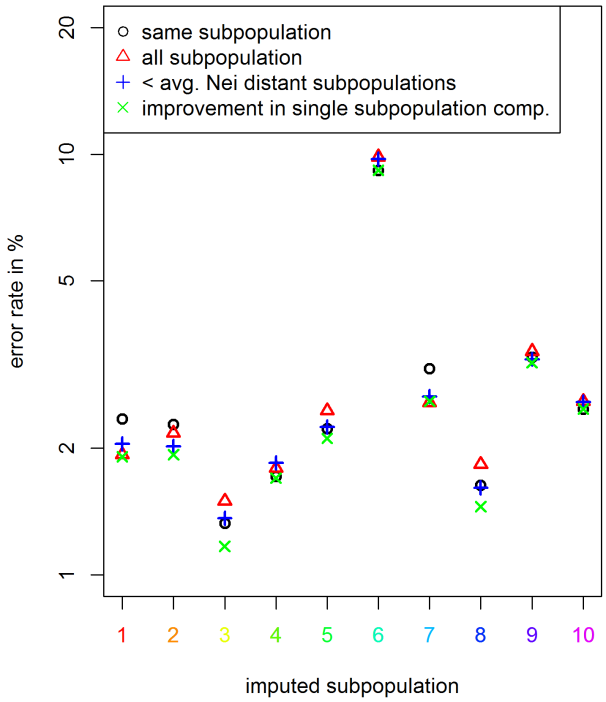

Figure 10 Comparison of error rates of UM imputation for different reference panels for the different subpopulations in the chicken diversity panel. Y-axis is log-scaled.

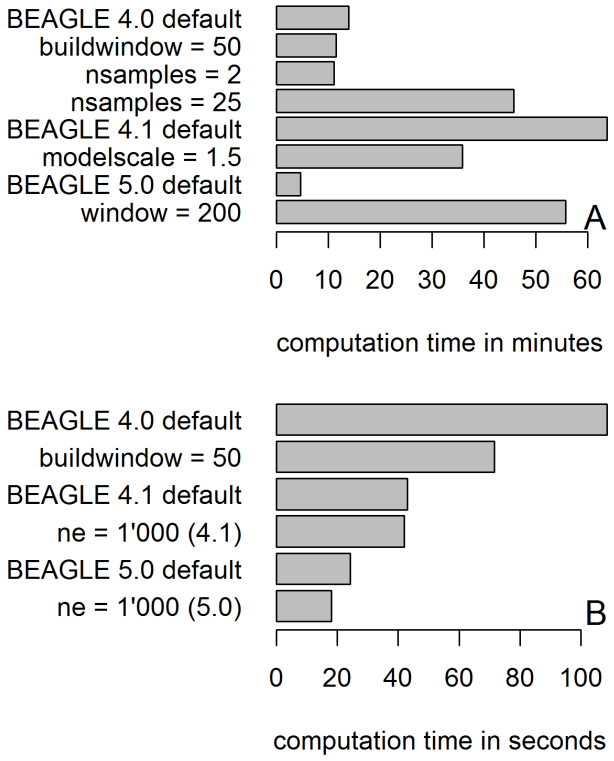

Figure 11 Computation time needed for performing inference (A) and UM imputation (B) for 64'080 biallelic markers in the maize data. For inference 501 DH-lines were used as the study sample. For UM imputation 50/350 DH-lines were used for study/reference sample.