Review of Genotype Imputation Algorithms

Darrell Aucoin (1003562316), Nicolas Brandt (1003649386)
Department of Computer Science
University of Toronto
27 King's College Cir, Toronto, ON M5S
{daucoin, nbrandt}@cs.toronto.edu

Abstract

In many genetic studies, especially genome wide association studies (GWAS), inference must be done on the whole spectrum of the genome. Because of the impractical cost of sequencing the whole genome for all individuals of a study, researchers often rely on sequencing only a portion of the genome and imputation methods to resolve the missing SNPs. These methods allow to reconstruct an estimation of the whole genome of an individual given a subset of it. Numerous softwares were developed to respond to this need and are regularly updated. In this paper, we aim to compare the most popular population-based imputation algorithms.

1 Introduction

Genotype imputation is the precursor to many genetic studies, particularly genome wide association studies (GWAS). This is because it is often impractical to sequence whole genomes for all individuals in the study due to the prohibitively high cost of time and money involved. In such cases, if the whole genome is needed, then the rest of the genome must be imputed. This is done with the aid of a database of known haplotypes (groups of genes that are inherited together), the polymorphic sites of study individuals, and an imputation algorithm. These algorithms perform a form of haplotype estimation ("phasing") to find the best haplotype for that region of the genome, this haplotype is then used for imputation. This report will be an overview of various genotype imputation algorithms and their strengths and weaknesses.

These genotype imputation algorithms typically start by with the known single-nucleotide polymorphisms (SNPs for all diploid chromosomes) of individuals in the study and a database of haplotypes, usually from HapMap 3, or the 1000 Genomes Project. The algorithm then tries matching these haplotypes to the known data often using a combination of MCMC, EM algorithm, and a hidden markov model (HMM). The imputations are then based off of these matched haplotypes.

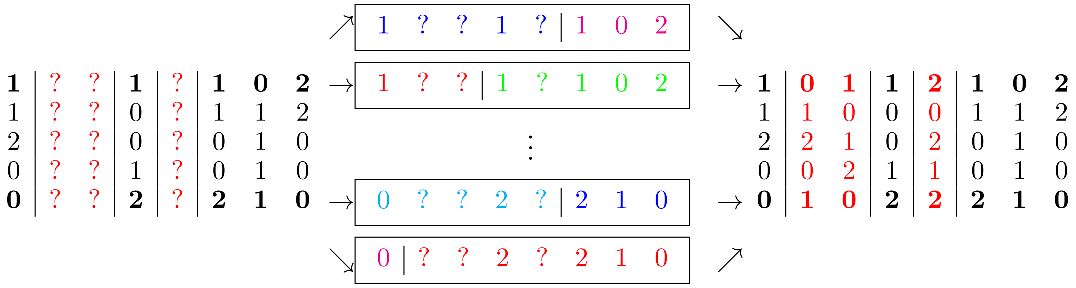


Figure 1: Genome Imputation Algorithm using Haplotypes

Intuitively choosing a good haplotype reference simply based on relevant details like ethnicity or geological region is difficult as humanity has had many migration and mixing events throughout it's history. This means that machine learning and probabilistic modeling is the best course of action for imputation. Also, it is usually better to provide all possible reference haplotypes, not just haplotypes clustered to that ethnicity or geological region. This is because although an individual's haplotypes are highly correlated with ethnicity, due to mixing, a larger mixture of haplotypes will likely tell a more complete story.

2 Current Approaches

Population-based algorithms maintain and improve multiple candidate solutions, each solution corresponding to a unique point in the search space of the problem. In our case, each solution correspond to an haplotype. A haplotype is a group of genes that frequently occur together.

Most of the literature for imputing genotypes uses variants of coalescent-based Hidden Markov Models (HMM). The hidden markov model basically works by estimating a hidden state of the genome of the individual (the haplotypes h) based on the sequence genome given and outputting the most likely genome given these haplotypes. These are coalescent, in that they incorporate the assumption that new haplotypes are produced by mutation and recombination of older haplotypes. Alleles close to each other are more likely to come from the same haplotype than alleles farther apart. We will be looking at various models like Beagle[1], IMPUTE v2[2], fastPHASE[3], and Minimac.

There are several measures of imputation correctness, but most are based around estimation of imputation accuracy, the correlation between true and imputed genotypes. Imputation accuracy has a linear relationship with genomic prediction accuracy, but due to the inability to know the true genotypes other measures are used to approximate accuracy: allelic R^2, standardized allele-frequency error, ratio of imputed allele dosage variance and true allele dosage variance, allele-frequency correlation. The different softwares are using different metrics to approximate accuracy. Nevertheless, a previous study [3] has revealed that those metrics are closely related and could be compared.

As we will see in the following parts, there are a few differences between the algorithms we studied. Nevertheless, as they all belong to population inference methods and, therefore, share also common points. The following table, extracted from whole-genome haplotyping approaches and genomic medicine[4] highlight these similarities.

Table 1: Pros and Cons of Population Inference Methods

Advantages	**Limitations**
Cost-effective	Can only phase common variants
Facilitates haplotype imputation in samples with low-density microarray panels	Difficult to impute private variants or rare haplotypes
Useful when family members cannot be ascertained	Limited by the accuracy and availability of suitable reference data
Large sample sizes increase accuracy	Generates short-range haplotypes
Good for large samples of unrelated individuals	Sample size impacts haplotype frequency estimations
Incorporation of family duos and trios improves accuracy	Methods are probabilistic and accuracy must be balanced against computational costs

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668004/

[2] http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000529

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1424677/

[4] https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-014-0073-7

2.1 Beagle

Beagle is based on a graphical model of haplotypes. A tree of haplotypes describing various sets of haplotypes where each edge is weighted by the number of haplotypes that pass along it. In each level pairs of nodes are compared in terms of downstream haplotype frequencies. If the sum of the squared differences of their downstream partial haplotype frequencies do not exceed a threshold then the nodes are combined[10]. This is essentially clustering haplotypes by how close they resemble each other.

The algorithm then iteratively prunes the tree of low information branches for the individual's dataset, resulting in a tree with few edges with low linkage disequilibrium (LD) and many edges with high LD[10]. Linkage disequilibrium is a measure of non-random association of alleles at two or more loci, this is affected by two parameters: the distance between SNPs and the recombination rate.

Beagle then fits the model to the current set of haplotypes and resamples new estimated haplotypes based on each individual's fitness model[10]. The probabilities of the missing genotypes are calculated from this final model.

This design incorporates the coalescent assumption of mixtures of haplotypes without the need to estimate any parameters while adapting to the local haplotype diversity in the individual's data.

2.2 FastPHASE

FastPHASE is another cluster-based model for haplotype variation. It is based on the observation that haplotypes tend to cluster into groups of closely related or similar haplotypes. It models implicitly the genealogy of chromosomes in a random sample from a population as a tree but summarizes all haplotype variation in the "tips" of the trees (cf. [25]).

FastPHASE was the first algorithm to make it possible to phase genome-wide SNP data (cf. [10]) and is really efficient for small sample sizes. For larger sizes, computational feasibility is maintained at the cost of loss of information and accuracy.

For most of the recent studies involving fastPHASE and another imputation algorithm, fastPHASE scored comparatively low on both accuracy and computation efficiency.

2.3 IMPUTE v2

IMPUTE2 uses a HMM where the hidden states emit the observed genotypes. The transition probabilities are based on recombination events and the emission probabilities reflecting the probability of mutation and error. Emission probability is constant if the mutation rate is constant and transmission probability is determined by the fine-scaled recombination map of human genome. During the imputation process, IMPUTE2 will exploit all the study and reference haplotypes. In order to reduce complexity, IMPUTE2 uses Hamming distance to select close haplotypes. This allows IMPUTE2 to be able to accommodate large reference panels.

The advantages of this method are that as IMPUTE2 is using all the haplotypes, where most of the other algorithms only allow for a subset of the haplotypes to be considered. Allowing for all haplotype considerations increases accuracy slightly. Moreover, as IMPUTE2 was mostly developed to be used on the human genome, it is highly accurate on many types of human populations. Furthermore, IMPUTE2 possess a relatively efficient computation by using approximation by only adding haplotypes when accuracy increases.

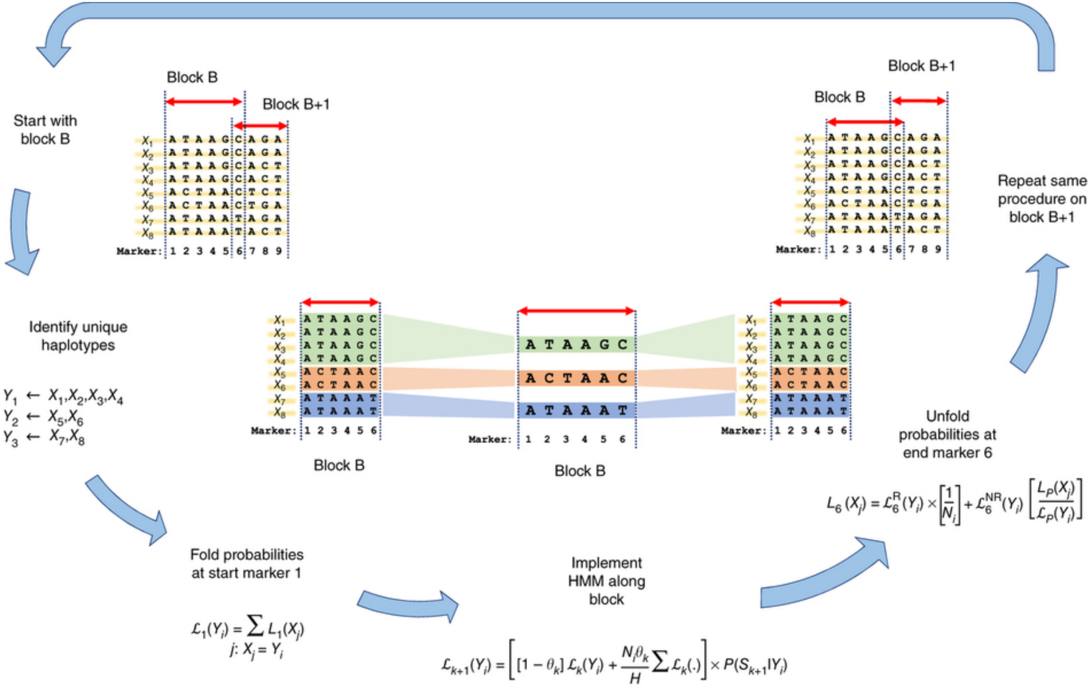
IMPUTE2 allows pre-phasing which decreases overall runtime at the expense of a small decrease in accuracy. Pre-phasing work by statistically phasing the study genotypes and imputing from the reference panel into the estimated study haplotypes. This is a new feature in IMPUTE2 comparative to IMPUTE1 but it is also present in all of the other algorithms.

2.4 MaCH/Minimac

The Imputation algorithm of MaCH and IMPUTE2 differ in two aspects. Where IMPUTE2 uses fixed values for the transmission and emission probabilities on the HMM, MaCH estimates these parameters using Baum-Welch algorithm. Moreover, while IMPUTE2 uses both study and reference haplotypes, MaCH randomly selects only 200 haplotypes. These features allow MaCH to work well with data where most genotypes are missing and need to be imputed.

Minimac is a low memory and optimized version of MaCH. Minimac exploits similarities among haplotypes in small genomic segments in order to reduce the number of states over which the HMM iterates. The whole process is represented in the following figure, extracted from Next-generation genotype imputation service and methods [23].

Figure 2: Overview of state space reduction



Start with block B
Block B
Block B+1
Marker: 1 2 3 4 5 6 7 8 9
Identify unique haplotypes
$Y_1 \leftarrow X_1, X_2, X_3, X_4$
$Y_2 \leftarrow X_5, X_6$
$Y_3 \leftarrow X_7, X_8$
A T A A G C
A C T A A C
A T A A A T
Marker: 1 2 3 4 5 6
Fold probabilities at start marker 1
$\mathcal{L}_1(Y_i) = \sum_{j:\, X_j = Y_i} L_1(X_j)$
Implement HMM along block
$\mathcal{L}_{k+1}(Y_i) = \left[[1-\theta_k]\, \mathcal{L}_k(Y_i) + \frac{N_i \theta_k}{H} \sum \mathcal{L}_k(.) \right] \times P(S_{k+1} \mid Y_i)$
Unfold probabilities at end marker 6
$L_6(X_j) = \mathcal{L}_6^{R}(Y_i) \times \left[\frac{1}{N_i}\right] + \mathcal{L}_6^{NR}(Y_i) \left[\frac{L_P(X_j)}{\mathcal{L}_P(Y_i)}\right]$
Repeat same procedure on block B+1

In the third version of Minimac, the complexity of the algorithm depends on the number of unique haplotypes in each genomic segment and the total number of such segments in the reference panel. As a result, it scales better than linearly over the range of reference panel sizes which is better than most of the current imputation algorithms.

3 Strengths and Weaknesses

The following table (Tables 2, 3) comes from a merging between the comparison tables present in *Genotype Imputation for Genome-Wide Association Studies* [3] and *Current Software for Genotype Imputation* [22] as well as our own study of each software.

In table 3, we compare error rate for two different scenario, A and B. Both scenarios and datasets come from *A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies* [14], written by the developers of IMPUTE2.

In scenario A, the SNPs in the dataset are partitioned in two disjoint subsets: the first is genotyped in all individuals and the second only in the haploid reference panel. The objective was to estimate the genotypes of SNPs in the later subset in the study sample.

In scenario B, the SNPs are partitioned in three disjoint subsets: the first is genotyped in all individuals, the second (U2) is genotyped in both the haploid and diploid reference panels but not the study sample, and the last (U1) is genotyped only in the haploid reference panel. In this scenario, we aim to estimate the genotypes of SNPs in U2 in the study sample and SNPs in U1 in both the study sample and, the diploid reference panel. The restricted scenario B is obtained by removing SNPs in U1.

Table 2: Comparisons of Imputation Algorithms

Properties	IMPUTE v1	IMPUTE v2	MaCH v1.0.16	fastPHASE v1.4.0 BIMBAM v0.99	Beagle v3.2
Operating System	Linux, Solaris, Windows, Mac	Linux, Solaris, Windows, Mac	Linux, Windows, Mac	BIMBAM (source code + Windows) fastPHASE (Linux, Solaris, Windows, Mac)	Java executable
License	Free for academic use	Free for academic use	Unclear	Commercial license	Free
Source code	No	No	Available	No	Available (Beagle v.4.1)
Documentation	Clearly structured	Clearly structured	Wiki and FAQ	Decent	Decent
Reference panels					
Reference panels available in correct format	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data	HapMap2	No
Can use a haplotype reference panel?	Yes	Yes	Yes	Yes	Yes
Can use a genotyped reference panel?	No	Yes	Yes	Yes	Yes
Can two haplotype or genotype references panels be used in the same run?	No	Yes	No	No	No
Output files					
Genotype posteriors produced?	Yes	Yes	Yes	Yes	Yes
Information measures?	Yes	Yes	Yes	No	Yes
Easiest use of output files to test association	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Feed files directly into SNPTEST. Test based on genotype posteriors, dosages or thresholded genotypes	Genotype dosage files can be fed into MACH2DAT or MACH2QTL	BIMBAM can produce file formats used by BIMBAM. fastPHASE out files need to be processed	Best-guess phased haplotypes can be tested in Beagle. Processing required to use genotype posteriors or dosage

Table 3: Comparisons of Imputation Algorithms: Study Samples, Program Options, and Performance

Properties	IMPUTE v1	IMPUTE v2	MaCH v1.0.16	fastPHASE v1.4.0 BIMBAM v0.99	Beagle v3.2
		Study samples			
Can take genotypes specified with uncertainty?	No	Yes	No	No	Yes
Can accommodate trios and related samples?	No	No	No	No	Trios and duos
Can impute into a study sample of autosomal haplotypes?	Yes	Yes	No	No	Yes
Can impute on the X chromosome?	Yes	Yes	No (Yes MiniMac3)	No	Yes
		Program options and features			
Does phasing as well as imputation?	No	Yes	Yes	Yes	Yes
Can impute sporadic missing genotypes?	No	Yes	Yes	Yes	Yes
Has internal performance assessment?	Yes	Yes	Yes	No	No
Can impute only in a specified interval?	Yes	Yes	No	No	No
Can handle strand alignment between data sets?	Yes	Yes	Yes	No	No
SNP and sample inclusion and exclusion options?	Yes	Yes	No	Yes	Yes
Joint model for imputation and association testing?	No	No	No	No	No
		Computational performance			
Assessment 1[5]	43m (1000 Mb)	75m (180 Mb)	105m (80 Mb)	855m (16 Mb)	56m (3100 Mb)
Assessment 2[6]	–	48m (115m)	–	157m (211m)	104m (234m)
Error rates					
Scenario A	5.42%	5.16%	5.46%	5.92%	6.33%
Scenario B (restricted)	–	3.4% (0.86%)	–	5.33% (1.32%)	3.46% (0.93%)
Scenario B (full) data sets	–	3.4% (0.86%)	–	–	4.01% (1.04%)

[5] Imputation of 1377 samples on the Affy500k chip from 120 CEU HapMap2 haplotypes; 7.5 Mb region. Data from [14].
[6] Imputation of 500 (1000) samples genotyped at 872 SNPs from 1000 haplotypes at 8712 SNPs in a 5 Mb region. Timings based on data sets simulated using HAPGEN and the pilot CEU haplotypes from the 1000 Genomes project in a 5 Mb region on chromosome 10.

[5]
[6]

The results obtained in term of computational performance were not necessarily reflecting the usual performance of each software. In particular, for IMPUTE1 and IMPUTE2 the efficiency in time and memory might come from the fact that only those 2 softwares can impute on just a segment of a chromosome whereas the other softwares are imputing on the whole chromosome.

Most of the papers we reviewed concentrated on comparing IMPUTE2 and Beagle, but never the same versions between them. Nevertheless, among all the comparisons, we managed to find a similar trend. IMPUTE2, Beagle and MaCH/Minimac seem to have a pretty high accuracy with fastPHASE lagging behind. In most of the cases, the accuracy for IMPUTE2 was the highest. Conversely, Beagle was often depicted as the fastest and the most memory-efficient.

For divergent sample mixes, Beagle seemed to do better than IMPUTE2. This comes from the fact that IMPUTE v2 is only considering haplotypes with small Hamming distance from the target sample during phasing. IMPUTE2 can improve its accuracy by increasing the amount of haplotypes to use as templates but at the cost of longer running time. On the contrary, IMPUTE2 can be used to only analyze a part of a chromosome, a feature not available in the other softwares. Thus, in some specific cases, IMPUTE2 can be accurate as well as time and memory efficient.

As mentioned previously, the various softwares are updated regularly. In particular, Beagle 4.1 and Minimac3 were released this year. In a recent comparison done by the developers of Minimac3 [23], both of them had better computational performance than IMPUTE2. It would be interesting to compare the 3 algorithms on different datasets in order to deepen the study.

3.1 Factors Influencing Imputation Accuracy

There are several factors that can influence the accuracy of the genotype imputation: SNP sample size, marker density, genotype accuracy, degree of relatedness, ethnicity of the individuals, and the allele frequency[10].

Sample Size: If the SNP sample size is increased, the haplotype phasing accuracy increases and thus genotype imputation accuracy increases.

Marker Density: With increased marker density, haplotype estimates give more locally accurate haplotypes. However, on a regional basis with an absolute accuracy measure this will create more opportunities for errors and thus lower accuracy. It is thus recommended to have a moderate marker density.

Genotype Accuracy: Genotype accuracy (controlled by the sequencing error rate and coverage) influences haplotype phase accuracy. For noisy or incomplete data, usually from low coverage, a solution for better accuracy is to phase genotype likelihoods to capture the uncertainty of the genotype data. The genotype and haplotype phase posteriors are estimated simultaneously, increasing the accuracy of both tasks.

Degree of Relatedness: If the study individuals are related, then this can be incorporated into the model to significantly increase imputation accuracy. Even if relatedness is not specified in the model, the number of haplotypes needed for imputation on the whole dataset is significantly less. And due to many of the algorithms' need to limit the number of haplotypes for efficient computation, the resulting haplotype phasing accuracy and genotype imputation accuracy will increase.

Sample Ethnicity: The haplotype diversity of various human subpopulations vary dramatically. For instance, African populations have significantly higher haplotype diversity than non-African populations, like Europeans. Allele frequencies and density of polymorphisms are greater confounding factors for haplotype dense vs haplotype thin ethnicities. A greater SNP sample size and/or larger haplotype database might be needed for the same imputation accuracy.

Allele Frequency: Rare allele variations are more difficult to phase because confidently phasing these alleles would require the variant to be seen several times within it's haplotype context. Thus, studies on rare genetic diseases must be handled with care, relying on relatedness of individuals to properly impute the missing genotype, or the use of experimental phasing methods.

3.2 Chromosome X Imputation

Imputation analysis of chromosome X is more difficult to impute than autosomes as the X chromosome is diploid in females and hemizygous for males. For imputation on males, if the X-chromosome data was just copied again to gain a full complementary set then the algorithms would give greater importance to the haplotypes associated with the X-chromosome. This could possibly affect imputation of the rest of the genome[26].

There is also less sequenced data available on the X-chromosome, meaning that there are less known haplotypes available and imputation accuracy is significantly weaker.

Most of the algorithms discussed now make adjustments to impute on chromosome X, based on if the individual is male or female. IMPUTE2 requires a separate file specifying which subjects are male or female. Beagle and the latest version of MaCH, minimac, also allows imputation on the X-chromosome.

4 Possible Improvements

In term of accessibility, we were unable to find an API for the algorithms studied to either R or Python to streamline the data analysis. A user-friendly API would help users to quickly integrate imputation analysis along with the analysis of the study. This kind of module could help researchers to easily change the imputation algorithm to reflect the researchers' needs.

We also saw previously that population-inference methods all shared common limitations (cf. table 1). An idea mentioned in *Combining Family and Population-Based Imputation Data for Association Analysis of Rare and Common Variants in Large Pedigrees* [15] was to combine two different kind of methods in order to increase accuracy. The result was that the combination of the two methods (GIGI and Beagle) was indeed more accurate than using Beagle alone. It would most likely be interesting to deepen this study and see if the combination of two methods can improve accuracy without degrading computation time and memory too much.

5 Conclusion

In this report we focused on 4 population inference algorithms. We reviewed the particularities of each one of them and tried to give an insight of their particular strengths and weaknesses. We referred to numerous comparisons to find which algorithm was the most efficient and accurate. Among Beagle, IMPUTE2 and MaCH/Minimac, none were really outperforming the others and the results mostly depended on the dataset. We brought to light different factors that could influence the accuracy of each software. Nevertheless, as Beagle and Minimac were updated recently (Beagle 4.1 and Minimac3), we didn't find enough information to really judge of the efficiency of the updated versions. We expect them to slightly outperform the previous releases. On the last part, we discussed on different means to improve the softwares. We considered that creating an API to R or Python might improve the accessibility of such softwares. Moreover, we saw that in some specific cases combining different kinds of methods would improve accuracy.

6 Contributions

In the beginning, we both took notes on various papers into a google document. We relied on this document for the individual write ups, as well as the other's assistance when needed. For the individual parts of the paper:

Darell: Introduction, Beagle, Factors Influencing Imputation Accuracy, and Chromosome X Imputation

Nicolas: Current Approaches, fastPHASE, IMPUTE2, MaCH/Minimac, Strengths and Weaknesses, Possible Improvements, and Conclusion

Darrell did the finishing touches on bringing the document together, LaTeXing it, and final editing.

Bibliography

[1] N. Landwehr, T. Mielikäinen, L. Eronen, H. Toivonen and H. Mannila (2007), *Constrained hidden Markov models for population-based haplotyping*. BMC Bioinformatics

[2] M. Cristani, A. Perina, L. Xumerle, V. Murino, P. F. Pignatti and G. Malerba (2008), *Fully non-homogeneous hidden Markov model double net: A generative model for haplotype reconstruction and block discovery*. Elsevier Ltd.

[3] J. Marchini and B. Howie (2010), *Genotype imputation for genome-wide association studies*. Macmillan Publishers Limited

[4] J. C. Huang, Q. D. Morris, T. R. Hughes and B. J. Frey (2005), *GenXHC: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data*. Oxford University Press

[5] M. Bink and R. van Binsbergen (2015), "Imputation of genotype data - Introduction to theory and implementation of Genomic Selection". ProCoGen 3rd Training Workshop "Bioinformatics and trees". Vienna. 11 March 2015

[6] O. Shai, Q. D. Morris, B. J. Blencowe and B. J. Frey (2006), *Inferring global levels of alternative splicing isoforms using a generative model of microarray data*. Oxford University Press

[7] T. S. Jaakkola and D. Haussler (1999), *Exploiting generative models in discriminative classifiers*

[8] A. Laughbaum (2013), "Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods for Accuracy, Computation Time, and Memory Usage", `http://blog.goldenhelix.com/`

[9] P. Ma, R. F. Brøndum, Q. Zhang, M.S. Lund and G. Su (2012), *Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle*. Elsevier Ltd.

[10] S. R. Browning and B. L. Browning (2011), *Haplotype phasing: Existing methods and new developments*. Nature Publishing Group

[11] J. Griesman (2012), *Imputing genotypes using regularized generalized linear regression models*. Joshua Griesman

[12] P. Scheet and M. Stephens (2008), *Documentation for fastPHASE 1.4*

[13] B. L. Browning and S. R. Browning (2009), *A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals*. Elsevier Ltd.

[14] B. N. Howie, P. Donnelly and J. Marchini (2009), *A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies*. PLoS Genetics

[15] M. Saad and E. M. Wijsman (2014), *Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees*. Genet Epidemiol.

[16] M. Saad, A. Q. Nato Jr, F. L. Grimson, S. M. Lewis, L. A. Brown, E. M. Blue, T. A. Thornton, E. A. Thompson and E. M. Wijsman (2016), *Identity-by-descent estimation with population- and pedigree-based imputation in admixed family data*. BMC Proceedings

[17] B. Howie and J. Marchini, "Impute2", `http://mathgen.stats.ox.ac.uk/impute/impute_v2.html`

[18] B. L. Browning and S. R. Browning (2016), *Genotype Imputation with Millions of Reference Samples*. Elsevier Ltd.

[19] G. Glusman, H. C. Cox and J. C. Roach (2014), *Whole-genome haplotyping approaches and genomic medicine*. BioMed Central Ltd.

[20] M. Nothnagel, D. Ellinghaus, S. Schreiber, M. Krawczak and A. Franke (2009), *A comprehensive evaluation of SNP genotype imputation*. Hum Genet

[21] L. Jostins, K. I. Morley and J. C. Barrett (2011), *Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets*. Macmillan Publishers Limited

[22] D. Ellinghaus, S. Schreiber, A. Franke and M. Nothnagel (2009), *Current software for genotype imputation*. Henry Stewart Publications

[23] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E Locke, A. Kwong, S. I Vrieze, E. Y Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. Loh, W. G Iacono, A. Swaroop, L. J Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R Abecasis and C. Fuchsberger (2016), *Next-generation genotype imputation service and methods*. Nature Genetics

[24] "Minimac", `http://genome.sph.umich.edu/wiki/Minimac`

[25] P. Scheet and M. Stephens (2006), A *Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase*. AM J Hum Genet

[26] A. L. Wise, L. Gyi, T.i A. Manolio (2013), *eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses*. Elsevier Ltd.