Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants

Kelly Swarts, Huihui Li, J. Alberto Romero Navarro, Dong An, Maria Cinta Romay, Sarah Hearne, Charlotte Acharya, Jeffrey C. Glaubitz, Sharon Mitchell, Robert J. Elshire, Edward S. Buckler, and Peter J. Bradbury*

Abstract

Next-generation sequencing technology such as genotyping-by-sequencing (GBS) made low-cost, but often low-coverage, whole-genome sequencing widely available. Extensive inbreeding in crop plants provides an untapped, high quality source of phased haplotypes for imputing missing genotypes. We introduce Full-Sib Family Haplotype Imputation (FSFHap), optimized for full-sib populations, and a generalized method, Fast Inbred Line Library ImputatioN (FILLIN), to rapidly and accurately impute missing genotypes in GBS-type data with ordered markers. FSFHap and FILLIN impute missing genotypes with high accuracy in GBS-genotyped maize (*Zea mays* L.) inbred lines and breeding populations, while Beagle v. 4 is still preferable for diverse heterozygous populations. FILLIN and FSFHap are implemented in TASSEL 5.0.

THE NUMBER of genotyped individuals available to researchers has vastly increased in recent years due to the advent of low-cost, genome-wide genotyping platforms, such as GBS (Elshire et al., 2011). Genotyping-by-sequencing provides a reduced representation of the genome by targeting sequences adjacent to restriction enzyme cut sites, enabling parity in read location across samples. By adding barcoded adaptor sequences to the restriction-digested DNA, up to 384 samples can be multiplexed in one flowcell lane. However, the resulting GBS data may have high rates of missingness and heterozygote undercalling, depending on genome size, genome structure, and the number of samples combined. To effectively use GBS sequence data while maintaining low costs, we need a mechanism to impute these missing genotypes.

K. Swarts, Dep. of Plant Breeding and Genetics, 175 Biotechnology Bldg., Cornell Univ., Ithaca, NY 14853; H. Li, Institute of Crop Science and CIMMYT-China Office, Chinese Academy of Agricultural Sciences, No. 12 Zhongguancun South St., Beijing 100081, China; J.A. Romero Navarro, Dep. of Plant Breeding and Genetics and CIMMYT, 175 Biotechnology Bldg., Cornell Univ., Ithaca, NY 14853; D. An, China Agricultural Univ., College of Information and Electrical Engineering, No. 17 Tinghua East Rd., Beijing, China 100083; M.C. Romay, C. Acharya, J. C. Glaubitz, and S. Mitchell, Inst. for Genomic Diversity, 175 Biotechnology Bldg., Cornell Univ., Ithaca, NY 14853; S. Hearne, CIMMYT, Km. 45 Carretera Mexico-Veracruz, El Batán, Texcoco, Edo. de Mexico, CP 56237, Mexico; R.J. Elshire, AgResearch Limited, Grasslands Research Centre, Tennent Dr., Palmerston North, New Zealand 4442; E.S. Buckler, USDA-ARS, Institute for Genomic Diversity, and Dep. of Plant Breeding and Genetics, 159 Biotechnology Bldg., Cornell Univ., Ithaca, NY 14853; P.J. Bradbury, USDA-ARS, 409 Bradfield Hall, Cornell Univ., Ithaca, NY 14853.
Received 23 May 2014. *Corresponding author (pjb39@cornell.edu).

Abbreviations: FILLIN, Fast Inbred Line Library ImputatioN; FSFHap, Full-Sib Family Haplotype Imputation; GBS, genotyping-by-sequencing; HMM, Hidden Markov Model; IBD, identical by descent; LD, linkage disequilibrium; MAF, minor allele frequency; NAM, nested association mapping; RIL, recombinant inbred line, SNP, single nucleotide polymorphism.

Published in The Plant Genome 7
doi: 10.3835/plantgenome2014.05.0023

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

Missing data is often a function of genome size and degree of multiplexing, where some sequences are simply not sampled when the genome size is large or many samples are combined in a flowcell. Heterozygote undercalling is also a function of low-coverage sampling; to call a heterozygote for a given genotype, that genotype must be covered by at least two reads, and those reads must be from different sister chromatids. In the case of maize, based on sampling from a Poisson distribution with a lambda value of 0.6 (the average coverage for maize GBS data), we expect only 12% of the genome to be sampled two or more times, providing an upper limit for correct heterozygous single nucleotide polymorphism (SNP) calls.

Missing data can also reflect true biologically missing sequence due to small insertions or deletions or larger structural variants in the genome. Because these missing data provide a real biological signal, it is desirable to capture this type of missing data in imputation. In maize, for example, not only is allelic diversity high (Vigouroux et al., 2008), but 70% of genes and 90% of the genome shows structural variation in a panel of only 103 diverse inbred maize and teosinte [*Z. mays* subsp. mexicana (Schrad.) H. H. Iltis] lines (Chia et al., 2012). Maize is not unique in this respect; many agronomically important crop plants show a similar pattern of high structural variation and allelic diversity (Bhullar et al., 2010; Hurwitz et al., 2010; Schnable et al., 2011; Ramu et al., 2013; Das et al., 2013).

Accurate imputation benefits downstream applications such as genome-wide association and linkage mapping studies by accurately identifying rare variants, which, in turn, increases the power to detect statistical associations (Spencer et al., 2009; Cleveland et al., 2011). Imputation is expected to provide the greatest benefit for mapping studies when linkage disequilibrium (LD) between markers is low. This is often the case in natural populations or primarily outcrossing species, since fewer markers are present on each haplotype to tag a statistical genotype-phenotype association (Spencer et al., 2009). Even crop species, which often have extended LD as a result of breeding efforts, only share short haplotypes when comparing distantly related individuals. Accurate imputation is thus critical for effectively using the output of low-coverage, low-cost genotyping platforms such as GBS.

While accurate imputation increases the value of low-cost, low-coverage genotyping, much of the available software for imputation has been tailored for humans (Howie et al., 2009; Browning and Yu, 2009; Liu et al., 2013). Humans, in contrast to many economically valuable plant species, are highly heterozygous, obligate outcrossers with no controlled mating designs, little inbreeding, and much less structural variation than that observed in crop plants (Ross-Ibarra et al., 2009; Buckler et al., 2009; Chia et al., 2012). Because of this, the publically available algorithms designed for humans are not optimized to accurately impute or leverage unique information from crop systems. That we can reasonably assume phase for inbred lines and inbred segments from crop plants suggests a different model for genotype imputation is needed. While there are crop specific algorithms that have been developed for unordered markers (Rutkoski et al., 2013), for known pedigrees (Meuwissen and Goddard, 2010), and in the context of genomic prediction (Daetwyler et al., 2011; Hickey et al., 2012), most of these are not publically available or do not output imputed genotypes.

A number of agronomically important systems use inbred lines extensively (e.g., maize, rice, wheat, soybean, barley, sorghum), and many of these have structured population resources for mapping traits of interest (McMullen et al., 2009; Diers et al., 2011; Mace et al., 2013). Very accurately mapping the recombination break points in these populations is desirable for fine mapping studies and for appropriately assigning effect estimates to the proper parent in association studies. We present here FSFHap, a fast and accurate imputation algorithm for ordered genotypes, optimized for full-sib families. FSFHap follows methods published for detecting recombination breakpoints in *Drosophila* (Andolfatto et al., 2011; King et al., 2012).

We also present FILLIN, a fast and accurate generalized imputation strategy built on the FSFHap algorithm that leverages inbred segments from large but sparse genotypic datasets to identify parental haplotypes and impute missing genotypes in systems with previously ordered markers. Both FILLIN and FSFHap allow for missing data in the imputed genotypes, capturing structural variation hidden in sparse genotypic data. We find from GBS data that many breeding programs have some small level of contamination and pedigrees are not always consistent with information on relatedness from genotypes. Because FILLIN and FSFHap do not require known parental genotypes, these algorithms provide pedigree independent imputation. Because populations in crop plants are often large, computational time for FILLIN is reduced by using bit level operations and multithreading, and is designed to scale linearly to enable fast breeding decisions. Both the FSFHap and FILLIN algorithms are implemented in TASSEL 5.0 (Bradbury et al., 2007).

We compare the speed and accuracy of our novel algorithms on inbred and outbred maize (*Zea mays*) genotyped using GBS (Elshire et al., 2011) with Beagle v. 4 (Browning and Browning, 2013). Beagle v. 4 was chosen because in early tests we found it to be the most powerful and comparable algorithm available; it does not require an external haplotype library, it accepts high levels of missing data, and it has the computational speed to impute whole genomes or chromosomes. We also compare accuracy between FSFHap, FILLIN, and Beagle v. 4 in a full-sib family recombinant inbred line (RIL) population.

Materials and Methods

Algorithms

Viterbi Algorithm

Both FSFHap and FILLIN rely on a Hidden Markov Model (HMM) to detect recombination break points between haplotypes. The HMMs define genotype as the true, unobserved genotype and the SNP or sequence

variant calls made by the sequencing pipeline as the observations. Using this formulation, the problem of imputation can be restated as the problem of determining the unobserved genotype that best explains the observed data. If for a given sample, **y** is a vector of the observed SNP calls; **g** is a vector of unknown, unobserved genotypes; and **M** is a probability model describing the data and the genotypes. Then, one approach is to seek to maximize the likelihood (L):

$$L(\mathbf{y}|\mathbf{g}, \mathbf{M}).$$

In general, maximizing L for genotypes of n sites requires evaluating this likelihood for each of 2^n possible genotypes, an impossible task for 100 markers, let alone for the hundreds of thousands of markers in the data described here. Fortunately, because nucleotides are arranged sequentially on a chromosome, the problem can be modeled as a Markov chain. Doing so allows the use of the Viterbi algorithm (Rabiner, 1989) to identify the genotype, **g**, that maximizes the likelihood of the observed data. The Viterbi algorithm only needs to evaluate a small, tractable subset of the potential genotypes to maximize the likelihood.

Applying the Viterbi algorithm requires defining two separate probability matrices. Taken together, these two probability matrices determine **M**, the probability model. The first is a transition probability matrix, which describes the probability of each possible genotype at a site given the genotype at the previous site for all possible genotypes at the previous site. The second is an emission probability matrix, which describes the probability of observing each possible allele call, given each possible genotypic state. This probability matrix has to capture both the probability of a genotyping error and the probability that only one of the two possible alleles was observed at a heterozygous site, which results in that site being incorrectly scored as homozygous. In a classic hidden Markov chain, both probability matrices are constant across all sites. In our application, we treat the emission probabilities as constant, but allow the transition probability to vary depending on distance between sites and location in the genome.

Both the transition and emission matrices are estimated from the data set being imputed, which is known to contain errors. An expectation-maximization method is used to improve that estimate. Because the imputed data provides a better indication of the actual genotypes than the original data, after the initial imputation, the imputed states can be used to make new estimates of the probability matrices. This process is repeated to convergence. Estimating the probability matrices is the expectation step. Applying the Viterbi algorithm constitutes the maximization step.

Initializing the Matrices

Many of the DNA samples in our data (as is common with other crop studies) were created by bulking DNA from several plants. Most were presumed to be homozygous but often had residual heterozygosity or heterogeneity. Typically, the bulked plants were progeny of a single self-pollinated plant. In that case, the progeny represented a random sample of 2n gametes from the parent, where n is the number of progeny bulked. As a result, for a single bulked sample, the minor allele frequency (MAF) at a heterozygous site, instead of being 0.5 as it would have been for a DNA sample of a single parent plant, ranged from 0 to 0.5 with probabilities equal to 2n draws from a binomial distribution with $p = 0.5$. Within a sample, the allele frequencies at adjacent segregating sites will be expected to be the same, since they all represent the same sample. To accommodate chromosome segments with different allele frequencies, we allow for five genotype states, representing homozygous A, 3A:1B, 1A:1B, 1A:3B, and homozygous B. The initial emission probabilities were set as shown in Table 1.

Table 1. The Viterbi algorithm initial emission probability matrix, P(Allele Call|State). Note that all of the rows sum to 1.

	Allele Call		
State	A	H	B
AA	0.998	0.001	0.001
3A:1B	0.6	0.2	0.2
1A:1B	0.4	0.2	0.4
1A:3B	0.2	0.2	0.6
BB	0.001	0.001	0.998

Table 2. The Viterbi algorithm transition probability matrix for Fast Inbred Line Library ImputatioN (FILLIN).

State	AA	3A:1B	1A:1B	1A:3B	BB
AA	0.999	0.0001	0.0003	0.0001	0.0005
3A:1B	0.0002	0.999	0.00005	0.00005	0.0002
1A:1B	0.0002	0.00005	0.999	0.00005	0.0002
1A:3B	0.0002	0.00005	0.00005	0.999	0.0002
BB	0.0005	0.0001	0.0003	0.0001	0.999

The transition probabilities between states at adjacent sites are calculated differently for the two algorithms we present here. For FILLIN, the transition matrix is fixed (Table 2). For FSFHap, the transition matrix is dependent on the expected rate of recombination and the distance between the sites. The transition probability between the different states was estimated using all intervals between nonmissing markers, then adjusted based on the ratio of the actual interval to the average interval length. An initial estimate was based on expected recombination rates. Convergence of the expectation-maximization algorithm is not dependent on the initial estimate as long as it is reasonable.

Full-Sib Family Haplotype Imputation

As with most imputation methods, FSFHap begins by identifying haplotypes. In the case of a biparental population when the objective is to find recombination

breakpoints, the interest is in determining which chromosome segments are identical by descent (IBD) from which parent. The algorithm thus attempts to identify exactly two parental haplotypes, ignoring any sites that happen to be heterozygous in either of the parents. The algorithm as written has not been tested for families derived from a cross between outbred, heterozygous parents but handles segments that are heterozygous in one parent by using only the homozygous sites in those segments.

One assumption of the Viterbi algorithm is that the probability of an error at one site is independent of other sites, which is not necessarily the case for sites in the same GBS sequence read. Consequently, before identifying parental haplotypes, if any SNP pair came from the same tag, one of the pair is deleted from the dataset. Next, within each biparental family, the algorithm clusters lines in a window of 50 variant sites at the beginning of a chromosome using a custom clustering method (described below). Large clusters identify parental haplotypes, while small clusters are generally heterozygous or contain individuals with genotyping errors. If the first window tested has more than two large clusters, the next adjacent window is checked until a window with only two large haplotype clusters is found. Once a window is found meeting that criterion, it serves as an anchor for determining the next haplotype block. Starting immediately after the anchor window, the allele calls for subsequent sites are evaluated one at a time. Labeling the two anchor haplotypes A and B, if a site's allele calls for RILs in haplotype A are mostly the same and the allele calls for the RILs in haplotype B are mostly different from the haplotype A majority allele, the site is assigned to the correct haplotype. Otherwise the site is removed from the dataset. The majority allele within the respective haplotypes is recorded and the next adjacent site evaluated. Once alleles have been assigned to A and B for 50 additional sites, these 50 sites become the new anchor window. Two clusters are formed of lines that are within a minimum distance of each of the two new haplotype sequences. The entire process is repeated to extend each haplotype another 50 sites until the end of the chromosome is reached.

Once the parental haplotypes are identified, the progeny are scored as Parent A, Parent B, or heterozygous at each site. Then, for each of the progeny individually, the Viterbi algorithm is applied to the nonmissing sites to determine the most likely genotype given the observations. The missing sites are then imputed based on the flanking nonmissing markers. If the flanking markers match (both are A, B, or H), then the missing site is imputed to the same value as the flanking markers; otherwise it is left missing. Finally, the sites are converted back to nucleotides by examining the original nucleotide calls of all the individuals in the A and B classes at each site.

Custom Clustering Method

Because of data scarcity, standard hierarchical clustering methods perform inadequately for classifying haplotypes formed from a limited number of sites. Of these, the *complete* method, which calculates distance between clusters as the maximum pairwise distance, gives the most useful results. The algorithm described here modifies that method by defining distance between two haplotypes as the sum of the differences across sites, where a site difference is 2 for different homozygotes, 1 if one site is homozygous and the other heterozygous, or 0 if either haplotype had a missing value. Further, each cluster contains all of the individuals less than a given distance from all the other individuals in the cluster. Because of missing data, this means some individuals could belong to more than one cluster. Each cluster is interpreted as all the individuals that could have the same genotype for the entire window. Because heterozygous sites are often called as one of the possible homozygotes at random, heterozygous individuals do not form large clusters. A useful measure of cluster size is the sum of 1/(the number of clusters to which an individual belonged) over the individuals in a cluster.

Fast, Inbred Line Library ImputatioN

FILLIN is optimized to leverage inbred segments for fast imputation in very large, sparse datasets. Like other algorithms, we separate haplotype generation and imputation (Howie et al., 2009; Liu et al., 2013). FILLIN first generates high coverage haplotypes from inbred lines and inbred segments by dividing the genome into nonoverlapping windows. Within each window, the Hamming distance is used to cluster sequences that share highly similar genotypes, and these clusters are then collapsed to generate higher-coverage haplotypes. To calculate distance, sites with a missing genotype call in either taxon are ignored and the distance between a heterozygous and homozygous genotype is considered to be half the distance of one homozygous genotype to the alternate homozygote. To best represent high levels of structural variation, we do not require complete coverage for the resulting haplotypes. Additionally, a small amount of residual heterozygosity propagates to the resulting haplotype donor files, as the algorithm makes no effort to phase residual heterozygous genotypes. This approach results in very fast haplotype generation (Fig. 1), but is less sensitive than other algorithms if the samples are highly heterozygous, since we make no effort to phase (Gusev et al., 2009; Browning and Browning, 2011, 2013). The haplotype generation step should always be performed with all of the samples available, as (i) small-scale haplotype windows may be replicated across even genetically distant individuals, and (ii) the algorithm requires at least two samples to generate a haplotype by default.

To impute these higher coverage haplotypes back to the target samples, FILLIN takes an iterative approach to imputation. First it selects possible donors based on shared minor alleles within each window. Shared minor alleles are particularly informative, since most of the minor allele states derive from more recent mutation and when two taxa share these alleles it suggests recent common ancestry. FILLIN then ranks haplotype donors by genetic distance to the taxon being imputed (again, looking only

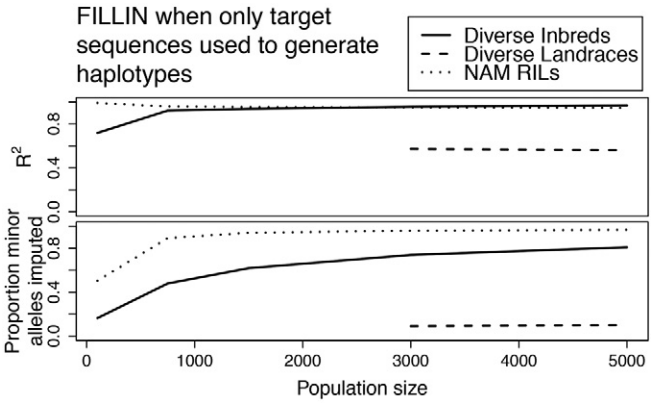

Figure 1. Fast Inbred Line Library ImputatioN (FILLIN) when only target genotypes for imputation used to generate donors, using the nested association mapping (NAM) recombinant inbred lines (RILs), diverse inbred lines, and heterozygous, diverse landraces. Within each population, each data set is a random subsample of the larger data set; only chromosome 10 was imputed.

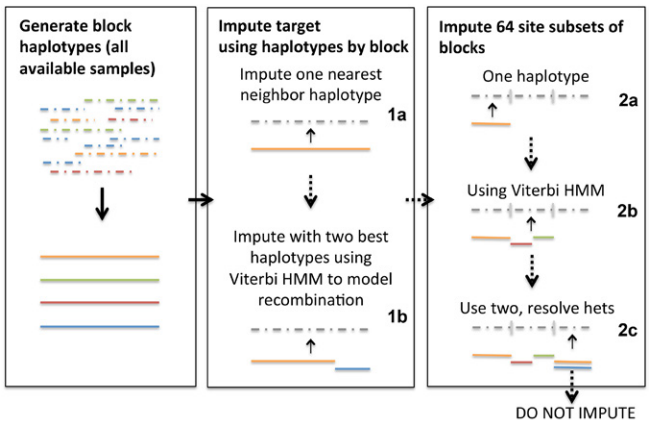

Figure 2. Fast Inbred Line Library ImputatioN (FILLIN) algorithm overview. FILLIN first tries to impute the entire site window with one (1a) or two (1b) haplotypes (using the Viterbi Hidden Markov Model [HMM] to model the recombination break point), then if that is unsuccessful tries to impute for smaller windows, first with one haplotype (2a), then two with Viterbi (2b), finally by combining two haplotypes to model heterozygosity (2c). If this does not satisfy (lower) error thresholds, the smaller window is not imputed. Refer to methods for detailed description of 1a–2c. Dashed arrows mean that the algorithm continues if conditions are not satisfied for imputation.

within the current window); if the distance falls below a user-specified threshold, it then imputes one haplotype to the entire window (1a in Fig. 2). If this fails, and the taxon is modeled as inbred based on global heterozygosity, the algorithm looks for two donors that can together adequately explain the minor alleles in the entire window (1b in Fig. 2). This assumes and models for a recombination break point between two known haplotypes, and it uses the above Viterbi Hidden Markov algorithm to decide where to switch. The Viterbi is run in both directions, with disagreements defaulting to the genotype with the longest path length (i.e., highest likelihood).

If one or two donors cannot be found to explain the entire window, the algorithm repeats this process for smaller, 64-site windows within the larger window. Each 64-site window serves as a focus, and the algorithm extends out right and left until this window (the "focus block") contains a minimum number of minor alleles to calculate Hamming distance. FILLIN then attempts to impute based on single haplotype (2a in Fig. 2) and the two-haplotype (2b in Fig. 2) Viterbi imputation, if distance between the donor and target falls below a threshold. If these attempts fail to explain sufficient minor alleles, the algorithm will then find two haplotypes that explain the minor alleles at a higher error threshold, combine these two haplotypes, and impute using this combined haplotype sequence, modeling the region as heterozygous (2c in Fig. 2). If *this* search fails, that 64-site window will not be imputed. Because low-coverage sequence data often results in undercalling heterozygotes, an option to resolve homozygotes predicted to be heterozygous is available for all imputation types except 2c.

The maximum genetic distance thresholds for the focus block are customizable by the user, but by default are set more stringently than those for the entire window since the focus blocks are shorter and are expected to contain fewer sequencing errors if the haplotype is truly IBD to the target. These thresholds are also different for outbred versus inbred taxa, since when two haplotypes explain the minor alleles of a target sequence in an outbred taxon, it is more probable that the target sequence is heterozygous rather than a segment containing a recombination between two inbred haplotypes. For a taxon that falls above a user-defined per taxon heterozygosity threshold (is outbred), the threshold for using Viterbi (2b) is set to 0. If a taxon is considered generally inbred, any discrepancy between the two combined haplotypes that generates a heterozygous genotype (2c), is set to missing.

Comparison with Existing Algorithms

FILLIN differs from other available algorithms, most of which have been designed for human-derived sequence data, primarily in its approach to haplotype generation, phasing, and inbreeding assumptions (Howie et al., 2009; Browning and Yu, 2009; Liu et al., 2013; Browning and Browning, 2013). Imputation algorithms either generate haplotypes de novo or rely on a densely genotyped reference panel, such as 1000 Genomes, which are not available for most species. The public algorithms that generate de novo haplotypes implicitly assume that the unimputed individuals have significant heterozygosity and must be phased (Browning and Yu, 2009; Browning and Browning, 2013). In the case of Beagle v. 4, this increases runtime exponentially by the number of samples (Fig. 1). However, if haplotypes really do only exist in the heterozygous state, Beagle's refined IBD (Browning and Browning, 2013) algorithm should find these segments better than FILLIN and thus make the extra computation worthwhile. In contrast, FILLIN allows for significant inbreeding in the target population and saves computational time by first checking for high similarity between one of the haplotypes and

the target sequence. FILLIN and Beagle v. 4 both differ from many other algorithms (Purcell et al., 2007; Howie et al., 2009; Liu et al., 2013) in that they can impute whole genomes or chromosomes in one run.

Test Datasets and Analysis Optimizations

Test Datasets

We tested the FILLIN and FSFHap algorithms against Beagle v. 4 (Browning and Browning, 2013), which we found in preliminary tests to be the most comparable available algorithm: it can generate haplotypes, tolerate high levels of missing data, and impute entire chromosomes with tens of thousands of markers in one run. We also compare against a naive imputation method, which imputes missing genotypes based solely on allele frequencies in the unimputed data. We compare results from three distinct maize datasets genotyped using maize GBS build v. 2.7 (Glaubitz et al., 2014): (i) 429 replicate samples representing 287 related temperate inbred Ex-PVP and Iowa breeding lines (temperate inbreds; Romay et al., 2013), (ii) 467 replicate samples from a panel of well-studied 282 diverse inbred lines from around the globe (Flint-Garcia et al., 2005; diverse inbreds), and (iii) 366 outbred (highly heterozygous) landraces where one half originate from the American Southwest, one quarter from the rest of the Americas (Heerwaarden et al., 2011), and one quarter from Spain (Revilla et al., 2003; diverse landraces). A fourth dataset, a RIL population of full-sib families from the maize nested association mapping (NAM) panel (McMullen et al., 2009), is used to compared FSFHap to FILLIN and Beagle v. 4. Each dataset is genome-wide and filtered so that only polymorphic sites with 10% minimum coverage and taxa with 10% sites present are retained.

FSFHap conducts additional filtering on the NAM RIL population before imputation. Because each of the individual NAM full-sib families was derived from three distinct F1 ears and because some parents had residual heterozygosity, any given site might be polymorphic in one subfamily and monomorphic in another. To deal with this after the parental haplotypes were identified for each family, the individual subpopulations were checked to make sure each site was more likely to be segregating 1:1 than to be monomorphic. Any site determined to be monomorphic in a subpopulation was set to missing within that subpopulation. At the same time, each site was checked to make sure it was in LD with all its neighbors within a 30-site window. Because of occasional contamination by foreign pollen during the inbreeding process, a few individual RILs carry substantial amounts of nonparental DNA. To find individuals containing significant amounts of nonparental DNA, after an initial imputation, individuals that were more than 30% heterozygous were removed from the data set and the families reimputed from the original data.

Haplotype Generation

FILLIN generates haplotypes using a GBS-derived dataset of 40,992 samples, one-eighth of which are outbred landrace accessions, and the rest are made up of diverse inbred maize lines, inbred teosinte, and biparental mapping and breeding populations. FILLIN generated haplotypes are publically available on panzea.org (verified 4 Sept. 2014) and will be updated periodically as new data becomes available. The populations tested in this manuscript are included in the 40k taxa dataset, increasing the probability that their haplotypes are included in the haplotype donor file. Beagle v. 4 has no mechanism to use FILLIN haplotypes as input due to residual missing data and could not internally generate haplotypes from such a large dataset. Thus, for all Beagle runs, Beagle uses haplotypes internally generated from the target samples only, but all available replicate samples for each inbred line were input into Beagle. We expect that FILLIN will typically use haplotypes generated from samples beyond just those targeted for imputation, but in Fig. 1 we provide results of FILLIN accuracies for inbred and outbred populations of various sizes imputed with haplotypes generated from the target samples only.

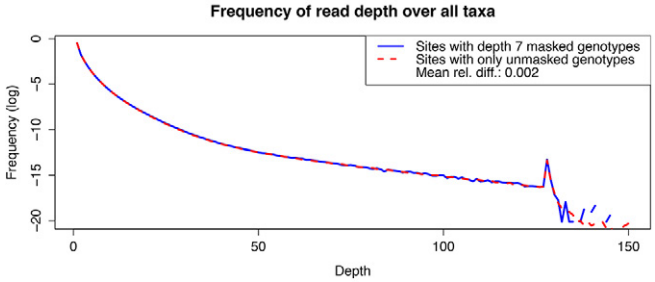

Figure 3. Sites that contain masked genotypes show the same read depth distribution as sites without masked genotypes, suggesting that masked genotypes are representative and acceptable for calculating accuracy.

Masking and Calculating Accuracy

To calculate accuracy, we masked a subset of known genotypes with high read-depth (exactly seven reads per site) and with a physical position divisible by seven. If GBS can be expected to sample either diploid chromosome equally, the probability that a heterozygous genotype with a read depth of seven is called as a homozygote for either the major or minor allele is $P(\text{AA}|\text{Het}) + P(\text{BB}|\text{Het}) = 0.5^7 + 0.5^7 = 0.0157$. Additionally, only heterozygote calls supported by at least two reads for both alleles were masked to exclude calls based on potential sequencing errors. Because we only masked a subset of genotypes with a read depth of seven, we can compare the distribution of read depths at sites where seven-read depth sites are masked versus sites where they are not. Figure 3 shows that the sites that contain masked genotypes have the same read depth distribution relative to sites without masked genotypes suggesting that sampling these genotypes for accuracy calculation is reasonable.

For all of the datasets tested, we chose to quantify accuracy using the coefficient of determination, R^2, versus a more simplistic measure, such as total percentage

Table 3. Coefficient of multiple determination (R^2), absolute proportion correct, and accuracies by known genotype, with the most accurate method for each class or population in italics. Note that the absolute proportion correct mirrors the accuracy for the major allele, while R^2 weights the minor alleles more heavily, which are more informative for downstream applications. For the diverse landraces and nested association mapping (NAM) recombinant inbred lines (RILs), Fast Inbred Line Library ImputatioN (FILLIN) is only more accurate for heterozygotes (Het) because it attempts many fewer imputations for that genotype class.

Test dataset	R^2	Absolute	Minor	Het	Major
Temperate inbreds					
Naive	0.046	0.642	0.116	0.348	0.748
Beagle v. 4	0.942	0.984	0.956	*0.452*	0.993
FILLIN	*0.986*	*0.996*	*0.993*	0.252	*0.999*
Diverse inbreds					
Naive	0.04	0.641	0.105	0.334	0.75
Beagle v. 4	0.883	0.97	0.905	*0.484*	0.986
FILLIN	*0.99*	*0.996*	*0.993*	0.322	*0.999*
Diverse landraces					
Naive	0.064	0.643	0.116	0.358	0.762
Beagle v. 4	*0.662*	*0.892*	*0.698*	0.656	*0.957*
FILLIN	0.583	0.85	0.57	*0.698*	0.905
NAM RILs					
FSFHap	*0.974*	0.99	0.968	0.846	*0.995*
FILLIN	0.971	*0.991*	*0.97*	*0.858*	*0.995*
Beagle v. 4	0.948	0.985	0.956	0.596	0.994

accuracy where known genotypes are coded as categorical variables. We did this because the great majority of genotypes masked are of the major allele, skewing the accuracy calculation towards imputation accuracies for this genotypic class (Table 3). Because minor alleles are of most interest in downstream applications and the harder allele to predict, we chose to use R^2 to better represent the capabilities of the different methods. To calculate R^2, we compared masked to imputed genotypes with the major allele coded as one, the minor as 0, and heterozygotes as 0.5. Unimputed genotypes are not reflected in the R^2 calculation except as a decrease in the number of genotypes compared.

Computational Time and Algorithm Parameters

FILLIN and Beagle v. 4 were run on two 6-core Intel Xeon E5 2620 with 2 GHz CPU, 4TB SATA HD, 1TB SSD HD, and 128GB RAM. Beagle v. 4 was run using the default parameters, with no external pedigree information or reference panel. FILLIN was run with a window size of 8000 sites. Haplotypes were required to have a minimum site presence of 0.6, and the maximum genetic divergence between samples to generate haplotypes was set to 0.01. For imputing haplotypes to the target sequences, 20 informative minor sites were required within a search window and up to 20 haplotype donor hypotheses were explored

Table 4. Raw datasets used for analysis.

Dataset	No. filtered taxa	No. segregating sites	Avg. proportion present	Avg. proportion heterozygous (± SE)
Temperate inbreds	429	443,036	0.431	0.003 ± 0.004
Diverse inbreds	467	545,154	0.462	0.003 ± 0.005
Diverse landraces	366	600,724	0.509	0.052 ± 0.017
NAM RILs†	4776	556,001	0.301	0.003 ± 0.002

† NAM, nested association mapping; RILs, recombinant inbred lines.

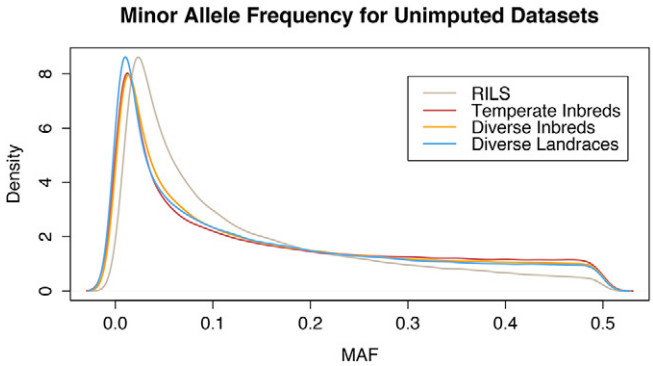

Figure 4. Minor allele frequency (MAF) densities for the unimputed datasets used for this study. For the recombinant inbred lines (RILs), the MAF was calculated in each biparental family separately, and combined. Populations sizes are as follows: temperate inbreds (429), diverse inbreds (367), diverse landraces (366), nested association mapping RILs (4667).

for a given window. The maximum genetic distance between the haplotype donor and target taxon to impute one haplotype for the entire sites window (1a in Fig. 2) was set to 0.01, and the maximum distance to impute two haplotypes was set to 0.003 (1b in Fig. 2). To impute donors to the smaller focus windows (64-site focus, but extended so that the focus window covers 20 informative sites) when the whole-block imputation thresholds were breached, the settings for imputing two haplotypes to inbred lines (with heterozygosity below 0.02) with Viterbi (2b in Fig. 2), one haplotype (2a in Fig. 2), or the combined hybrid haplotype (2c in Fig. 2) were set to 0.001, 0.003, and 0.01. For heterozygous genotypes, these thresholds were set to 0, 0.001, and 0.01. Genotypes were not imputed if these thresholds were not met. All of these thresholds values can be optimized for different needs, but the values above are the default values for FILLIN.

Results

The NAM RILs consist of 25 biparental families with around 200 F6 progeny each (Table 4). They have an average of 0.3X coverage per site, are polymorphic in at least one family at 556,000 sites across the genome, and are highly inbred. Average heterozygosity per line before imputation is approximately 0.001. While for each family we expect MAF of 0.5, across the whole population, MAFs are very low (Fig. 4). To test FILLIN versus Beagle, we test three maize datasets differing in degree

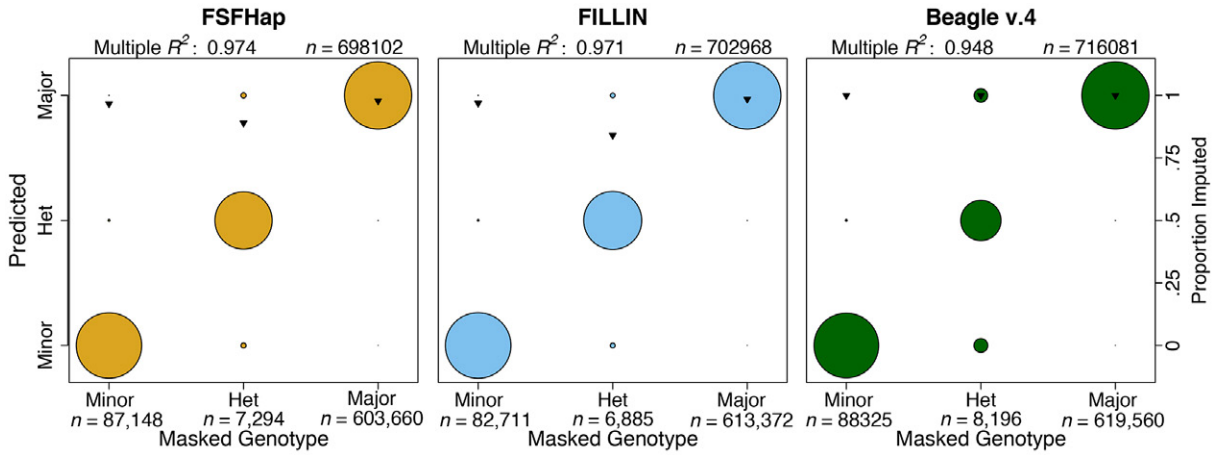

Figure 5. Accuracy comparison between Full-Sib Family Haplotype Imputation (FSFHap), Fast Inbred Line Library ImputatioN (FILLIN), and Beagle v. 4 for full-sib nested association mapping (NAM) recombinant inbred lines (RILs). The diameter of each circle represents the proportion within each known genotype class, the values imputed for each class are indicated at the bottom of the plot. Triangles mark proportion imputed by each known class; Beagle imputes 100% of missing genotypes. Note that FSFHap imputes more heterozygous sites than FILLIN.

of inbreeding and haplotype diversity. All have approximately 0.5X coverage per site and range in average heterozygosity from 0.001 to 0.029 (Table 4). The number of polymorphic sites across the genome ranges from 433,000 to 600,000. As the datasets become more diverse, there is an increased skew towards rare alleles (Fig. 4).

For the full-sib NAM RILs, FSFHap and FILLIN performed very similarly, but FSFHap imputed more heterozygous sites with increased accuracy. Both algorithms outperformed Beagle v. 4 (Fig. 5). As expected, all algorithms performed far better than the naive allele frequency imputation for all datasets tested with FILLIN and Beagle v. 4 (Fig. 6). FILLIN outperformed Beagle v. 4 for closely related and diverse inbred lines (Fig. 6A,B), but Beagle v. 4 outperforms FILLIN for heterozygous landraces (Fig. 6C), as well as the few residual heterozygous sites in inbred lines (Fig. 7).

For all of the inbred datasets (diverse, temperate, and RILs), FILLIN most often imputed the whole site window with one haplotype or two, using the Viterbi algorithm to model recombination breakpoints (Table 5). The temperate inbreds, although they are more closely related, use the focus block imputation more often than the diverse inbreds, and this may explain their slightly decreased accuracy and suggest more residual heterozygosity in these lines. The landraces almost never impute using the whole site window, which is expected given their high heterozygosity and increased historical recombination. The landraces also use the two combination haplotype modes more often, and set more focus blocks to missing, reflecting a lack of accurate haplotypes.

The stringency settings chosen for FILLIN, which are also the defaults for the algorithm, were decided empirically based on these data, and were optimized for accuracy in inbred and breeding populations. Changing these thresholds leads to an increased number of genotypes imputed, but at the cost of accuracy. For the landrace populations especially, loosening the requirements rapidly leads to decreased accuracies while never imputing >60% of the minor alleles.

The gain in accuracy for FILLIN derived from more accurate imputation of minor alleles (Fig. 7). Figure 8 suggests that the increase in accuracy for minor alleles derives from FILLIN's insensitivity to the MAF. Gain in accuracy from accurate imputation of minor alleles is especially true for inbred lines (Fig. 7) and suggests that MAF insensitivity results from imputing one haplotype onto the inbred regions of the target taxon. For Beagle v. 4 and FILLIN in heterozygous populations, imputation accuracy is otherwise a function of the MAF, where lower frequency variants are imputed less accurately (Fig. 8).

These tests suggest that Beagle's advantage in heterozygous populations lies in their haplotype generation and phasing RefinedIBD (Browning and Browning, 2013) algorithm. FILLIN only draws haplotypes from shared identical-by-state segments, implicitly assuming that the haplotypes present in heterozygous lines are present as inbred regions somewhere in the dataset. This is not necessarily true, either because the inbred line or segment containing that haplotype may not have been sampled, or because it may not exist in a homozygous state because it contains a fatal deleterious allele. To test whether FILLIN was imputing landraces with haplotypes derived from inbred lines, we trained haplotypes for landrace imputation on only landraces and found that, when given no external information, FILLIN could not generate any haplotypes from the landraces until 3000 samples were input. For 3000 samples, resulting R^2 accuracies are <0.1 (Fig. 1). FILLIN achieves accuracies of around 0.5 when haplotypes are trained on the entire 4k sample dataset (Fig. 7). This suggests that only half of the haplotypes present in the landraces are present as inbred segments in GBS genotyped maize samples, and highlights FILLIN's inability to phase heterozygotes. However, we

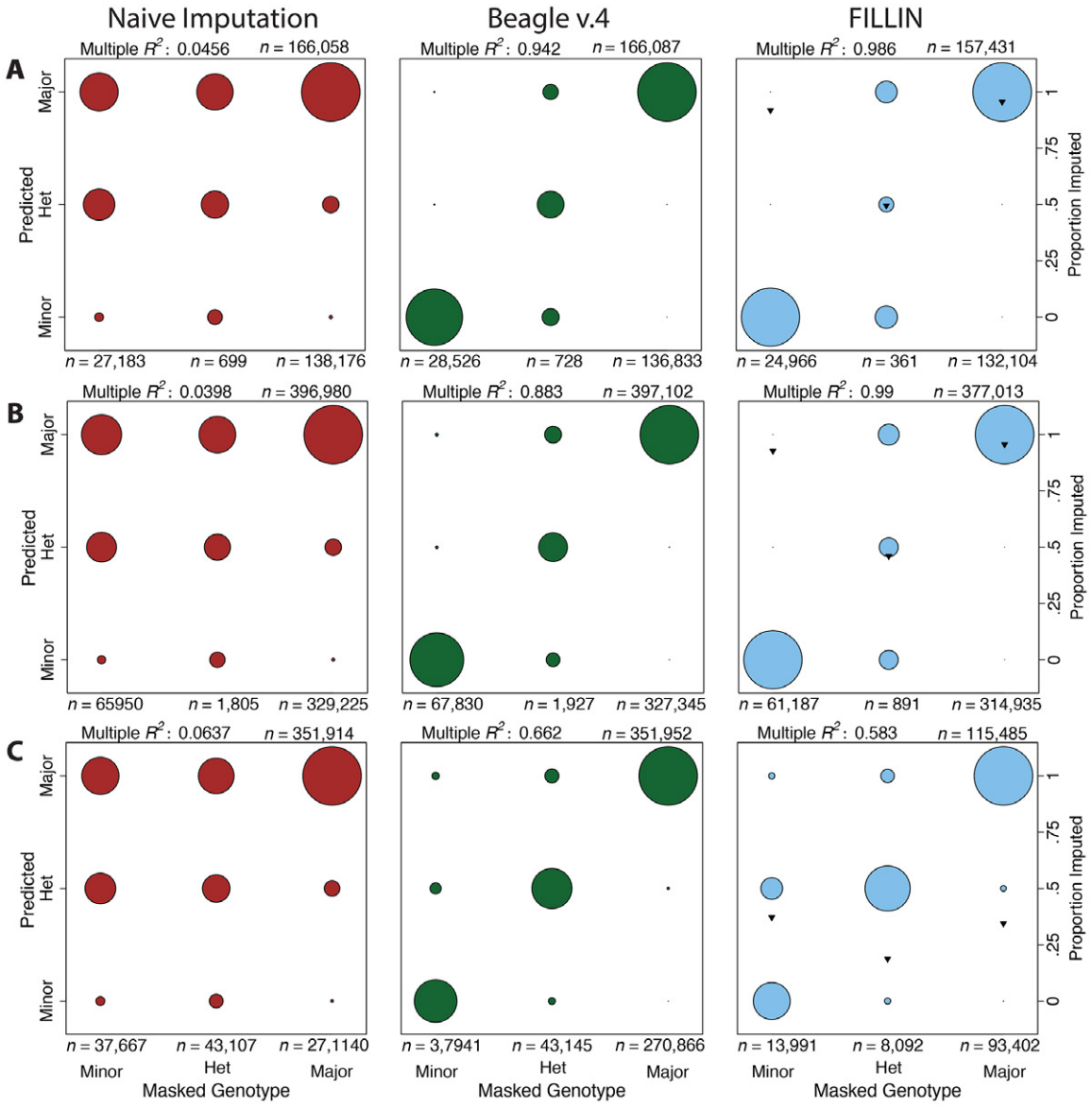

Figure 6. Accuracy for (A) temperate inbreds, (B) diverse inbreds, and (C) diverse landraces. The diameter of each circle represents the proportion within each known genotype class; the number of genotypes from each genotype class is indicated at the bottom of each known class. Note that the two inbred datasets both contain very few heterozygous genotypes relative to homozygous, which explains is why the R^2 value is affected very little by FILLIN's poor performance on heterozygous genotypes. Triangles mark proportion imputed by each known class; Beagle and the naive allele frequency imputation impute 100% of missing genotypes.

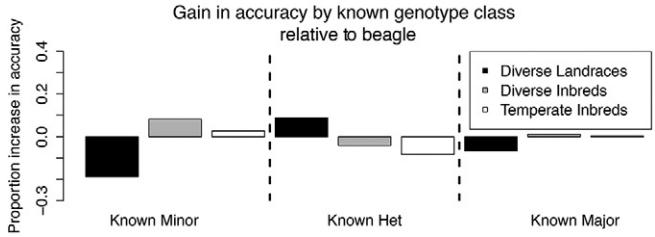

Figure 7. Gain in accuracy by Fast Inbred Line Library ImputatioN (FILLIN) relative to Beagle v. 4. If negative, Beagle is more accurate. FILLIN imputes very few of the masked heterozygotes for the Diverse Landraces, so the number of samples for comparison is very low, accounting for the apparent gain in accuracy for FILLIN.

found that FILLIN performed very well for inbred lines and biparental families, albeit not as well as when provided additional information. For inbred lines and families, accuracy as well as proportion of minor alleles imputed increased with sample size (Fig. 1).

Another feature of FILLIN is that it requires at least two taxa in a given block to share a haplotype to generate a donor. This is done to increase haplotype coverage, and to increase the robustness of the donor haplotypes, but means that diverse taxa at low coverage may not be represented in the donor file well (5.68% of masked, unimputed polymorphic sites are monomorphic in the donor file for the diverse landraces). Figure 1 shows that the more diverse the dataset, the more samples are required to adequately generate haplotypes.

Table 5. Method of imputation for windows averaged across taxa (±1 SE) and proportion of each method for those site windows that go into focus block imputation. All populations have 116 windows.

	Mode for window		Proportion of focus block			
Test dataset	Site window	Focus block	Inbred	Viterbi	Combo	Missing
Temperate inbreds	102.76 ± 0.97	13.24 ± 0.97	0.03 ± 0.0017	0 ± 0	0.84 ± 0.0038	0.13 ± 0.0038
Diverse inbreds	109.54 ± 0.77	6.46 ± 0.77	0.08 ± 0.0025	0 ± 0	0.77 ± 0.0043	0.15 ± 0.0049
NAM RILs†	115.05 ± 0.05	0.93 ± 0.05	0.04 ± 0.0008	0 ± 0	0.63 ± 0.0027	0.34 ± 0.0029
Diverse landraces	2.15 ± 0.37	113.84 ± 0.37	0 ± 0.0001	0 ± 0	0.35 ± 0.0059	0.65 ± 0.0059

† NAM, nested association mapping; RILs, recombinant inbred lines.

Table 6. Accuracy (R^2) for both imputations separately, the consensus imputation, and the accuracy for Beagle when the inbred imputation chooses not to impute. The consensus imputation is always more accurate than either method alone, but imputes the fewest missing genotypes.

	R^2		
Algorithm(s) used	Diverse landraces	Diverse inbreds	Temperate inbreds
FILLIN†	0.583 (115,485)	0.990 (377,013)	0.986 (157,431)
Beagle	0.737 (352,033)	0.891 (397,163)	0.949 (166,097)
Both agree	0.809 (98,325)	0.995 (367,828)	0.995 (155,591)
Beagle, when inbred does not impute	0.606 (236,518)	0.753 (20,106)	0.810 (8,661)

† Fast Inbred Line Library Imputation.

For all tested datasets, imputation accuracy improved when only the consensus genotypes from both imputation methods were accepted (Table 6). A consensus approach gains from the strengths in each imputation method: FILLIN's sensitivity to inbred segments and Beagle's to highly heterozygous regions. While it is very difficult to accurately identify and mask structurally missing variation in GBS data, the lower R^2 for Beagle at sites that FILLIN chooses not to impute suggest that FILLIN does provide sensitivity to structural variation by allowing for residual missing data in the haplotypes. The consensus approach also reduces the potential for overimputation, which is important in species with high structural variation such as crop plants (Chia et al., 2012).

Tests with different sized datasets suggested that the gain in computational time by FILLIN relative to Beagle increased with sample number: where FILLIN scales linearly with sample size, while Beagle runtime increases exponentially (Fig. 9). Here again, Beagle performs better on heterozygous taxa than inbreds, and vice versa for FILLIN, as shown by the change in rank between heterozygous and inbred datasets by method.

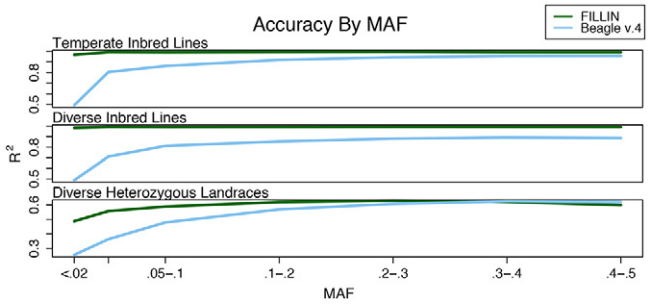

Figure 8. Accuracy for sites with different minor allele frequencies (MAF). The MAF for Fast Inbred Line Library ImputatioN (FILLIN) is taken from the donor haplotype files, while the MAF for Beagle v. 4 and Fast Inbred Line Library ImputatioN is taken from the unimputed input data.

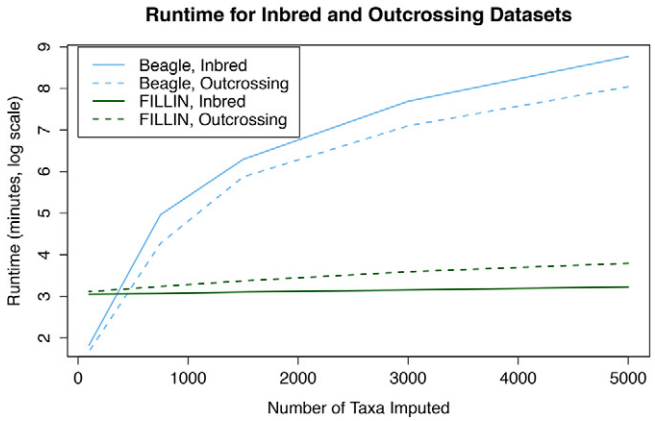

Figure 9. Computational time for Fast Inbred Line Library ImputatioN (FILLIN) and Beagle v. 4. Only one chromosome (chromosome 10) was compared, since computing the whole genome with >1000 samples using Beagle v. 4 is intractable unless parallelized. Each subset is a random sample of the larger taxa set.

Discussion

FILLIN and FSFHap produce highly accurate imputed genotypes, especially for closely related populations with replicated inbred segments. Cross contamination is difficult to completely exclude in even controlled crosses, and a number of maize GBS genotyped lines contain errors in pedigree. Because FILLIN and FSFHap do not require known parental genotypes, these algorithms provide a pedigree independent imputation method. If imputing full-sib families, FSFHap is optimized for modeling recombination, which allows it to more accurately impute heterozygotes. Beagle v. 4 provides more accurate imputation for highly heterozygous populations.

Accurate and complete haplotype generation is critical to high accuracies in both Beagle v. 4 and FILLIN imputed datasets. Results in Fig. 8 suggest that the relative gain in accuracies for the two algorithms for different types of datasets directly reflects the strengths of the two algorithms in haplotype generation. It is impossible in both

algorithms to impute a minor allele correctly if that variant does not exist in the haplotypes. Beagle is better able to phase and extract haplotypes from heterozygous taxa, and consequently imputes heterozygous datasets better for both heterozygous genotypes and genotypes homozygous for the minor allele. FILLIN focuses on extracting phased haplotypes from inbred segments, and imputes minor alleles better for homozygous datasets.

Overall, FILLIN provides rapid imputation of large sample size, low-coverage, whole-genome sequence data from predominantly inbred or breeding populations with high overall accuracy. For full-sib families where the objective is to find recombination breakpoints or to do linkage analysis, which requires IBD information, FSFHap provides sensitive and accurate imputation. For highly heterozygous samples with unknown segregating parental haplotypes, we recommend at this time that researchers use Beagle v. 4 (Browning and Browning, 2013). If highly accurate imputation is required, taking a consensus imputation will provide the most accurate results (Table 6).

Together, these three algorithms, Beagle v. 4, FILLIN and FSFHap, provide robust imputation of low-coverage GBS data from diverse populations. High quality haplotypes are required for accurate imputation by any of the algorithms presented here. Thus, if genotyping unrelated inbred lines or heterozygous populations in a species without available haplotype panels, resources should be expended to genotype a subset of individuals covering the diversity of haplotypes at higher coverage to ensure accurate haplotype generation and subsequent imputation. These results suggest that even one generation of selfing, in species where that is possible, can aid in accurate imputation of low-coverage genotyped populations. For breeders or researchers desiring to use GBS for genotyping highly related breeding populations, these results suggest that very-low-coverage genotyping, combined with FILLIN or FSFHap imputation, will provide highly accurate results at low cost. This is true even if the parents have not been sampled elsewhere, since numerous low-coverage replicates of each haplotype are expected within the population. Well-thought-out experimental design can help keep genotyping costs low (<$20 per sample), which enables efficient breeding and conservation biology decisions to be made.

Acknowledgments

This work was supported by the National Science Foundation (DBI-0820619, DBI-0922493, NSF IOS-1238014, IOS-0965342), the USDA-ARS, and The National 973 Program of China (Project No. 2011CB100106). The authors also wish to acknowledge SAGARPA (La Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación), Mexico for funding under the MasAgro (Sustainable Modernization of Traditional Agriculture) initiative. The Spanish landrace accessions are maintained by Misión Biológica de Galicia and were provided by Dr. Amando Ordás.

References

Andolfatto, P., D. Davison, D. Erezyilmaz, T.T. Hu, J. Mast, T. Sunayama-Morita, and D.L. Stern. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21:610–617. doi:10.1101/gr.115402.110

Bhullar, N.K., M. Mackay, and B. Keller. 2010. Genetic diversity of the Pm3 powdery mildew resistance alleles in wheat gene bank accessions as assessed by molecular markers. Diversity 2:768–786. doi:10.3390/d2050768

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Browning, B.L., and S.R. Browning. 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88:173–182. doi:10.1016/j.ajhg.2011.01.010

Browning, B.L., and S.R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471. doi:10.1534/genetics.113.150029

Browning, B.L., and Z. Yu. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. 85:847–861. doi:10.1016/j.ajhg.2009.11.004

Buckler, E.S., J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J.C. Glaubitz, M.M. Goodman, C. Harjes, K. Guill, D.E. Kroon, S. Larsson, N.K. Lepak, H. Li, S.E. Mitchell, G. Pressoir, J.A. Peiffer, M.O. Rosas, T.R. Rocheford, M.C. Romay, S. Romero, S. Salvo, H.S. Villeda, H.S. da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M.D. McMullen. 2009. The genetic architecture of maize flowering time. Science 325:714–718. doi:10.1126/science.1174276

Chia, J.-M., C. Song, P.J. Bradbury, D. Costich, N. de Leon, J. Doebley, R.J. Elshire, B. Gaut, L. Geller, J.C. Glaubitz, M. Gore, K.E. Guill, J. Holland, M.B. Hufford, J. Lai, M. Li, X. Liu, Y. Lu, R. McCombie, R. Nelson, J. Poland, B.M. Prasanna, T. Pyhäjärvi, T. Rong, R.S. Sekhon, Q. Sun, M.I. Tenaillon, F. Tian, J. Wang, X. Xu, Z. Zhang, S.M. Kaeppler, J. Ross-Ibarra, M.D. McMullen, E.S. Buckler, G. Zhang, Y. Xu, and D. Ware. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44:803–807. doi:10.1038/ng.2313

Cleveland, M.A., J.M. Hickey, and B.P. Kinghorn. 2011. Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. BMC Proc. 5(Suppl. 3):S6. doi:10.1186/1753-6561-5-S3-S6

Daetwyler, H.D., G.R. Wiggans, B.J. Hayes, J.A. Woolliams, and M.E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317–327. doi:10.1534/genetics.111.128082

Das, B., S. Sengupta, S.K. Parida, B. Roy, M. Ghosh, M. Prasad, and T.K. Ghose. 2013. Genetic diversity and population structure of rice landraces from eastern and north eastern states of India. BMC Genet. 14:71. doi:10.1186/1471-2156-14-71

Diers, B., J. Specht, D. Hyten, R. Nelson, and B. Beavis. 2011. Soybean nested association mapping. Soybean Breeder's Workshop, St. Louis, MO.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):E19379. doi:10.1371/journal.pone.0019379

Flint-Garcia, S., A. Thuillet, J. Yu, G. Pressoir, S. Romero, S. Mitchell, J. Doebley, S. Kresovich, M. Goodman, and E. Buckler. 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J. 44:1054–1064. doi:10.1111/j.1365-313X.2005.02591.x

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9(2):E90346. doi:10.1371/journal.pone.0090346

Gusev, A., J.K. Lowe, M. Stoffel, M.J. Daly, D. Altshuler, J.L. Breslow, J.M. Friedman, and I. Pe'er. 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19:318–326. doi:10.1101/gr.081398.108

Heerwaarden, J. van, J. Doebley, W.H. Briggs, J.C. Glaubitz, M.M. Goodman, J. de J.S. Gonzalez, and J. Ross-Ibarra. 2011. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. USA 108:1088–1092. doi:10.1073/pnas.1013011108

Hickey, J.M., B.P. Kinghorn, B. Tier, J.H. van der Werf, and M.A. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44(1):9. doi:10.1186/1297-9686-44-9

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 Genet. 5(6):E1000529. doi:10.1371/journal.pgen.1000529

Hurwitz, B.L., D. Kudrna, Y. Yu, A. Sebastian, A. Zuccolo, S.A. Jackson, D. Ware, R.A. Wing, and L. Stein. 2010. Rice structural variation: A comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. Plant J. Cell. Mol. Biol. 63:990–1003.

King, E.G., S.J. Macdonald, and A.D. Long. 2012. Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics 191:935–949. doi:10.1534/genetics.112.138537

Liu, E.Y., M. Li, W. Wang, and Y. Li. 2013. MaCH-admix: Genotype imputation for admixed populations. Genet. Epidemiol. 37:25–37. doi:10.1002/gepi.21690

Mace, E.S., C.H. Hunt, and D.R. Jordan. 2013. Supermodels: Sorghum and maize provide mutual insight into the genetics of flowering time. Theor. Appl. Genet. 126:1377–1395. doi:10.1007/s00122-013-2059-z

McMullen, M.D., S. Kresovich, H.S. Villeda, P. Bradbury, H. Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, P. Brown, C. Browne, M. Eller, K. Guill, C. Harjes, D. Kroon, N. Lepak, S.E. Mitchell, B. Peterson, G. Pressoir, S. Romero, M.O. Rosas, S. Salvo, H. Yates, M. Hanson, E. Jones, S. Smith, J.C. Glaubitz, M. Goodman, D. Ware, J.B. Holland, and E.S. Buckler. 2009. Genetic properties of the maize nested association mapping population. Science 325:737–740. doi:10.1126/science.1174320

Meuwissen, T., and M. Goddard. 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441–1449. doi:10.1534/genetics.110.113936

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77:257–285. doi:10.1109/5.18626

Ramu, P., C. Billot, J.-F. Rami, S. Senthilvel, H.D. Upadhyaya, L. Ananda Reddy, and C.T. Hash. 2013. Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. Theor. Appl. Genet. 126:2051–2064. doi:10.1007/s00122-013-2117-6

Revilla, P., P. Soengas, M.E. Cartea, R.A. Malvar, and A. Ordas. 2003. Isozyme variability among European maize populations and the introduction of maize in Europe. Maydica 48:141–152.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, R.J. Elshire, C.B. Acharya, S.E. Mitchell, S.A. Flint-Garcia, M.D. McMullen, J.B. Holland, E.S. Buckler, and C.A. Gardner. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14(6):R55. doi:10.1186/gb-2013-14-6-r55

Ross-Ibarra, J., M. Tenaillon, and B.S. Gaut. 2009. Historical divergence and gene flow in the genus *Zea*. Genetics 181:1399–1413. doi:10.1534/genetics.108.097238

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes, Genomes, Genet. 3:427–439.

Schnable, J.C., N.M. Springer, and M. Freeling. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA 108:4069–4074. doi:10.1073/pnas.1101368108

Spencer, C.C.A., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5(5):E1000477. doi:10.1371/journal.pgen.1000477

Vigouroux, Y., J.C. Glaubitz, Y. Matsuoka, M.M. Goodman, J. Sánchez G., and J. Doebley. 2008. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Am. J. Bot. 95:1240–1253. doi:10.3732/ajb.0800097