A comparative study of genotype imputation programs

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Xuguang Yang

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9
Or
Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Background Genotype imputation infers missing genotypic data computationally, and has been reported to be highly useful in various genetic studies; e.g., genome-wide association studies and genomic selection.

Motivation While various genotype imputation programs have been evaluated via different measurements, some, such as Pearson correlation, may not be appropriate for a given context and may result in misleading results. Further, most evaluations of genotype imputation programs are focused on human data. Finally, the most commonly used measurement, concordance, is unable to determine a difference in performance in some cases.

Research Questions (1) How do popular genotype imputation programs (i.e., Minimac and Beagle) perform on plant data as compared to human data? (2) Can we find measures that better discriminate imputation performance when concordance does not? and (3) What do alternate measures indicate for the performance of these imputation programs?

Methods Since *Kullback-Leibler divergence* (K-L divergence) and *Hellinger distance* can aid in ranking statistical inference methods, they can be highly useful in our study. To amplify signals from K-L divergence and Hellinger distance, we obtain their negative logarithmic values (i.e., negative logarithmic K-L divergence (NLKLD) and negative logarithmic Hellinger distance (NLHD)) so that larger values indicate better imputation results. With NLKLD and NLHD, we investigate the performance of two existing genotype imputation programs (i.e., Beagle and Minimac) on data from plants, specifically *Arabidopsis thaliana* and rice, as well as human. For each pair of organisms to be compared, we select data from one chromosome of each organism such that approximately the same number of samples/participants and SNPs are present for each organism. Finally, we apply different missing rates for target datasets and different sample size ratios between reference and target datasets for sensitivity analysis of the imputation programs.

Results We demonstrate that in a general case where single nucleotide polymorphisms (SNPs) with different minor allele frequencies (MAFs) are imputed at the same concordance, both NLKLD and NLHD capture a difference in the imputation performance. Such a difference reflects not only the difference of correspondence between the known and imputed MAFs, but also the difference of chance agreement between the known and imputed genotypes. Additionaly, neither Minimac nor Beagle performs better on either *A. thaliana* or human data. However, Beagle performs better on human data than on rice data. Finally, the majority of both NLKLD and NLHD results from all experimental data indicate that Minimac outperforms Beagle.

Conclusions (1) Although neither Minimac nor Beagle consistently performs better on either plant or human data, Beagle evidently performs better on human data than on rice data; (2) NLKLD and NLHD can be more discriminating than concordance and should be considered in comparing different genotype imputation programs to determine superior imputation methods; and (3) the NLKLD and NLHD results suggest that Minimac's imputation method is superior to Beagle's. Further study can involve confirming these trends with runs on more experimental data.

Acknowledgements

I would first like to thank my thesis advisor Prof. Tony Kusalik of the Department of Computer Science at the University of Saskatchewan. The door to Tony's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work and steered me in the right direction whenever he thought I needed it.

I would also like to thank the statistics experts who offered me helpful advice on how to present my results informatively: Dr. Longhai Li and Dr. Juxin Liu (ordered alphabetically by the last name) of the Department of Mathematics and Statistics at the University of Saskatchewan. Without their passionate participation and input, the validation of my research would barely have been successfully conducted.

Moreover, I would like to acknowledge Dr. Matthew Links and Dr. Ian McQuillan (ordered alphabetically by the last name) of the Department of Computer Science at the University of Saskatchewan as my thesis advisory committee members, and I am gratefully indebted to their valuable comments on this thesis.

Further, I would like to thank the Department of Computer Science and the Plant Phenotyping and Imaging Research Centre for funding my thesis project. Without them, I would never have had this great opportunity to conduct any great project or meet so many great researchers and fellow students from whom I learned a great deal.

Finally, I must express my very profound gratitude to my family and to my laboratory colleagues, Jason Bernard, Lingling Jin, Farhad Maleki, Kimberly Mackay, Daniel Hogan and Morgan Kirzinger for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Dedication

To my parents and brother

Contents

List of Tables

List of Figures

List of Abbreviations

A. thaliana	*Arabidopsis thaliana*
GS	Genomic selection
GWAS	Genome-wide association study
HMM	Hidden Markov model
indel	insertion/deletion
IQS	Imputation quality score
K-L distance/divergence	Kullback-Leibler distance/divergence
MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
mrate	missing rate
NGS	Next-generation sequencing
NLHD	Negative logarithmic Hellinger distance
NLKLD	Negative logarithm Kullback-Leibler distance/divergence
ref:tgt	sample size ratio between reference (ref) and target (tgt)
SNP	Single nucleotide polymorhism
SSE	sum of squared estimates of errors
VCF	Variant call format

Chapter 1

Introduction

Recently, the food industry has seen increased utilization of plant-based products. Plants such as wheat and rice are important components in our everyday diets, and their high productivity is partly a result of plant phenotyping and genotyping studies. One type of plant study investigates the whole genomes of a set of samples to find genetic variants that have potential associations with a certain phenotypic trait. However, such association studies rely heavily on the whole genome datasets [11,15], and due to technological constraints such as genotyping errors and limited sets of genetic markers, missing genotypic records become commonplace. In light of this, genotype imputation aims to infer the missing portion of genotypes in a group of study samples via computational tools.

Many genotype imputation studies have been done to compare performance (e.g., running time, memory usage and imputation accuracy) of different genotype imputation programs in the context of human data [11, 15]. However, acquisition of human data often raises ethical concerns, and the fundamental differences between human and plant genomes complicate whether the imputation results from human carry over to plants. This question may be answered by evaluating genotype imputation performance using plant data. Its result may be highly beneficial for future plant research. Further, statistical measures such as Pearson correlation that have been used to evaluate imputation results might not be appropriate in a given context and might lead to misleading results. In addition, the most commonly used measure, concordance, may not be sufficiently discriminating to rank imputation performance for different imputation methods in some cases. Hence, it is necessary to seek more discriminating statistical measures to rank imputation methods. With more discriminating statistical measures, we can compare performance of a genotype imputation program between plant and human, and rank imputation performance between genotype imputation programs.

Genetic data has been widely utilized in various fields such as human disease susceptibility research, genomic selection programs and quantitative trait loci detection for cattle and plant breeding programs. Genetic data comes in various forms, including whole genome sequencing data and genotypic data. Genetic data is typically obtained via next-generation sequencing technologies. However, numerous technical constraints result in errors and missing values in the data. While addressing the error rate of genetic data is beyond the scope of this study, reducing the missing rate of genetic data can be achieved by genotype imputation. Genotype imputation infers missing genotypic data computationally, and has been reported to be highly useful in various genetic studies; e.g., genome-wide association studies and genomic selection. For example, Li et al. [22] and Orho-Melander et al. [34] demonstrated that common missense mutations imputed by genotype imputation showed strong association with their studied diseases. They concluded that genotype imputation could accelerate genotype-phenotype studies. Additionally, Willer et al. [39] and Kathiresan et al. [21] showed that genotype imputation increased the significance of specific genotyped markers in their respective genome-wide association studies (GWAS). Therefore, genotype imputation can boost the power of GWAS.

Unfortunately, the performance of genotype imputation programs has been either poorly understood or subject to misinterpretation because of inappropriate statistical measures. For instance, Pearson correlation has been applied to ascertain the agreement between the known and imputed genotypes. However, genotypic data does not necessarily follow a normal distribution, which is a requirement of the measure, and when the imputed genotypes have zero variance, a divide-by-zero error occurs in the calculation. In addition, concordance (described in section 2.4) is sometimes insufficient to discriminate imputation results with different MAFs. Finally, evaluations have often been restricted to human data and it is unclear whether these imputation programs perform well on non-human datasets.

In our study, we evaluate two genotype imputation programs, Beagle [11] and Minimac [15], based on their performance on human, *Arabidopsis thaliana*, and rice data. We measure performance using concordance, IQS, negative logarithmic Kullback-Leibler divergence (NLKLD), and negative logarithmic Hellinger distance (NLHD). Specifically, our research answers the following research questions:

1. How do popular genotype imputation programs (i.e., Minimac and Beagle) perform on plant data as compared to human data?

2. Can we find measures that better discriminate imputation performance when concordance does not? and

3. What do alternate measures indicate for the performance of these imputation programs?

In this thesis we present the results of our study. We also explain why Pearson correlation is a misleading measurement when used for evaluating imputation programs, and why NLKLD and NLHD are more appropriate to evaluate performance of genotype imputation programs. Finally, we discuss problems with existing genotype imputation programs, and provide recommendations on how they might be addressed so as to facilitate improved genotype imputation.

Chapter 2

Background

In this chapter, we provide background information of this study. In section 2.1, we explain genomic related terminology. In section 2.2, we discuss circumstances in which missing genotypic data could occur. In section 2.3, we explain how various imputation methods work. In sections 2.4 and 2.5, we explain previously-used measures, i.e., concordance and imputation quality score. In section 2.6, we explain why Pearson correlation might not be an appropriate measure for genotype imputation programs. Finally, in section 2.7, we discuss studies that compared genotype imputation programs, and we focus on the measures that were used in those studies.

2.1 Terminology

In this section, we assume that readers know common terms such as DNA and chromosome.

Diploid Diploid is a cell or an organism that has paired chromosomes and each chromosome is from one parent. The context of this study revolves around diploid organisms in order to conform with most genotype imputation studies.

Gene Genes are basic physical units of inheritance that contain the information needed to specify traits. Genes are linearly organized on structures called chromosomes [5].

Genetic variation Genetic variation is differences of genes between individuals or populations. A genetic variation can be a mutation of a single base pair in a DNA sequence, but also a large-scale difference involving thousands of base pairs [6].

Allele Alleles are different versions of a genetic variant such as a SNP or indel. An allele can either refer to a single DNA position or a portion of a DNA sequence. A base allele is a single version of the genetic variant. For a SNP, each base allele is a pair of nucleotides. Since most genotype imputation studies are restricted to bi-allelic sites where only two different alleles, i.e., reference and alternate alleles, are possible, we focus on the bi-allelic sites of SNP data. The allele with lower frequency is the minor allele and the one with higher frequency is the major allele. In other words, the reference allele is not necessarily the major allele and the alternate allele is not necessarily the minor allele.

Phenotype A phenotype is a set of an individual/sample's observable traits; e.g., height, eye colour, biomass, etc [9].

Genotype Genotype is a set of individual/sample's genes [8]. A genotype is one combination of alleles in a specific region of a chromosome for one sample/participant. For example, in a diploid organism where sexual reproduction is involved, suppose that on the reference sequence at location L we have an allele, and that alleles A (from the paternal chromosome) and a (from the maternal chromosome) are present in the genome of individual P at L. Then we say that P has a genotype Aa.

Single nucleotide polymorphism A single nucleotide polymorphism (SNP) is a difference between individuals/samples in a single DNA building block, called a nucleotide; i.e., adenine, thymine, cytosine, or guanine [10]. SNPs, along with other common genetic variants such as insertion/deletion (indel) and repetition, may directly associate with a specific phenotypic trait, and hence have become highly useful in genetic studies.

Minor allele frequency Suppose A represents the major allele and a represents the minor allele. Then, P_{AA}, P_{Aa}, and P_{aa} are calculated probabilities of three genotypes for a specific SNP from a given set of genotypic data. The minor allele frequency (MAF) of such a SNP is $0 \times P_{AA} + 0.5 \times P_{Aa} + P_{aa}$.

Genome-wide association study Genome-wide association study (GWAS) is an approach to discover potential genetic variations that contribute to specific genotypic traits [7].

Imputed dosage An imputed dosage is a linear transformation of the highest posterior probability of the imputed genotype at one specific genetic location for one sample/participant. For a diploid organism, a genotype can be AA (two major alleles), Aa (one major allele, one minor allele), or aa (two minor alleles). Typically, we assign 0 to the major allele and 1 to the minor allele so that AA becomes 0, Aa becomes 1 and aa becomes 2. Suppose that AA has a posterior probability 0.1, Aa 0.1 and aa 0.8. Since aa has the highest posterior probability, the imputed dosage is $2 \times 0.8 = 1.6$. Suppose instead that AA has a posterior probability 0.1, Aa 0.8 and aa 0.1. Since Aa has the highest posterior probability, the imputed dosage is $1 \times 0.8 = 0.8$.

Genetic marker Genetic markers are short DNA sequences with known physical locations on chromosomes. In a genetic marker, DNA segments are close to each other and are likely to be inherited together. Genetic markers are highly useful in detecting SNPs.

Variant calling SNP data is a product of variant calling, which detects genetic variants in the population of a given species by aligning target DNA sequences to a reference DNA sequence.

Genotyping Genotyping (also known as genotype calling) identifies alleles of each individual at specific genetic locations where SNPs are detected [33]. Genotyping is not necessarily performed by variant calling.

Haplotyping Haplotyping extracts alleles that belong to individual chromosomes.

Haplotype block A haplotype block is a set of genetic variants in close proximity on a chromosome, where such genetic variants come from a single chromosome without the effect of chromosome recombination.

Phasing For a diploid organism, phasing assigns alleles to the chromosome where genetic variants occur. In a genotypic data file, a vertical bar between two alleles is used for phased genotypes while a slash is used for unphased genotypes. Let 0 represent the reference allele and 1 represent the alternate allele. Then, for a given SNP, four genotype combinations are present in phased genotypic data: $0|0$, $0|1$, $1|0$ and $1|1$, whereas three genotype combinations are present in unphased genotypic data: $0/0$, $0/1$ and $1/1$.

Synthetic data Synthetic data, in the context of our study, are obtained via computational processes instead of genotyping.

Genotyping platform Genotyping platforms are devices that detect and generate genotypic data.

2.2 Missing genotypic data

Missing genotypic data is common and can happen in one or more of the following occasions [17]:

- Missing genotypic data can occur due to biological reasons; e.g., individuals/samples have different numbers of copies of genes.

- When a genotype calling process is assigning alleles to all individuals at a given genetic location, the genotyping platform may encounter low certainty in determining which alleles should be assigned to some individuals and will mark such alleles as missing.

- When samples/participants come from different study groups and are genotyped on different genotyping platforms, the varied genetic markers between the platforms may lead to missing genotypic data. For instance, suppose group A (GA) uses platform PA with genetic markers MA and group B (GB) uses platform PB with genetic markers MB. Suppose MA and MB have a shared subset of markers MC. From PA we obtain a set of genotypic data (DA) and from PB we obtain another set of genotypic data (DB). When we combine DA and DB into one, samples/participants from GA have missing genotypes at markers MA−MC (MA excluding markers from MC), and samples/participants from GB have missing genotypes at markers MB−MC (MB excluding markers from MC).

- Two groups of samples/participants are genotyped on the same platform. However, A uses a platform with a dense set of genetic markers whereas B uses a sparse set of genetic markers. In this case, samples/participants from group B have missing genotypes.

Typically, reference genotypic data does not have missing genotypes and missing genotypes occur in target genotypic data.

The following two paragraphs explain what are entirely and partially missing genotypes.

Entirely missing genotypes When target samples/participants are genotyped on a platform with less dense genetic markers as compared to the references, certain SNPs that are present in the references do not appear in the targets. In this case, the target samples/participants have entirely missing genotypes for such SNPs.

Partially missing genotypes During a genotyping process, some samples/participants have undetected alleles for a given SNP. In this case, the target SNP has partially missing genotypes.

2.3 Genotype imputation

The basic operation of a reference-based genotype imputation program is as follows. It first reconstructs haplotype blocks in both reference and target genotypic data, and then infers missing genotypes for the target genotypic data. Figure 2.1 shows the framework of a generic reference-based genotype imputation program.

Panel A shows the original genotypic data of both references and targets (where missing alleles are marked with dots). In common cases, however, only a minority portion of a genotype is missing. Each column represents a genetic location whereas each row represents a partial haplotype of one sample/participant. Panel B shows matched haplotype blocks between reference and target genotypic data. Between a pair of reference and target, the likely matched haplotype blocks have the same colour. However, as can be seen in Panel B, haplotype blocks may vary in size and the size of a haplotype block is determined by the closest match between reference and target data. Panel C shows that the target haplotype blocks are fully imputed and the imputed alleles are the same as in their reference haplotype blocks.

Over the past decade, numerous genotype imputation programs have either emerged or been upgraded from their older versions. In 2001, Stephens and colleagues [38] applied a Bayesian statistical method to develop PHASE for haplotype reconstruction. In 2006, considering that similar haplotype blocks exist over a short region within a chromosome, Scheet and Stephens [37] utilized a hidden Markov model (HMM) in their fastPHASE. In 2007, Marchini and colleagues [27] also implemented an HMM in their IMPUTE program,

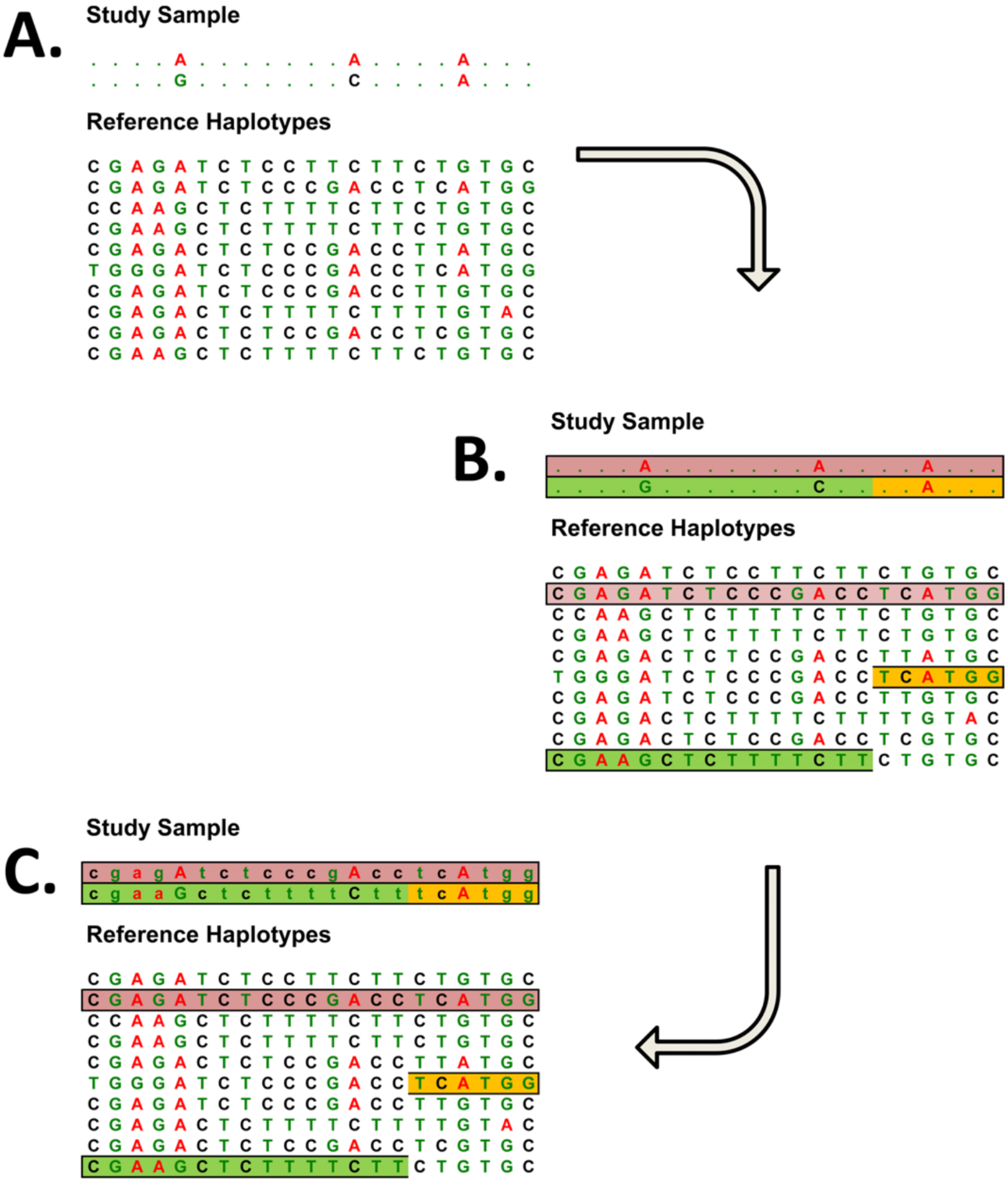
A.
Study Sample
. . . . A . . . . A . . . . A . . .
. . . . G . . . . C . . . . A . . .
Reference Haplotypes
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
B.
Study Sample
. . . . A . . . . A . . . . A . . .
. . . . G . . . . C . . . . A . . .
Reference Haplotypes
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
C.
Study Sample
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
Reference Haplotypes
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

Figure 2.1: Imputation of missing genotypes by inference from common haplotype blocks (from Li et al. [23]. Imputed alleles are designated with lowercase letters.

and validated their imputed results. However, later in the same year, Browning and Browning [12] argued that these imputation programs did not scale well for large datasets involving thousands of individuals. In light of this, they developed a localized haplotype-cluster model and incorporated it in their Beagle program. In 2009, Howie et al. [20] upgraded IMPUTE to its second version. In 2010, Li et al. [23] developed MaCH using a Markov chain haplotyping model. Such a model was highly similar to an HMM. A year later, Howie and colleagues [19] implemented a Markov chain Monte Carlo (MCMC) algorithm in their IMPUTE2, which was not only capable of identifying which parent chromosomes a set of genotypes came from, but also capable of inferring genotypes that had not been called. However, all the above-mentioned genotype imputation programs heavily depend on reference genotypic data that has high fidelity and no missing genetic information. In other words, these programs cannot run without the presence of reference genotypic data, and hence become nonapplicable for imputation practices where genotypic reference data is unavailable. To address this issue, in 2015 Money and colleagues [28] developed LinkImpute utilizing a k-nearest neighbor genotype imputation method. In 2016, Browning and Browning [11] upgraded Beagle using parallelization to make it memory efficient for imputation from millions of reference samples. In the same year, Das et al. [15] incorporated an expectation-maximization algorithm along with MCMC in their Minimac3 program.

2.4 Concordance

Concordance is the percent agreement between the known and imputed genotypes for each SNP, and is the most intuitive way to measure the accuracy of imputation results. However, we observe that genetic variant data is highly dense at SNPs with low MAFs (<0.1), and that imputation programs tend to assign the major alleles to all samples/participants. Such a situation would result in high concordance, yet the imputation results might not be useful. However, we still report concordance in this paper to be comparable with other studies.

2.5 Imputation quality score (IQS)

IQS [24], based on Cohen's kappa (κ), determines agreement between raters for categorical (or qualitative) variables [3]. In this case, raters are a genotyping platform that generates the ground-truth genotypes and an imputation program that produces imputed genotypes. Categorical variables are the ground-truth and imputed genotypes.

The definition of κ is as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the concordance and p_e is the hypothetical chance agreement probability. In Table 2.1 are the counts of the agreement and disagreement between the known and imputed genotypes. We define p_0 by the following equation:

$$p_0 = \frac{\sum_i n_{ii}}{N}$$

and p_e is defined as:

$$p_e = \frac{\sum_i n_{i.} n_{.i}}{N^2}$$

.

Table 2.1: Counts of agreement and disagreement between the known (j) and imputed (i) genotypes. $n_{i,j}$ corresponds to the count of a particular pair of known and imputed genotypes, $n_{i.}$ corresponds to the marginal count between the imputed and all known genotypes, $n_{.j}$ corresponds to the marginal count between the known and all imputed genotypes, and N is the total number of genotypes for the given SNP.

		Known genotypes			
		AA	AB	BB	Total
Imputed genotypes	AA	n_{11}	n_{12}	n_{13}	$n_{1.}$
	AB	n_{21}	n_{22}	n_{23}	$n_{2.}$
	BB	n_{31}	n_{32}	n_{33}	$n_{3.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

To avoid divided-by-zero situations, we add $\epsilon = 0.00001$ to both the numerator and denominator in the calculation of κ. Hence, we formulate IQS as follows:

$$IQS = \frac{p_0 - p_e + \epsilon}{1 - p_e + \epsilon}$$

As illustrated by Lin et al. [24], $IQS = 1$ means complete agreement, $IQS = 0$ means complete disagreement, and $IQS < 0$ means the results are worse than a random guess.

2.6 Misuse of Pearson correlation coefficient

The Pearson correlation coefficient has been extensively used to ascertain the agreement between actual and imputed genotypes. As exemplified by Howie et al. [19], a Pearson correlation coefficient is calculated based on the known genotypes and the imputed dosages at each genetic locus. The known genotypes take values in $\{0, 1, 2\}$, and the imputed dosages take values in $[0, 2]$. However, using dosages and known genotypes to calculate Pearson correlation might be problematic both theoretically and practically.

First, Pearson correlation coefficient is defined as follows:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{2.1}$$

and its estimate is defined as follows:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2.2}$$

where x_i's and y_i's are known genotypes and imputed dosages, respectively, and n is the number of samples/participants. According to Eq 2.2, when either the imputed dosages or the known genotypes have zero variance, a divide-by-zero error will occur. During our study, such an error happened numerous times when imputed dosages from one SNP with low MAF were all zeros, while the references had non-zero alleles. As a result, the correlation rates were undefined.

Typically, Pearson correlation coefficient is applied to determine the extent of a linear relationship between two normally distributed continuous variables. In section 2.1, we determined that the known genotype was categorical and not a normally distributed continuous variable. Hence, using Pearson correlation between the known and imputed genotypes violates its assumptions.

Therefore, from both practical and theoretical perspectives, Pearson correlation is not appropriate to evaluate agreement between the known and imputed genotypes and should not be used in genotype imputation studies.

2.7 Related work

There have been many studies evaluating genotype imputation programs. Although these studies made numerous contributions to the field of genetic studies, they also had shortcomings.

In 2008, Zhao and colleagues [40] conducted a sensitivity study on the IMPUTE program to investigate how the program responded to different SNP missing rates (either entirely or partially missing genotypes for each SNP). However, they used only concordance to measure genotype imputation performance. A single measure might not thoroughly capture the goodness/badness of the program, especially since concordance can easily be inflated in low MAF regions [36].

In their study, Lin and colleagues [24] used IQS to evaluate the performance of the IMPUTE program. Also, they performed downstream genome-wide association studies for further validation. Unfortunately, they did not perform replicates to rule out the possibility that imputation results from the same input data with the same program settings might not be identical.

In their study, Hancock and colleagues [16] compared IMPUTE2, BEAGLE, MaCH, and MaCH-Admix. The authors reported that MaCH and IMPUTE2 had the highest overall concordance, and that IMPUTE2 had the highest IQS and average r2hat, which is an estimated Pearson correlation between known and imputed genotypes. Interestingly, the authors used data from 2 human chromosomes to test consistency of the MaCH program, yet used data from only one chromosome to test all the other imputation programs. Additionally, the authors did not examine their results at different MAFs. It was probable that concordance was inflated at SNPs with low MAFs, since for SNPs with low MAFs the program tended to assign major alleles to all participants and reach high percent agreement. Moreover, they applied a 2% missing rate but did not explain well why they chose such a missing rate. Finally, their use of Pearson correlation to evaluate genotype imputation programs was problematic.

Hickey and colleagues [17] used maize data to test the IMPUTE2 program. Their work used proportion of correctly imputed genotypes, which is similar to concordance, and Pearson correlation. Again the use of Pearson correlation raises concerns regarding the validity of their results.

Liu et *al.* [25] used human data to compare IMPUTE2, Minimac, and Beagle. Their study dataset included 90 participants as targets and 379 participants as references. However, numerous studies [11, 15, 19] have suggested that practical imputation programs should be able to handle data involving thousands of participants/samples (references and targets combined). Finally and unfortunately, they used Pearson correlation to evaluate the imputed results.

In their study, Ramnarine and colleagues [36] compared various measurements (i.e., concordance, squared correlation, IQS and program built-in measurements) to evaluate genotype imputation programs (i.e., Beagle v3.3.2 and IMPUTE2) using human data. Their study focused on whether results from alternative measurements were consistent with each other. Unfortunately, they did not notice that squared correlation, which is Pearson correlation, was not appropriate when evaluating imputation results.

Chapter 3

Methods and Materials

In this section, we present the materials we used, including programs (section 3.1), computing facilities, and data (section 3.3), as well as the methodologies and experiment designs (section 3.2). Section 3.4 presents detailed steps to make data suitable for a mask-impute scenario. In subsection 3.4.1, we discuss how we filtered SNP data, why we chose SNP data on specific chromosomes of different organisms, as well as how we created reference and target data before inputting it to the genotype imputation programs. In subsection 3.4.2, we present how we made the filtered data ground truth to which we compared the imputed results. In subsection 3.4.3, we present how we separated datasets into references and targets, as well as how we made the non-missing target datasets "missing". Section 3.5 presents the two measures that have not been considered in previous studies but might be highly useful in measuring the performance of genotype imputation programs. Section 3.6 presents our approach to analyze the imputation results.

3.1 Genotype imputation program selection

Our preliminary examinations of numerous genotype imputation programs showed that Minimac and Beagle were the most recently upgraded programs among all such software, and that they both had versatile input data format compatibility. We therefore selected Beagle and Minimac as the target programs in our study. However, since both Beagle and Minimac are reference-based genotype imputation programs, we also considered a non-reference-based genotype imputation program (i.e., LinkImpute) in our study to compare it with reference-based programs. Therefore, we selected the following genotype imputation programs:

1. Minimac3 (`https://genome.sph.umich.edu/wiki/Minimac3`, accessed on November 15, 2017.).

2. Beagle v.4.1 (`https://faculty.washington.edu/browning/beagle/b4_1.html`, accessed on November 15, 2017.)

3. LinkImpute (`http://www.cultivatingdiversity.org/software.html`, accessed on April 15, 2018.)

Additionally, we used the following programs for specific, limited purposes:

1. Eagle2 (`https://data.broadinstitute.org/alkesgroup/Eagle/downloads/`, accessed on April 19, 2018.) was selected for phasing the *A. thaliana* and rice datasets.

2. PLINK v.1.9 (`http://s3.amazonaws.com/plink1-assets/plink_linux_x86_64_20181202.zip`, accessed on April 19, 2018.) was used to convert data formats between the variant call format (VCF) and the PLINK ped format due to data format compatibility issues arising from the LinkImpute program.

3.2 Experiment layout

In this section, we describe experiment settings along with detailed steps to answer the research questions. Recall that the three research questions are: (1) How do Minimac and Beagle perform on plant data as compared to human data? (2) Can we find measures that better discriminate imputation performance when concordance does not? and (3) What do alternate measures indicate for the performance of these imputation programs? To answer the first question, we compared performance of imputation programs on human and *A. thaliana*, and on human and rice. To assess performance, we needed some form of ground truth data to compare with the imputed data. Therefore, we used synthetic data. The synthetic data in our study were generated via filtering and phasing, which were common processes in previous genotype imputation studies [11, 15]. We then needed appropriate measures to compare imputation results between different organisms. As discussed in section 2.6, Pearson correlation was an inappropriate measure. Also, concordance was insufficient to measure imputation performance. Hence, we needed to find alternate measures to interpret imputation results from different organisms. Such effort also corresponded to research questions (2) and (3).

Since this study mainly focuses on comparing human and plant imputation results, we did not design experiments to compare rice and *A. thaliana*. Considering the possibility that each imputation process could take a significant amount of elapsed time, CPU time and memory, as well as the difficulty of scheduling all individual computational jobs on our server, we chose to run the computational experiments on Compute Canada with fixed numbers of replicates.

We designed the following experiment steps to answer our research questions.

```
filter data;
phase data;
select samples/participants to make them references and targets with separation ratios of 5:5, 6:4, 7:3,
  8:2, and 9:1;
for separation ratio in [5:5, 6:4, 7:3, 8:2, 9:1] do
    make a copy of the original target dataset for accuracy measurement purposes;
    for missing rate in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8. 0.9] do
        generate synthetic data by selecting and masking SNP sites in genomes from the target;
        for iteration in [1...10] do
            input the reference and modified target datasets as well as any other required information to
              Beagle and Minimac;
            run Beagle and Minimac;
            aggregate the imputed SNP sites by logarithm-base-two of MAFs, filter out SNP sites with MAF
            less than 0.5% and then measure imputation performance using concordance, IQS, NLKLD, and
            NLHD;
        end for
    end for
end for
```

The separation ratio and missing rate numbers were chosen arbitrarily since there was no consensus in the literature on how to select ratios between reference and target genotypic data, or on typical missing rates of target genotypic data. Additionally, we visualized the above experiment layout in Figure 3.1. Finally, we ran all imputation programs with the same parameters settings (same CPU, memory and time limit configurations) on the Cedar cluster of Compute Canada (`cedar.computecanada.ca`).

3.3 Data acquisition

Numerous studies have suggested that a well-scaled imputation program should be able to handle data involving thousands of participants/samples. Therefore, we only selected data containing more than a thousand participants/samples. The following well-curated datasets (based on their frequent mention in numerous studies) were therefore acquired. The SNPs in the different datasets were determined via different genotyping processes.

- Human, 1000 Genomes [1], 2504 human participants at time of data acquisition;
- *Arabidopsis thaliana (A. thaliana)*, 1001 Genomes [2], 1135 samples at time data acquisition; and
- Rice, The European Bioinformatics Institute [4], 3000 samples at time of data acquisition.

All of the above datasets were downloaded in our study. In addition, the SNPs in these different datasets were determined in different ways. Moreover, since whole genome size varied between organisms, to compare performance of genotype imputation programs between plants and humans we randomly selected samples/participants so as to have the same number of samples/participants within each pair of organisms. Additionally, we selected data from chromosomes having approximately the same size within each pair of organisms. Such a data selection strategy was to minimize the possibility that the imputation results might be affected by the numbers of samples/individuals and SNP sites. Therefore, to compare results from human and *A. thaliana* for both target programs, we randomly selected 1,135 human participants from chromosome 22 data with 606,756 SNP sites, and selected all 1,135 *A. thaliana* samples from chromosome 4 data with 651,406 SNP sites; to compare results from human and rice for both target programs, we selected all 2,504 human participants from chromosome 13 data with 1,536,078 SNP sites, and randomly selected 2,504 rice samples from chromosome 12 data with 1,592,744 SNP sites. The human dataset had zero missing rate for

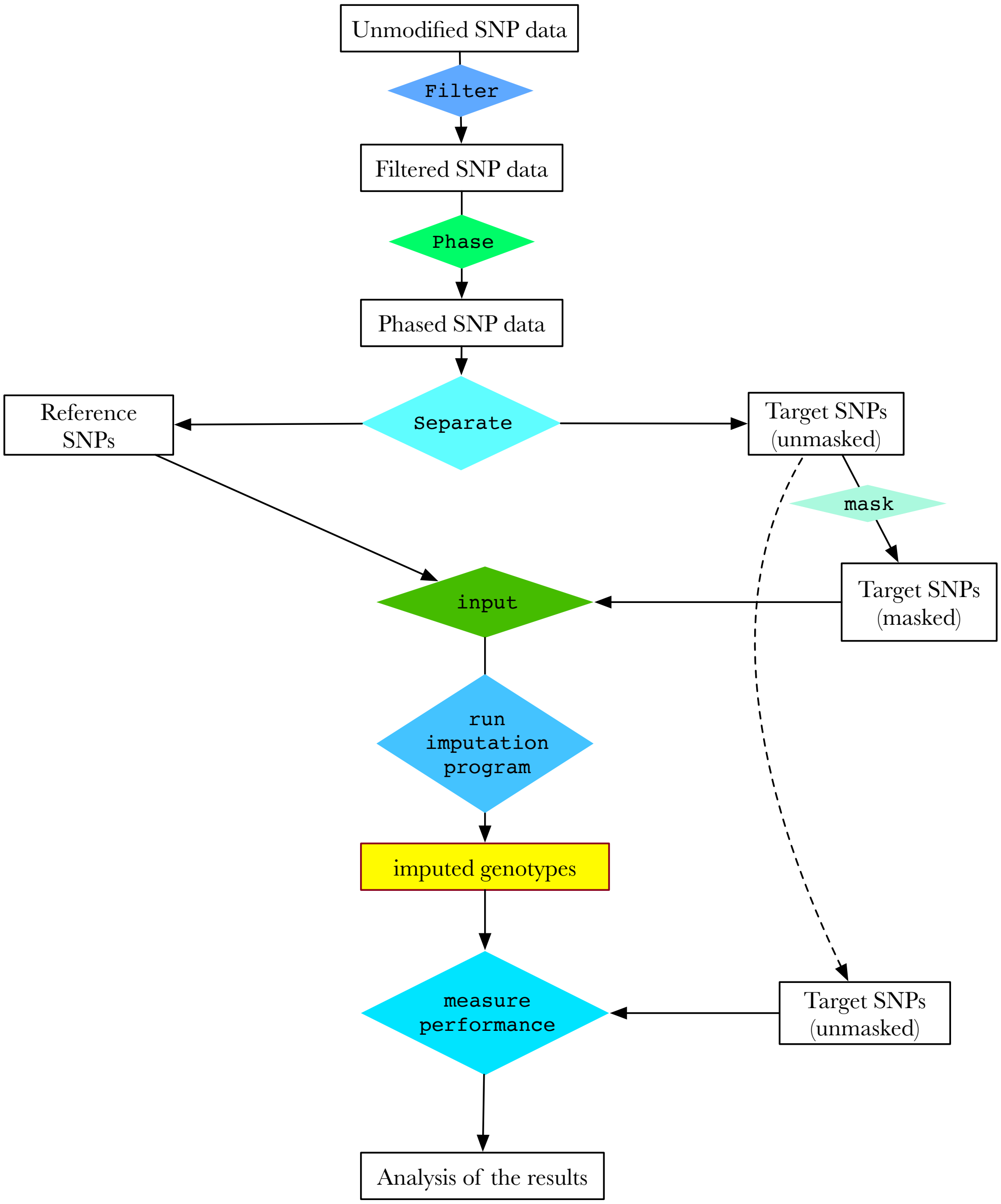

Figure 3.1: Workflow of imputation experiments ignoring iteration. Each rectangle represents data whereas each diamond represents an action on data. The arrows correspond to the directions in which data flows. Different colours are used only to label different data operations and data at different stages.

all SNP sites and all SNP sites were phased. However, neither of the two plant datasets had non-missing SNP sites and neither was fully phased.

3.4 Data preparation

In this section, we present detailed steps to prepare data for input to the genotype imputation programs. In subsections 3.4.1 and 3.4.2, we present how we created ground-truth SNP data to which we compared the imputed results. During these steps, we used phasing tools to make the plant data consistent with the human data. In subsection 3.4.3, we present how we separated the ground-truth data into references and targets, which were required input data of Beagle and Minimac.

3.4.1 Filter

Since numerous studies tested imputation programs with bi-allelic SNP sites, we chose only bi-allelic sites (step 1 in section 3.2). Additionally, as exemplified in previous studies [11, 15], we did not input SNP data for all chromosomes from each organism to the imputation programs since different organisms have different numbers of chromosomes and large differences in data size could potentially lead to misleading results. Instead, we only selected SNP data from a specific chromosome for testing. As mentioned in section 3.3, the plant datasets were not fully genotyped (i.e., all individuals were missing allelic information for at least one SNP position). Therefore, we filtered out SNP sites with high missing rates to make ground-truth data. (We later used the ground-truth data to validate the imputed results.) Additionally, since the human data was fully genotyped (i.e., 0% missing rate), we filtered the SNP sites in the *A. thaliana* and rice datasets to reduce the missing rate for the remaining SNPs. Ultimately, we only kept sites that had less than 5% missing rate for the following reasons:

1. We aimed to minimize the missing rate of the original SNP data so that we could maximize the number of ground-truth SNP sites. Since we set a series of missing rates for generating synthetic data (step 6 in section 3.2) starting at 10%, the missing rate threshold of filtering the original SNP data should be under 10%.

2. We filtered SNP sites with various, increasing missing rate thresholds starting at 0 in order to obtain a sufficient amount of data:

    - First, we filtered SNP sites by applying a missing rate threshold 0; i.e., only sites with a 0% missing rate were retained. However, this resulted in an insufficient number of SNP sites (<50 SNP sites on the *A. thaliana* chromosome 4 data).
    - We then applied a 1% missing rate threshold, yet the number of SNP sites did not increase significantly.

- Finally, we applied a 5% missing rate threshold and this time we were able to find chromosome data that had approximately the same number of SNP sites between organisms. Specifically, with a 5% missing rate threshold, we obtained 228,168 SNP sites for the *A. thaliana* chromosome 4 data, 220,336 SNP sites for the human chromosome 22 data, 383,467 SNP sites for the rice chromosome 12 data, and 559,611 SNP sites for the human chromosome 13 data.

Thus, the 5% missing rate threshold was adopted.

3.4.2 Making ground-truth data

Unlike the human SNP data, where all participants were fully genotyped, the *A. thaliana* and rice data still had numerous sites with partially missing genotypes after the SNP filtering process. In addition, both Minimac and Beagle required the reference genotypic data to be fully phased and the *A. thaliana* and rice data were not phased. Moreover, both programs required that the reference genotypic data to be non-missing. However, the plant data could not meet such a requirement so the data in our study needed to be further refined. Hence, phasing was an inevitable step (step 2 in section 3.2) in this study. Numerous phasing programs could not only phase SNP data but also impute missing genotypes. To keep datasets from different organisms as consistent as possible, we chose Eagle2 [26] to phase both the *A. thaliana* and rice datasets for the following reasons:

- We tried Shapeit [18], yet it had the following two major issues:
  1. it was developed by the authors of one of the selected imputation programs (i.e., Minimac) and could potentially introduce bias in further imputation results; and
  2. Shapeit had a fatal memory leak issue that eventually caused a system crash.
- Beagle could also be used for phasing, yet it was not considered for the following two reasons:
  1. similar to Shapeit, it was exactly one of the imputation programs to be tested so it could introduce bias; and
  2. Beagle did not work properly for phasing when individual samples had missing SNP sites in the reference data.
- Finally, Eagle2 was efficient at phasing huge datasets and easy to use.

The ground-truth genotypic data is also called the known genotypes in this document.

3.4.3 Separating ground-truth data to references and targets

For each comparison pair, we applied the following ratios to create subsets of randomly selected samples to make references and targets (steps 3 and 4 in section 3.2): 5:5, 6:4, 7:3, 8:2, and 9:1; at each separation ratio, we randomly selected SNP sites to make them “missing” with each of nine missing rates [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%] (step 6 in section 3.2). Here, making SNP sites missing is identical to masking. Finally, we verified that the imputation results were identical with ten repetitions (steps 8 through 12 in section 3.2) at the same separation ratio and at the same masking rate so that we ruled out the possibility that imputation programs could yield different results each time with the same input.

3.5 Measures for imputation results

Statistical measures (step 11 in section 3.2) interpret imputation results from different perspectives and form an essential part of genotype imputation studies. In this section, as alternatives to the commonly used measures such as concordance (section 2.4) and IQS (section 2.5), we define the more discriminating measures negative logarithmic K-L divergence (NLKLD) and negative logarithmic Hellinger distance (NLHD).

3.5.1 Negative logarithmic Kullback-Leibler divergence

Since it can be disproportionately affected by a low MAF, concordance alone is not sufficient to ascertain the overall performance of different imputation programs [36]. In addition, Pearson correlation coefficient is problematic when applied to assess correspondence between imputed and actual genotypes at individual genetic loci (section 2.6). Therefore, measurements that are more suitable need to be introduced for the analysis. The *relative entropy* or *Kullback-Leibler divergence (K-L divergence)* measures the dissimilarity between two probability distributions P and Q, and it is commonly used in statistical inference measurements [29]. Here, P corresponds to the estimated probability distribution of the known genotypes whereas Q corresponds to the estimated probability distribution of the imputed genotypes. In this study, since the organisms we chose are diploid, P and Q both consist of three probabilities, each of which corresponds to one genotype, i.e., AA, Aa, and aa (where A is the major allele and a is the minor allele), and all three probabilitie add up to one. Although we phased SNP data in our study, for this measurement, whether or not the data was correctly phased was not highly relevant. The smaller the K-L divergence is, the better the imputation results are. As given by Cover and Thomas [14],the formula of K-L divergence between two probability mass functions $P(x)$ and $Q(x)$ is as follows:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log_2 (P(x)/Q(x)) \tag{3.1}$$

Here the x's are genotypes of SNP X, $P(x)$ is the probability mass function of the known genotypes, and $Q(x)$ is the probability mass function of the imputed genotypes. The range of $D_{KL}(P||Q)$ is $[0, \infty)$ where 0 means perfect correspondence.

Preliminary work using the K-L divergence to measure agreement between actual and imputed genotypes resulted in very small values (less than 1) and a small range of values. To amplify the range we applied a logarithm function to the raw K-L divergence values. Further, to have the values convey a more intuitive meaning (positive values indicating good performance, negative values indicating poor performance), a unary negation operation was applied. Ultimately, we used *negative logarithmic Kullback-Leibler divergence (NLKLD)* as given in Eq 3.2 to measure the agreement between the estimated probability distributions of the imputed and known genotypes.

$$\begin{aligned} NLD_{KL}(P||Q) &= -\log_{10} D_{KL}(P||Q) \\ &= -\log_{10} \sum_{x \in X} P(x) \log_2(P(x)/Q(x)) \end{aligned} \tag{3.2}$$

According to Eq 3.2, the range of NLKLD is (-∞, ∞) and the bigger the NLKLD value is, the better the imputation results are. Table 3.1 gives some examples of how NLKLD behaves in different situations. As can be seen, the rows where the NLKLD is ∞ show the perfect-correspondence cases whereas the rows where the NLKLD is $-\infty$ show the zero-correspondence cases. In situations where the NLKLD is a negative value (e.g., 5th and 6th rows), the imputation results are worse than random guess. The remaining rows show different situations, and a larger NLKLD value means a better result.

3.5.2 Negative logarithmic Hellinger distance

Although a K-L divergence describes the amount of information loss from letting probability distribution Q approximate P [13], it is an asymmetric measure as $D_{KL}(P||Q)$ does not always equal to $D_{KL}(Q||P)$, and hence can not be used as a true metric [29]. In light of this issue, we applied Hellinger distance to quantify the similarity between two probability distributions [30]. By definition, a Hellinger distance between two discrete probability distributions P and Q is formulated as follows:

$$H(P,Q) = \sqrt{\sum_{x \in X} (\sqrt{P(x)} - \sqrt{Q(x)})^2/2} \tag{3.3}$$

(subsection 3.5.1 for the meaning of P and Q). As can be seen from the above equation, the range of $H(P,Q)$ is [0, 1], where 0 means perfect correspondence and 1 means no correspondence.

Similar to the calculation of K-L divergence, in preliminary work the Hellinger distances calculated from the above equation were small values (much less than 1). Hence we used the negative logarithmic values of the Hellinger distances as a similarity measure between the probability distributions of the known and imputed SNP sites. Thus, we define the *negative logarithmic Hellinger distance (NLHD)* as follows:

$$\begin{aligned} NLHD(P,Q) &= -\log_{10} H(P,Q) \\ &= -\log_{10} \sqrt{\sum_{x \in X} (\sqrt{P(x)} - \sqrt{Q(x)})^2/2} \end{aligned} \tag{3.4}$$

Given Eq 3.4, the range of NLHD is [0, ∞), where 0 means zero correspondence. Table 3.1 gives some examples of how NLHD behaves in different situations. As can be seen, the rows where NLHD is ∞ show the perfect-correspondence cases whereas the rows where the NLHD is 0 show the zero-correspondence cases. The remaining rows show different situations where correspondences are in an ascending order, and a larger NLHD value means a better result.

3.6 Analysis of imputation results

Since low-MAF SNP sites are commonplace in genotypic data, high concordance of the imputed SNP sites with low MAFs may not be appropriate to determine the performance of an imputation program. For example, if more than 90% of the missing SNP sites in the target data have MAFs less than 1% in the reference data, an imputation program may assign major alleles to all targets. In this case, the concordance of the imputed results could easily reach 99% for these low-MAF SNP sites. And since the majority of SNP sites have low MAFs the overall concordance wouldn't be low.

During our analysis of the imputation results, we did not consider results from the SNP sites with MAFs less than 0.5% for the following reasons:

1. numerous studies have already reported that SNP sites below the 0.5% MAF threshold are poorly imputable;

2. SNP sites with MAFs less than 0.5% do not constitute the majority of the results but could potentially affect the whole analysis process; and finally,

3. genetic studies (e.g., GWAS and genomic selection) tend to filter out such SNP sites, and these rare SNP sites tend to be investigated differently.

Further, we binned our results based on the logarithm-base-two of their MAFs. This scale setting was determined based on our observation that MAF distribution almost followed a negative exponential function in the range of [0.005, 0.5]. Hence using a logarithm-base-two function to bin results could potentially allow the results to be evenly distributed. Such a method is effective and can be seen from the actual MAF frequencies as in Figures 3.2 through 3.5, where the y-axes show the percentage of MAFs falling in particular bins.

Table 3.1: Simulation of NLKLD and NLHD for different imputation results.

G_{known}			$G_{imputed}$			NLKLD	NLHD
AA	Aa	aa	AA	Aa	aa		
P_{known}			$P_{imputed}$				
0.0	0.0	1.0	0.0	0.0	1.0	∞	∞
0.0	0.0	1.0	0.0	1.0	0.0	$-\infty$	0
0.0	0.0	1.0	1.0	0.0	0.0	$-\infty$	0
0.1	0.1	0.8	0.1	0.1	0.8	∞	∞
0.1	0.1	0.8	0.1	0.8	0.1	-0.322	0.238
0.1	0.1	0.8	0.8	0.1	0.1	-0.322	0.238
0.15	0.15	0.70	0.15	0.15	0.70	∞	∞
0.15	0.15	0.70	0.15	0.70	0.15	-0.087	0.347
0.15	0.15	0.70	0.70	0.15	0.15	-0.087	0.347
0.2	0.2	0.6	0.2	0.2	0.6	∞	∞
0.2	0.2	0.6	0.2	0.6	0.2	0.198	0.485
0.2	0.2	0.6	0.6	0.2	0.2	0.198	0.485
0.25	0.25	0.50	0.25	0.25	0.50	∞	∞
0.25	0.25	0.50	0.25	0.50	0.25	0.602	0.684
0.25	0.25	0.50	0.50	0.25	0.25	0.602	0.684
0.3	0.3	0.4	0.3	0.3	0.4	∞	∞
0.3	0.3	0.4	0.3	0.4	0.3	1.382	1.072
0.3	0.3	0.4	0.4	0.3	0.3	1.382	1.072
0.33	0.33	0.34	0.33	0.33	0.34	∞	∞
0.33	0.33	0.34	0.33	0.34	0.33	3.366	2.064
0.33	0.33	0.34	0.34	0.33	0.33	3.366	2.064

A represents the major allele and a represents the minor allele. G_{known} and $G_{imputed}$: the known and imputed genotypes; P_{known} and $P_{imputed}$: probability of the known and imputed genotype. For both NLKLD and NLHD, since this study focuses on performance of imputation programs instead of phasing accuracy, whether or not the SNP data is correctly phased is not highly relevant; i.e., Aa and aA are treated the same.

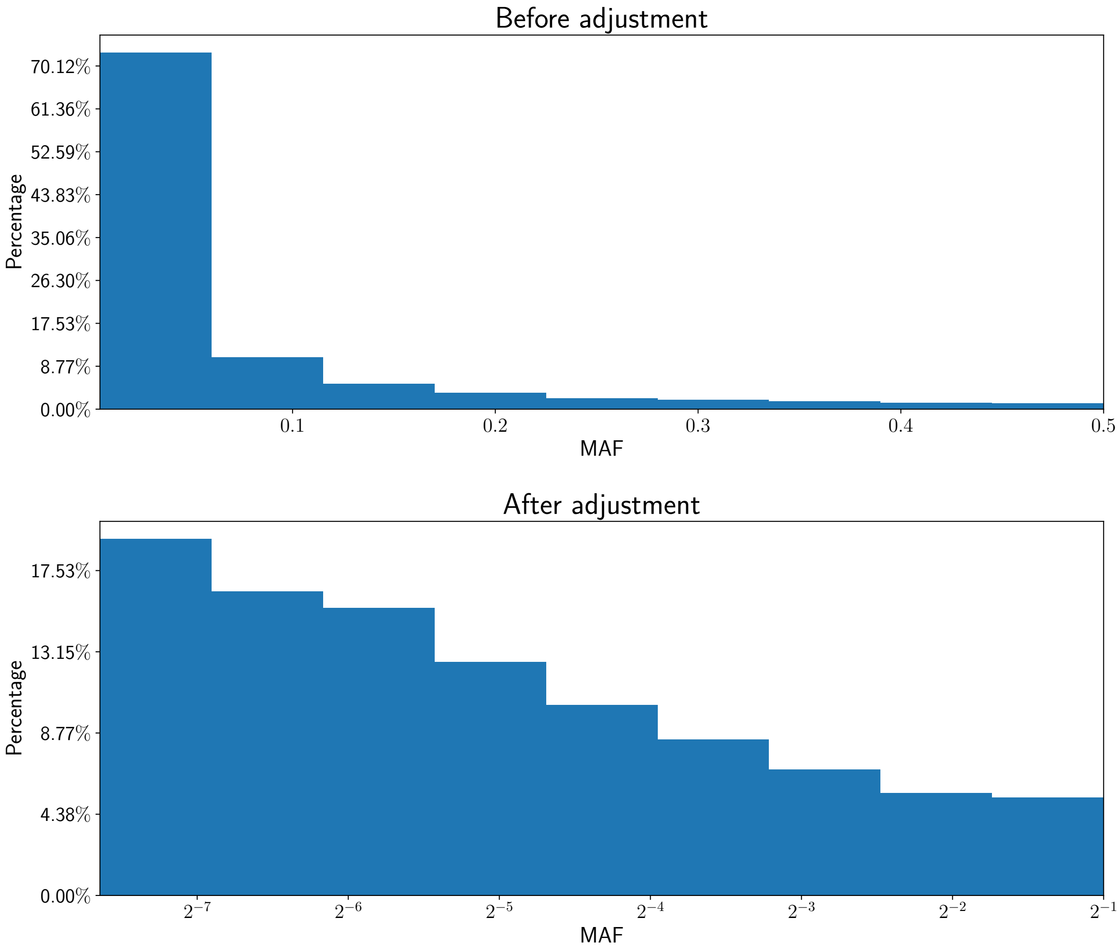

Figure 3.2: MAF distributions using linear-spaced and logarithm-base-two-spaced bins on the *A. thaliana* chromosome 4 dataset. The y-axes show the percentages of MAFs falling in particular bins. The upper figure shows the MAF distribution throughout the range [0.005, 0.5] in linear-spaced bins. Each bin has the same width. The lower figure shows the MAF distribution throughout the same range in logarithm-base-two-spaced bins.

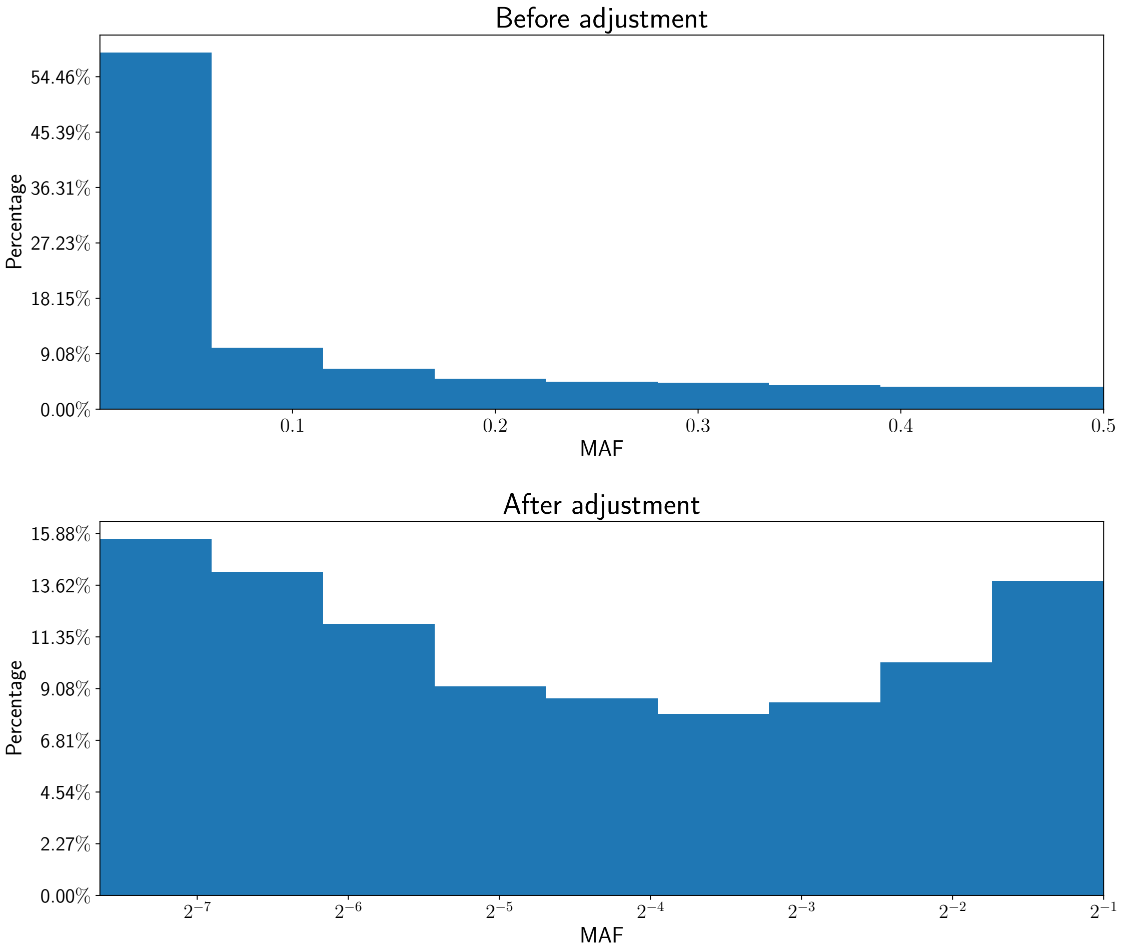

Figure 3.3: MAF distributions using linear-spaced and logarithm-base-two-spaced bins on the human chromosome 22 dataset. The y-axes show the percentages of MAFs falling in particular bins. The upper figure shows the MAF distribution throughout the range [0.005, 0.5] in linear-spaced bins. Each bin has the same width. The lower figure shows the MAF distribution throughout the same range in logarithm-base-two-spaced bins.

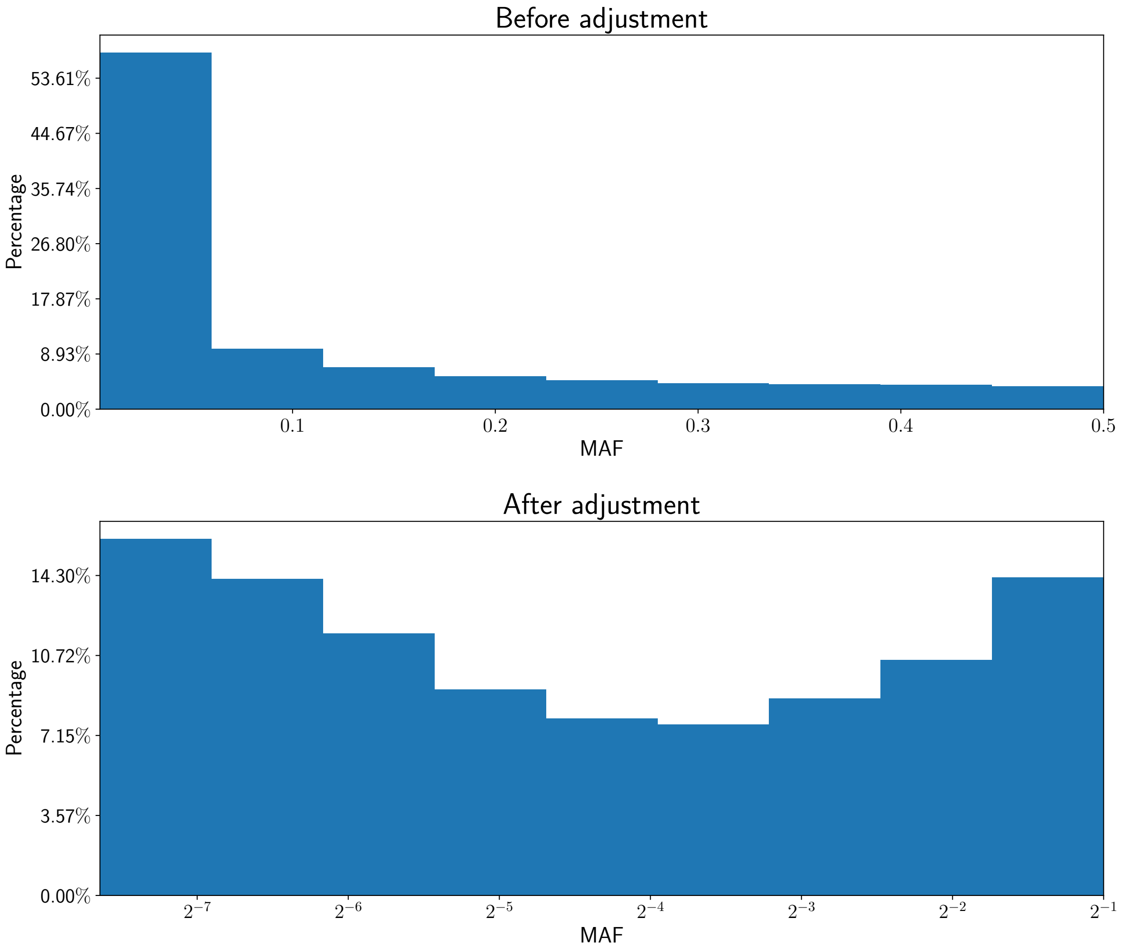

Figure 3.4: MAF distributions using linear-spaced and logarithm-base-two-spaced bins on the human chromosome 13 dataset. The y-axes show the percentages of MAFs falling in particular bins. The upper figure shows the MAF distribution throughout the range [0.005, 0.5] in linear-spaced bins. Each bin has the same width. The lower figure shows the MAF distribution throughout the same range in logarithm-base-two-spaced bins.

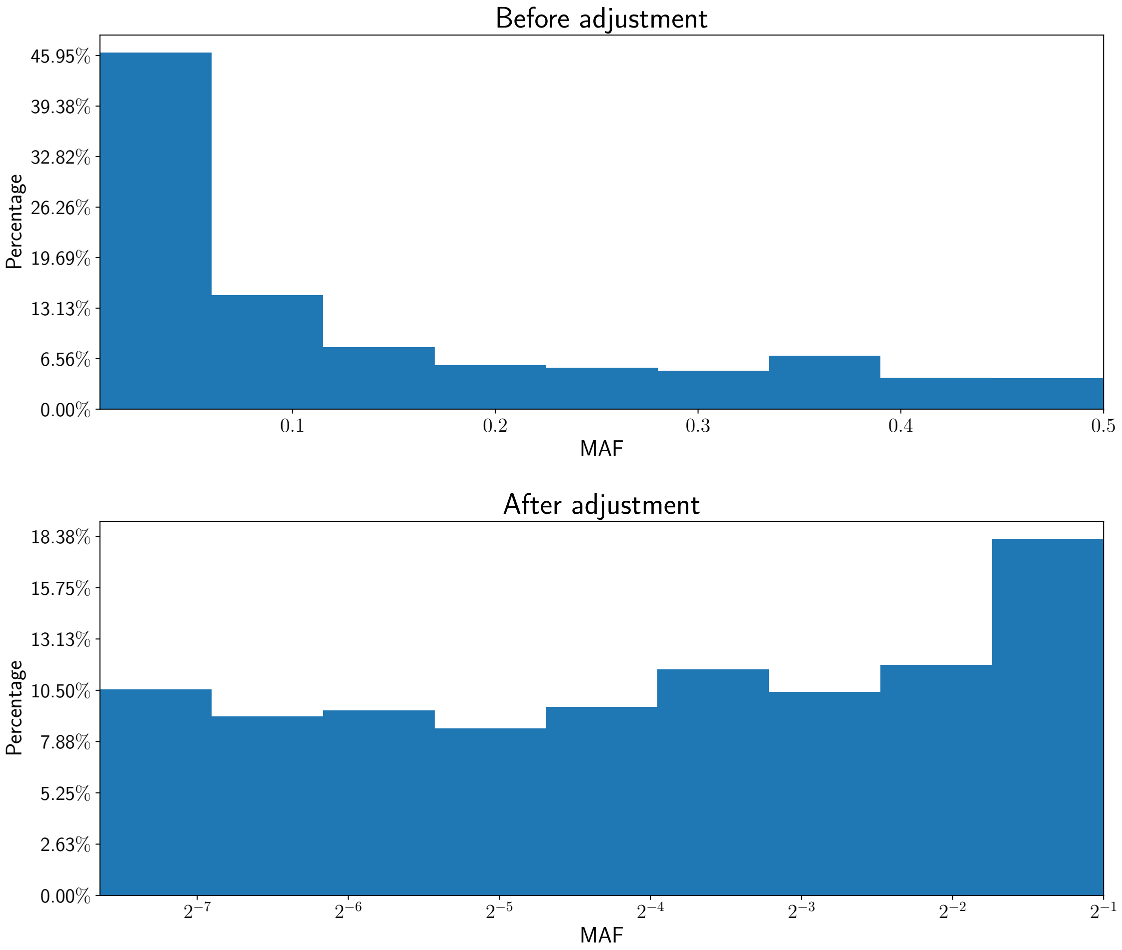

Figure 3.5: MAF distributions using linear-spaced and logarithm-base-two-spaced bins on the rice chromosome 12 dataset. The y-axes show the percentages of MAFs falling in particular bins. The upper figure shows the MAF distribution throughout the range [0.005, 0.5] in linear-spaced bins. Each bin has the same width. The lower figure shows the MAF distribution throughout the same range in logarithm-base-two-spaced bins.

Chapter 4

Results

In this chapter, we present all our results aggregated in two different ways: bar graphs that show the patterns of each measurement against MAFs and scatter plots that provide closer details of the comparisons between the imputation programs, and between plants and humans. To be comparable with the previous imputation studies [11, 15] regarding accuracy measurement, we originally determined both concordance and Pearson correlation coefficient between genotyped and imputed SNP sites. However, these two measurements were insufficient to assess imputation accuracy when the MAFs were low. In light of this, we used IQS (section 2.5) to investigate the correspondence between the genotyped and imputed SNP sites. Further, we calculated a negative logarithmic Kullback-Leibler divergence (NLKLD) (subsection 3.5.1) and a negative logarithmic Hellinger distance (NLHD) (seubsection 3.5.2) to investigate the correspondence between probability distributions of the known and imputed SNP sites. Again, we generated bar graphs and scatter plots where each set of measurement results were compared against each other between two imputation programs. In section 4.2, we present the detailed observations. In sections 4.3 and 4.4, we combine the results from Figures 4.1 through 4.16 and present the implications of the results in terms of differences in imputation performance. Data of results used to generate the graphs are in Appendix .

4.1 Ranking imputation results with NLKLD and NLHD

Table 4.1 demonstrates that in a general case where SNP sites with different MAFs are imputed at the same concordance of 90%, IQS, NLKLD, and NLHD all capture a difference in the imputation performance. Overall, IQS agrees with NLKLD and NLHD except for the first two imputation cases, where both NLKLD and NLHD agree that the second one is the superior imputation result whereas IQS ranks the first one as superior. The second row shows the minimum discrepancy between the known and imputed MAFs and should be considered the supreme result among all rows. Taking a closer look at the IQS and MAF_{known} columns, one may notice that these two are positively correlated. Such a correlation indicates that IQS can be biased in ranking imputation results as it reflects the correspondence between the known and imputed genotypes yet does not reflect the correspondence between the MAFs of the known and imputed SNP sites. In contrast, both NLKLD and NLHD reflect not only this correspondence, but also the chance agreement between the known and imputed genotypes. Thus, NLKLD and NLHD are more appropriate in ranking imputation results.

Table 4.1: Ranking imputation results with the same concordance (90%) using IQS, NLKLD, and NLHD.

G_{known}			$G_{imputed}$			MAF_{known}	$MAF_{imputed}$	IQS	NLKLD	NLHD
AA	Aa	aa	AA	Aa	aa					
P_{known}			$P_{imputed}$							
0.3	0.5	0.2	0.25	0.45	0.3	0.45	0.475	0.844	1.421	0.931
0.5	0.28	0.22	0.44	0.38	0.18	0.36	0.37	0.842	1.488	0.969
0.6	0.3	0.1	0.7	0.2	0.1	0.25	0.2	0.804	1.376	0.928
0.7	0.2	0.1	0.8	0.1	0.1	0.2	0.15	0.756	1.186	0.844
0.8	0.1	0.1	0.75	0.05	0.20	0.15	0.225	0.733	1.128	0.788
0.88	0.07	0.05	0.98	0.02	0.0	0.085	0.01	0.266	0.219	0.584

A represents the major allele and a represents the minor allele. G_{known} and $G_{imputed}$: the known and imputed genotypes; P_{known} and $P_{imputed}$: probability of the known and imputed genotype; MAF_{known} and $MAF_{imputed}$: minor allele frequency of the known and imputed SNP sites.

4.2 Imputation results on all experimental data

In this section, we use bar graphs to present imputation results on all experimental data, i.e., *A. thaliana* chromosome 4, human chromosomes 22 and 13, and rice chromosome 12. Additionally, it is worth noting that some bars are missing from the graphs especially in the MAF range [0.005, 0.0083]. Such a situation happened when there were no imputed SNP sites in such a MAF range at a particular missing rate.

4.2.1 Imputation results on *A. thaliana*

Figures 4.1 through 4.4 show the results from runs on the *A. thaliana* chromosome 4 experimental data.

Concordance In Figure 4.1, concordance drops as MAF increases for both Minimac and Beagle. This means that for the *A. thaliana* data, both programs have difficulty inferring individual genotypes correctly at SNP sites with high MAFs. The Minimac concordance drops more rapidly than that of Beagle. This means that Minimac's percent agreement is more sensitive to MAF than Beagle's.

IQS According to Figure 4.2, the Minimac imputation results have high chance agreement between the known and imputed genotypes given the range of Minimac IQS. Also, the Minimac IQS values show an inconsistent pattern. Some colours (representing different missing rates) are even seemingly missing from the stacked barplot. This is because in the same MAF range, IQS results at different missing rates could have both positive and negative values. Such an inconsistent pattern of IQS values indicates that the program is inconsistent at chance agreement control in terms of different missing rates of SNPs for the *A. thaliana* data. For instance, at separation ratio 8:2, IQS values in MAF ranges other than [0.005, 0.0083] are positive so all bars are above the $y = 0$ line. In contrast, at the same separation ratio, all IQS values in the MAF range [0.005, 0.0083] are negative so that all bars for that range are below the $y = 0$ line. Meanwhile, at separation ratio 9:1, in the MAF range [0.039, 0.065], IQS at missing rate 10% is negative and all other IQS values in the same MAF range are positive. In this case, bars with positive IQS values are stacked on top of the bar with a negative IQS value so that the red bar looks "missing" from the graph. In addition, the Minimac IQS does not increase with MAF whereas the Beagle IQS does. Such an observation indicates that Beagle's chance agreement between the known and imputed SNPs drops as MAF increases, whereas Minimac is more likely to infer SNPs by accident.

NLKLD In Figure 4.3, both programs' NLKLD values generally increase with MAF. This means that the probability distribution of the imputed SNPs is reflective of the probability distribution of the known SNPs as MAF increases for the *A. thaliana* data. Both programs have superior performance on data at sample size ratios between reference and target 5:5, 6:4, and 7:3.

NLHD Figure 4.4 displays similar patterns as in Figure 4.3. This indicates that as MAF increases, probability distributions of the known and imputed SNPs become closer for the *A. thaliana* data. However, the NLHD results have less variance as compared to the NLKLD ones. Also, that both programs perform better on data at sample size ratios between reference and target 5:5, 6:4, and 7:3 is less distinguishable.

4.2.2 Imputation results on human

Figures 4.5 through 4.12 show the results from imputation runs on human data.

Imputation results on human chromosome 22

Figures 4.5 through 4.8 show the results from the Minimac and Beagle imputation runs on the human chromosome 22 experimental data.

Concordance Figure 4.5 manifests almost the identical pattern as in Figure 4.1, where concordance drops as MAF increases. Again, this means that both programs have difficulty inferring individual genotypes correctly at SNP sites with high MAFs for the human chromosome 22 data.

IQS In Figure 4.6, Minimac has an erratic IQS pattern similar to the one in Figure 4.2 whereas the Beagle IQS overall increases with MAF. Likewise, this inconsistent pattern of IQS values indicates that the program is inconsistent at chance agreement control in terms of different missing rates of SNPs in the human chromosome 22 data. Again, some coloured bars seem to be missing because the “missing” bars correspond to negative values and they are covered by bars that correspond to positive values.

NLKLD In Figure 4.7, both Minimac and Beagle show an erratic pattern in their NLKLD results. This indicates that the probability distribution of the imputed SNPs is not reflective of the probability distribution of the known SNPs as MAF increases for the human chromosome 22 data. More interestingly, neither program shows superior results on SNP sites with high MAFs between 0.18 and 0.5, as compared to the results on *A. thaliana*. On the contrary, imputation results within such a range are the poorest.

NLHD In Figure 4.8, similar to the pattern in Figure 4.7, the NLHD results from both programs do not show strong correlation with MAF. This means that probability distributions of the imputed and known SNPs do not become closer as MAF increases for the human chromosome 22 data. However, imputation results with MAF between 0.18 and 0.5 are not evidently the poorest. Moreover, the NLHD results are closer to each other as opposed to the NLKLD results.

Imputation results on human chromosome 13

Figures 4.9 through 4.12 show the results from imputation results on the human chromosome 13 experimental data. In Figures 4.9, 4.10, and 4.12, both Minimac and Beagle show almost the identical patterns as in Figures 4.5, 4.6 and 4.8 respectively, for the concordance, IQS, and NLHD results. However, for the NLKLD results as in Figure 4.11, both programs show an overall decreasing pattern with MAF. Such an observation indicates that the probability distribution of the imputed SNPs becomes inconsistent with the probability distribution of the known SNPs as MAF increases for the human chromosome 13 data.

4.2.3 Imputation results on rice

Figures 4.13 through 4.16 show the results from imputation runs on the rice experimental data. In Figures 4.13 and 4.14, both programs' concordance and IQS values show little difference as in results from other experimental data. This means that both programs have difficulty correctly imputing SNPs with high MAFs for the rice data. Also, Minimac performs poorly in controlling chance agreement whereas Beagle's chance agreement drops as MAF increases. In Figure 4.15, Minimac shows an overall increasing pattern in its NLKLD results, meaning that Minimac's probability distribution of the imputed SNPs becomes reflective of the probability distribution of the known SNPs for the rice data. In addition, Beagle shows an overall decreasing pattern in NLKLD, meaning that Beagle's probability distribution of the imputed SNPs becomes inconsistent with the probability distribution of the known SNPs for the rice data. In Figure 4.16, Minimac shows an inconsistent pattern in its NLHD results, meaning that the probability distribution of Minimac's imputed SNPs do not strongly correlate to the probability distribution of the known SNPs for the rice data. In addition, Beagle shows an overall decreasing pattern in both NLKLD and NLHD. This means that for the rice data, Beagle's probability distribution of the imputed SNPs becomes dissimilar from the probability distribution of the known SNPs as MAF increases, and that Beagle's probability distribution of the imputed SNPs becomes more distant from the probability distribution of the known SNPs as MAF increases.

4.3 Comparisons of imputation results between plant and human

Figures 4.17 through 4.20 show the comparisons of imputation results between plant and human. Since this section mainly compares performance of imputation programs between plant and human, how the performance of imputation programs interacts with MAF, missing rates, and separation ratios between reference and target is not addressed. Additionally, in one ideal case where one organism has superior results over the other, in the scatterplots we should see that all dots are on one side of the diagonal lines. In another ideal case where both organisms have equivalent results, we should see that all dots are on the diagonal lines.

Concordance Figure 4.17 combines the results shown in Figures 4.1, 4.5, 4.9, and 4.13. In Figure 4.17, Minimac shows superior performance on plant over human whereas Beagle shows the opposite overall. Interestingly, at separation ratio 9:1 with a 30% missing rate, Beagle shows a superior performance on *A. thaliana* over human.

IQS Figure 4.18 combines the results shown in Figures 4.2, 4.6, 4.10, and 4.14. In Figure 4.18, data points are equally distributed in the Minimac column for both organisms. This indicates that the chance agreement of the Minimac imputation results is almost insensitive to the imputation results of either data. In contrast, almost all data points are on the human side of the Beagle column, meaning that Beagle has lower chance agreement in its human imputation results than in its plant imputation results.

NLKLD Figure 4.19 combines the results shown in Figures 4.3, 4.7, 4.11, and 4.15. In Figure 4.19, the upper row shows that both Minimac and Beagle have superior results on *A. thaliana* over human. Meanwhile, the lower row shows that Minimac performs equally well on rice and human whereas Beagle has superior performance on human over rice.

NLHD Figure 4.20 combines the results shown in Figures 4.4, 4.8, 4.12, and 4.16. In Figure 4.20, the upper row shows that Beagle performs slightly better on human over *A. thaliana* whereas Minimac performs equally well on both datasets. In contrast, the lower row shows that both Minimac and Beagle have superior performance on human over rice.

4.4 Comparisons of imputation results between Minimac and Beagle

Figures 4.21 through 4.24 show the comparisons of imputation results between Minimac and Beagle with varying missing rates and sample size ratios between reference and target. In one ideal case where one program has superior performance over the other, in the scatterplots we should see that all dots are on one side of the diagonal lines and in the bar-plots we should see that all bars are either above or below $y = 0$. In another ideal case where both programs have the same performance, in the scatterplots we should see that all dots are on the diagonal lines and the bar-plots should be empty.

Concordance Figure 4.21 combines the results shown in Figures 4.1, 4.5, 4.9, and 4.13. In Figure 4.21, Beagle shows superior performance over Minimac in terms of concordance. According to the bar-plots, Beagle's superiority over Minimac increases with MAF.

IQS Figure 4.22 combines the results shown in Figures 4.2, 4.6, 4.10, and 4.14. Figure 4.22 shows the similar pattern as in Figure 4.21. However, the Beagle IQS shows more overwhelming superiority over Minimac especially on imputation results with lower MAFs (0.005 – 0.11). This means that Beagle has a much superior control at chance agreement over Minimac.

NLKLD Figure 4.23 combines the results shown in Figures 4.3, 4.7, 4.11, and 4.15. In Figure 4.23, Minimac shows an overall superior performance over Beagle according to NLKLD. This means that Minimac has effective probability distributions of the imputed SNPs to represent the probability of the known SNPs. However, $NLKLD_{Minimac-Beagle}$ increases with MAF only on the rice data and does not have the same consistent pattern on other data. Such a difference means that for the rice data, Minimac's probability distribution of the imputed SNPs is increasingly reflective of the probability distribution of the known SNPs over Beagle's as MAF increases.

NLHD Figure 4.24 combines the results shown in Figures 4.4, 4.8, 4.12, and 4.16. Figure 4.24 shows that NLHD overall agrees with NLKLD as in Figure 4.23. However, as can be seen in the bar-plots, Minimac has more overwhelming superiority over Beagle as the bars are higher above the y=0 line as compared to the ones in Figure 4.23. This means that Minimac's probability distribution of the imputed SNPs is closer to the probability distribution of the known SNPs than Beagle's.

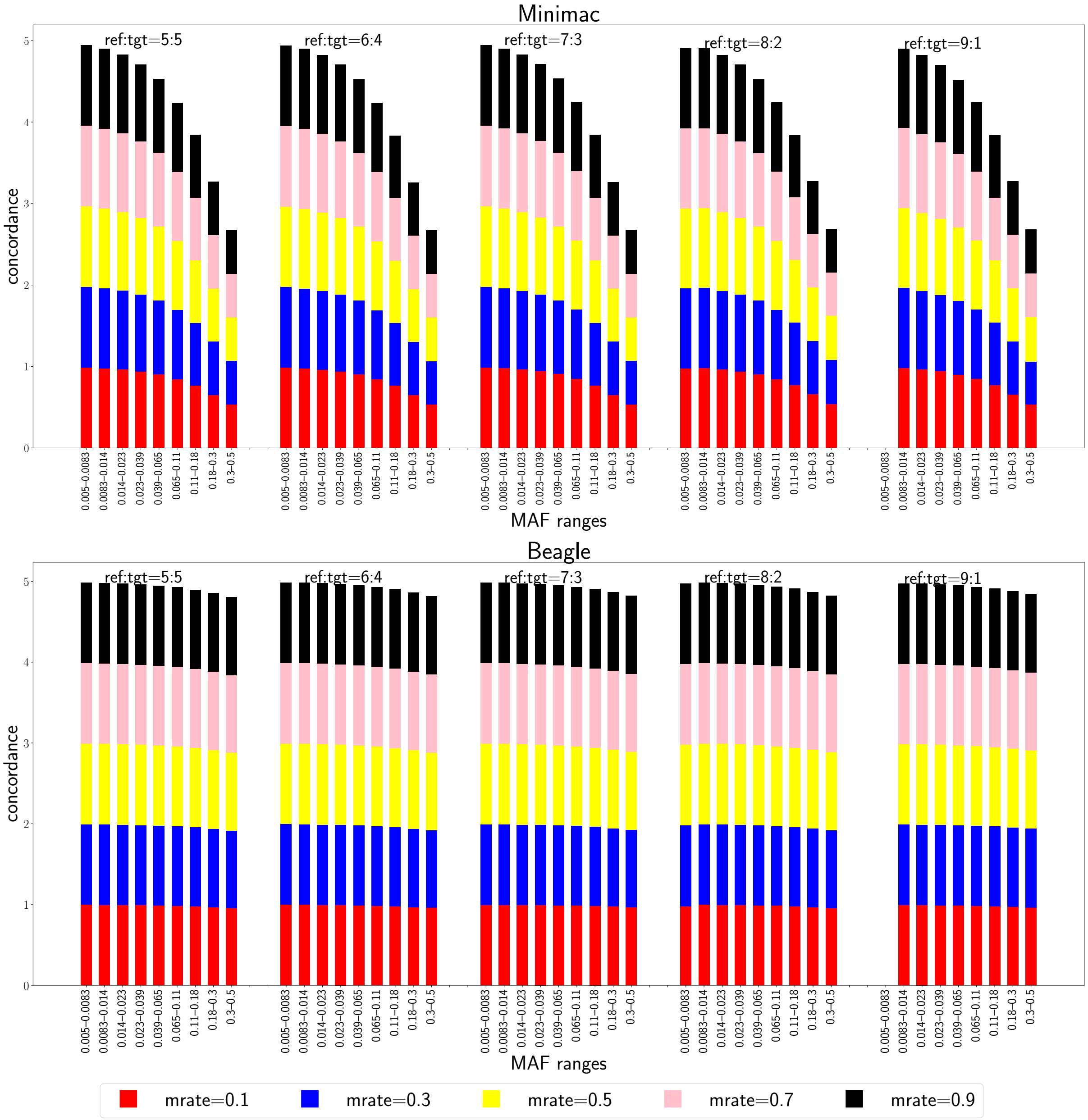

Figure 4.1: Concordance results for the *Arabidopsis thaliana* chromosome 4 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for both programs: [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

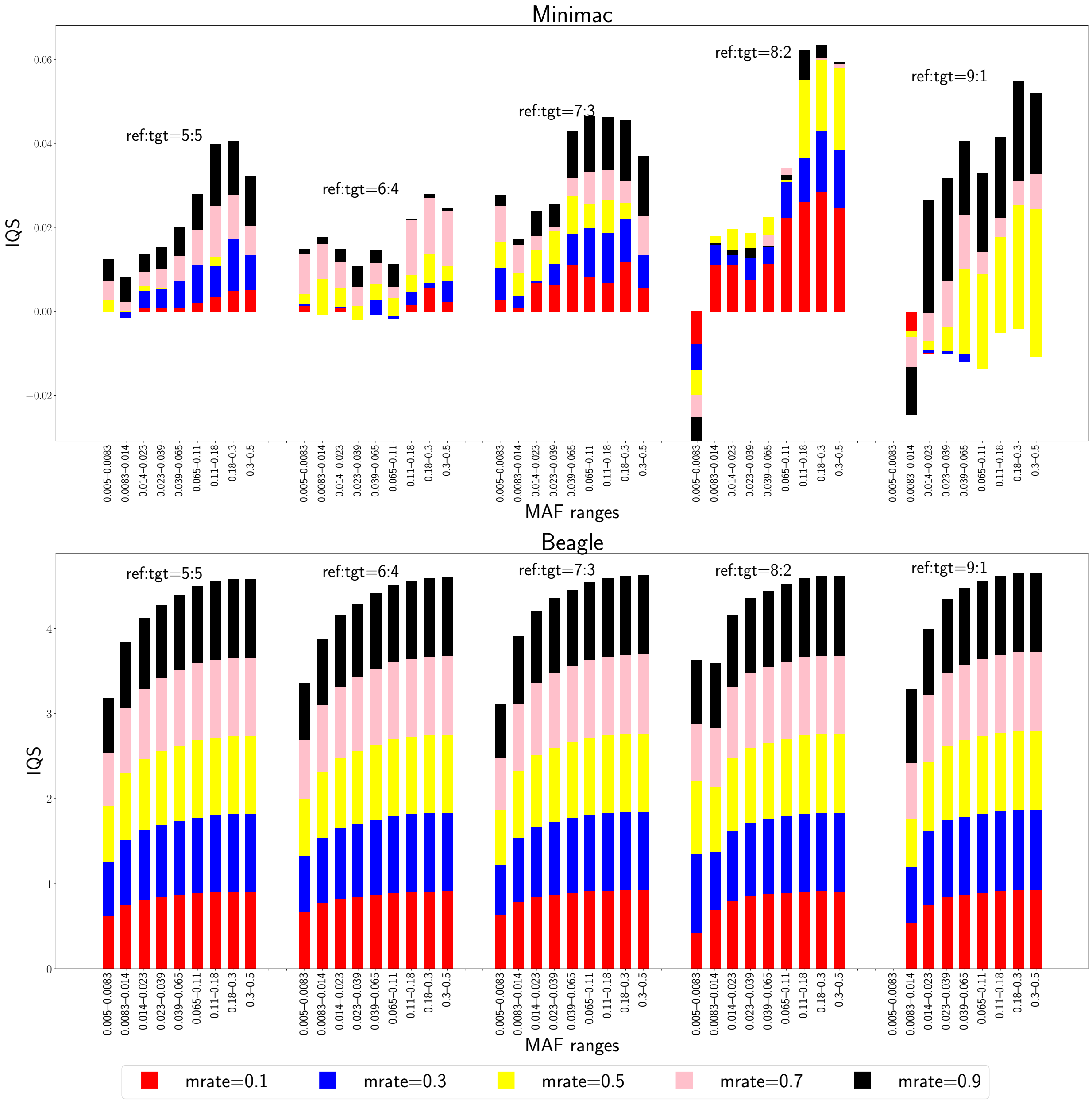

Figure 4.2: IQS results for the *A. thaliana* chromosome 4 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [-0.02, 0.06]; y-axis range for Beagle (lower row): [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

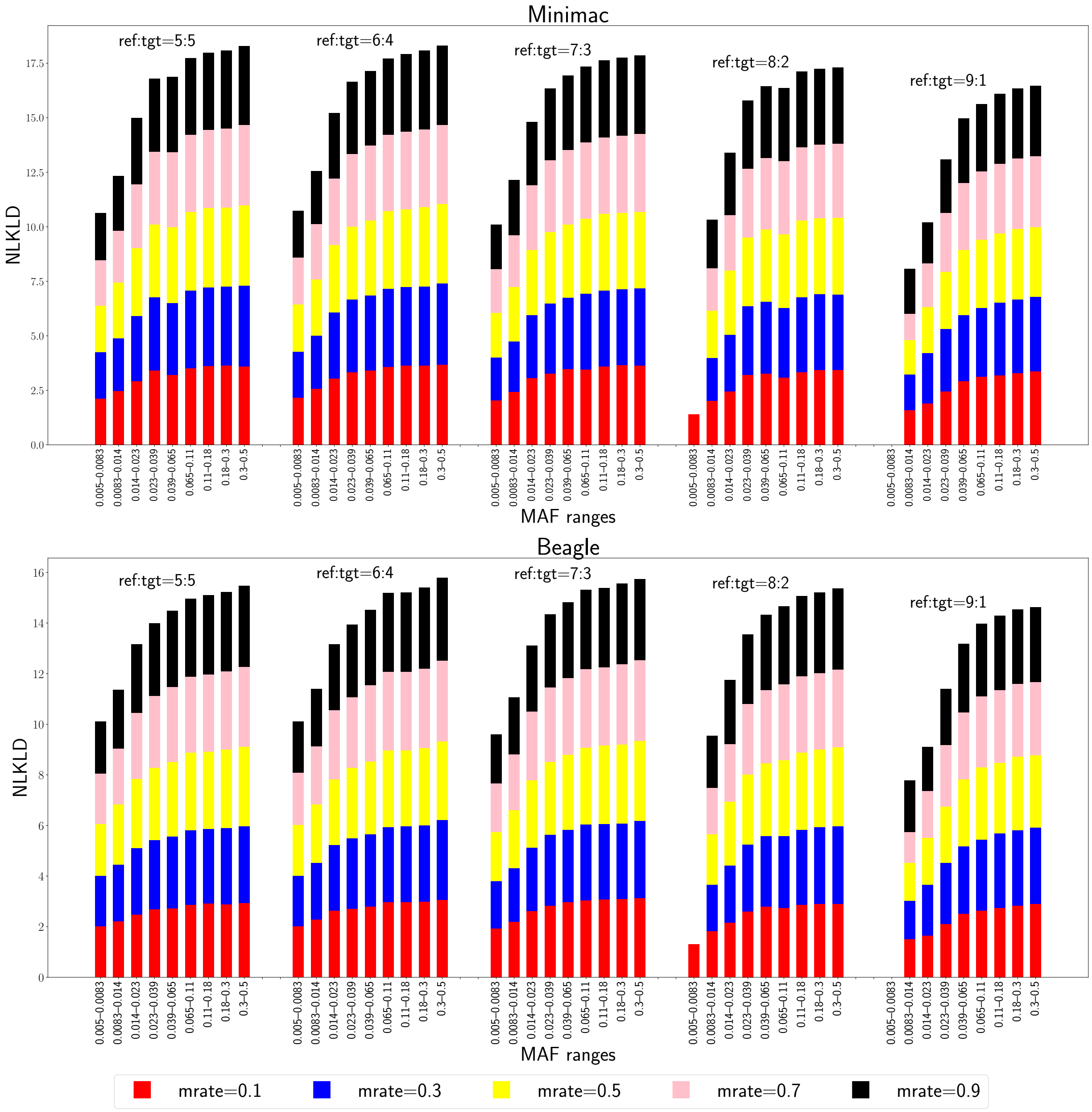

Figure 4.3: NLKLD results for the *A. thaliana* chromosome 4 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 17.5]; y-axis range for Beagle (lower row): [0, 16]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

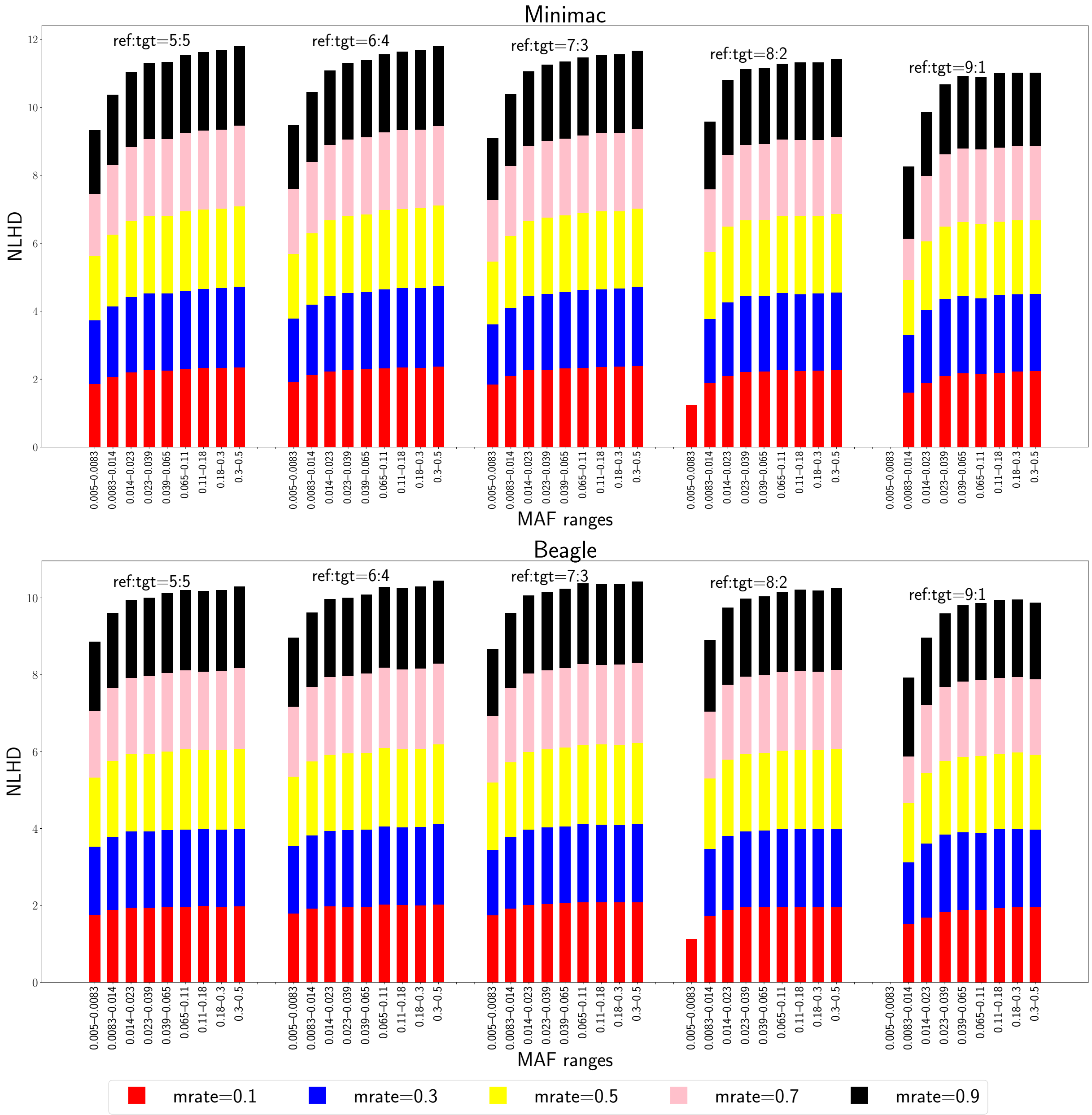

Figure 4.4: NLHD results for the *A. thaliana* chromosome 4 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 12]; y-axis range for Beagle (lower row): [0, 10]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

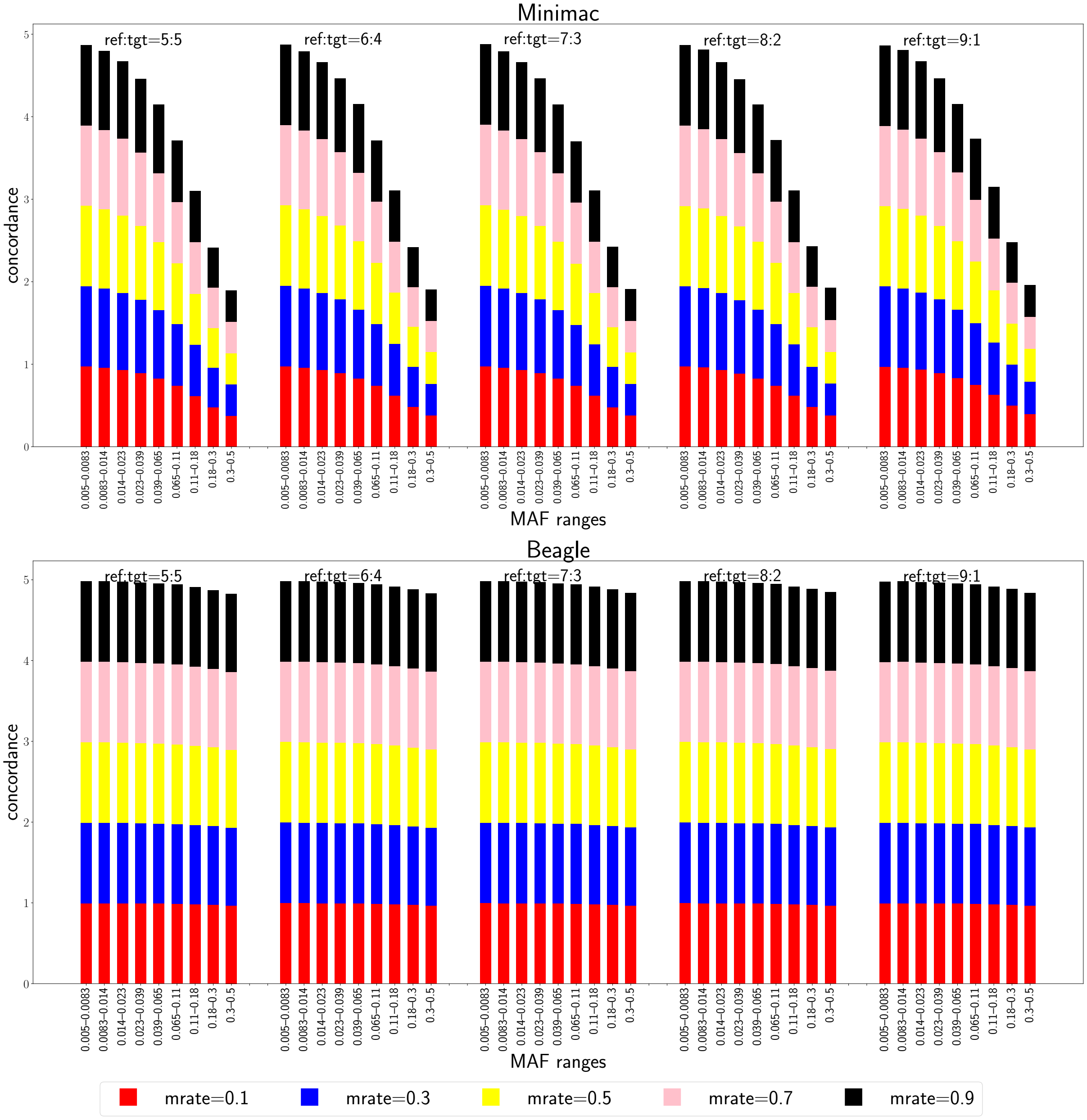

Figure 4.5: Concordance results for the human chromosome 22 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for both programs: [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

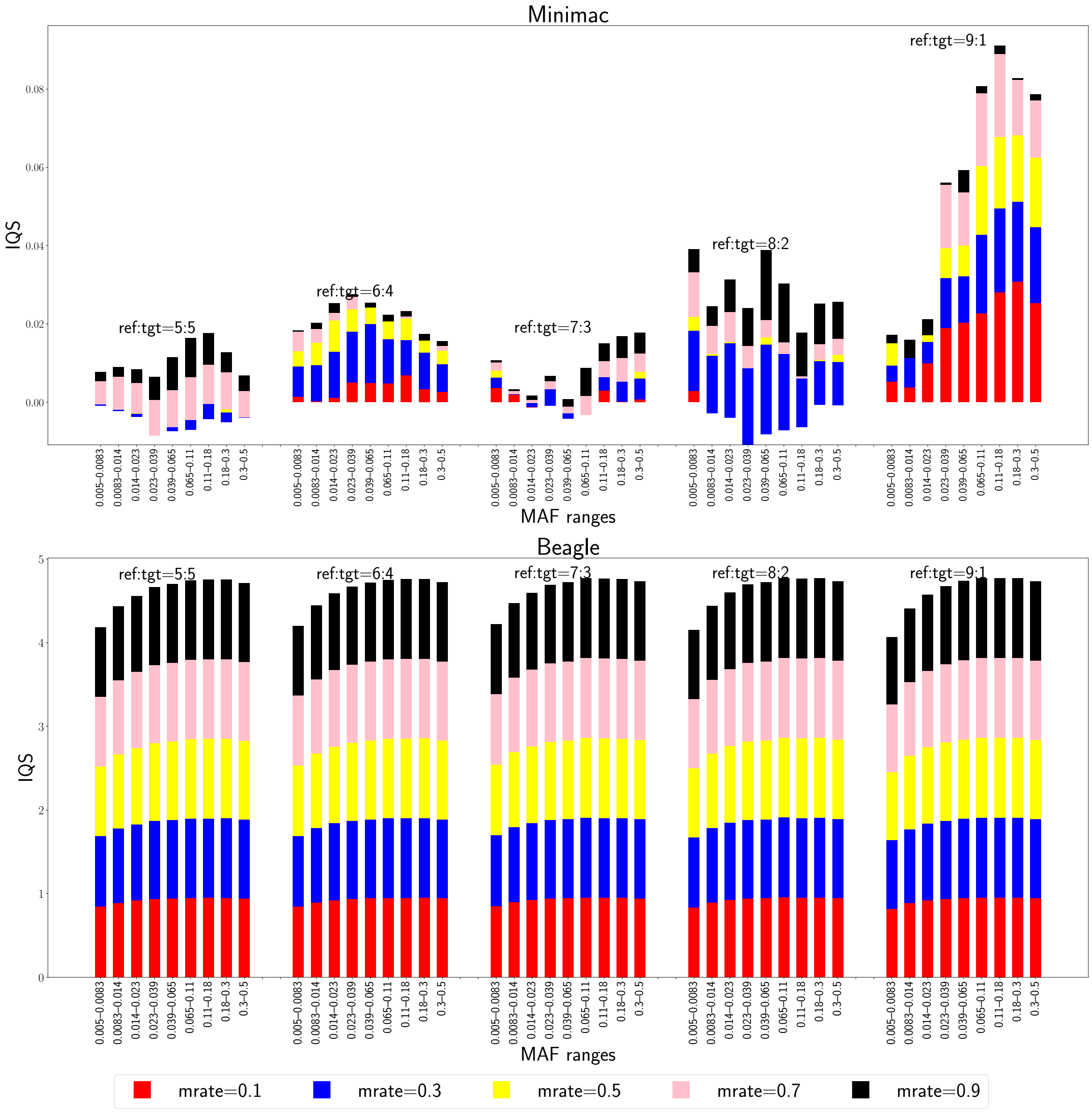

Figure 4.6: IQS results for the human chromosome 22 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 0.08]; y-axis range for Beagle (lower row): [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

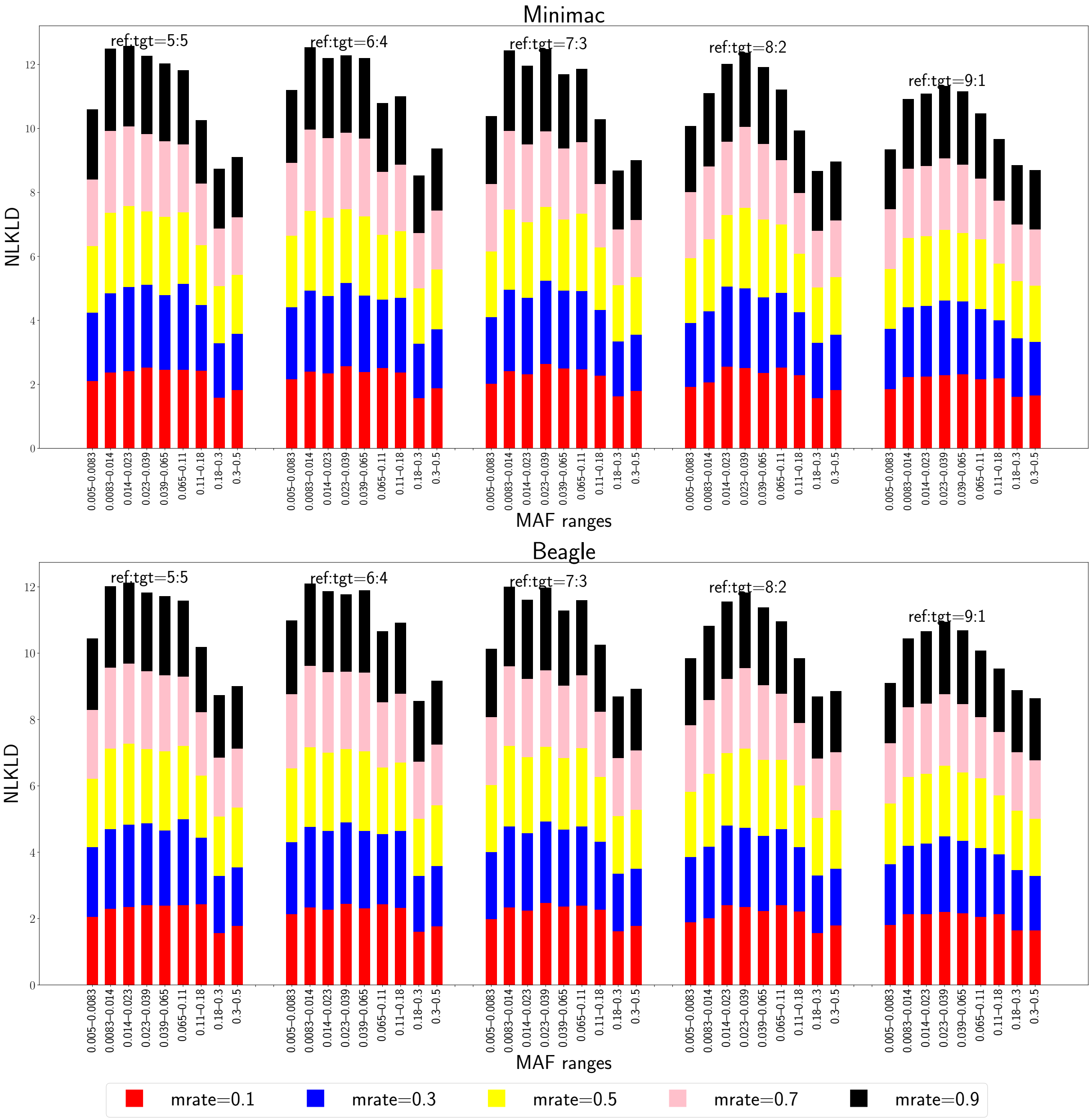

Figure 4.7: NLKLD results for the human chromosome 22 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 12]; y-axis range for Beagle (lower row): [0, 12]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

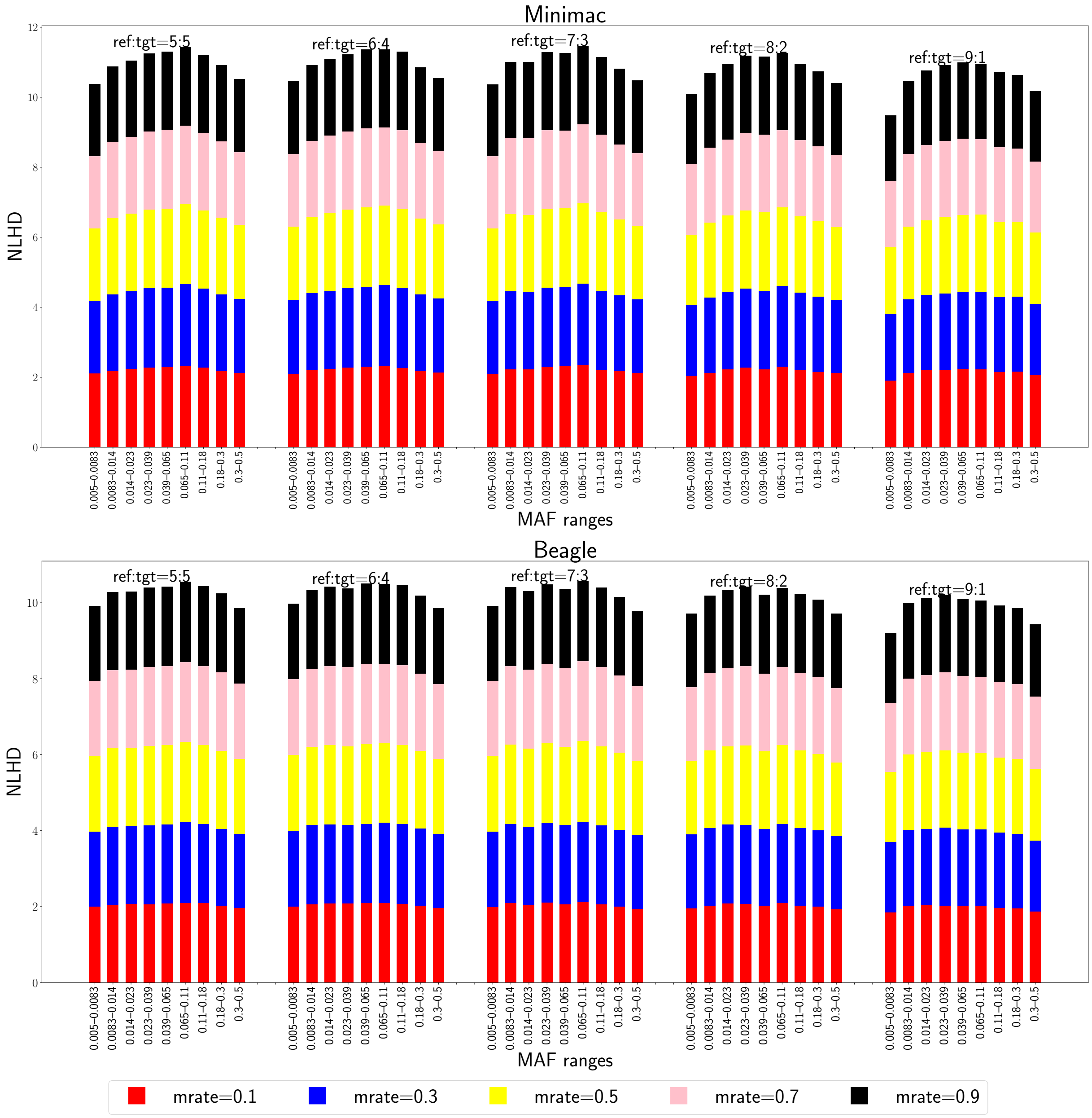

Figure 4.8: NLHD results for the human chromosome 22 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 12]; y-axis range for Beagle (lower row): [0, 10]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

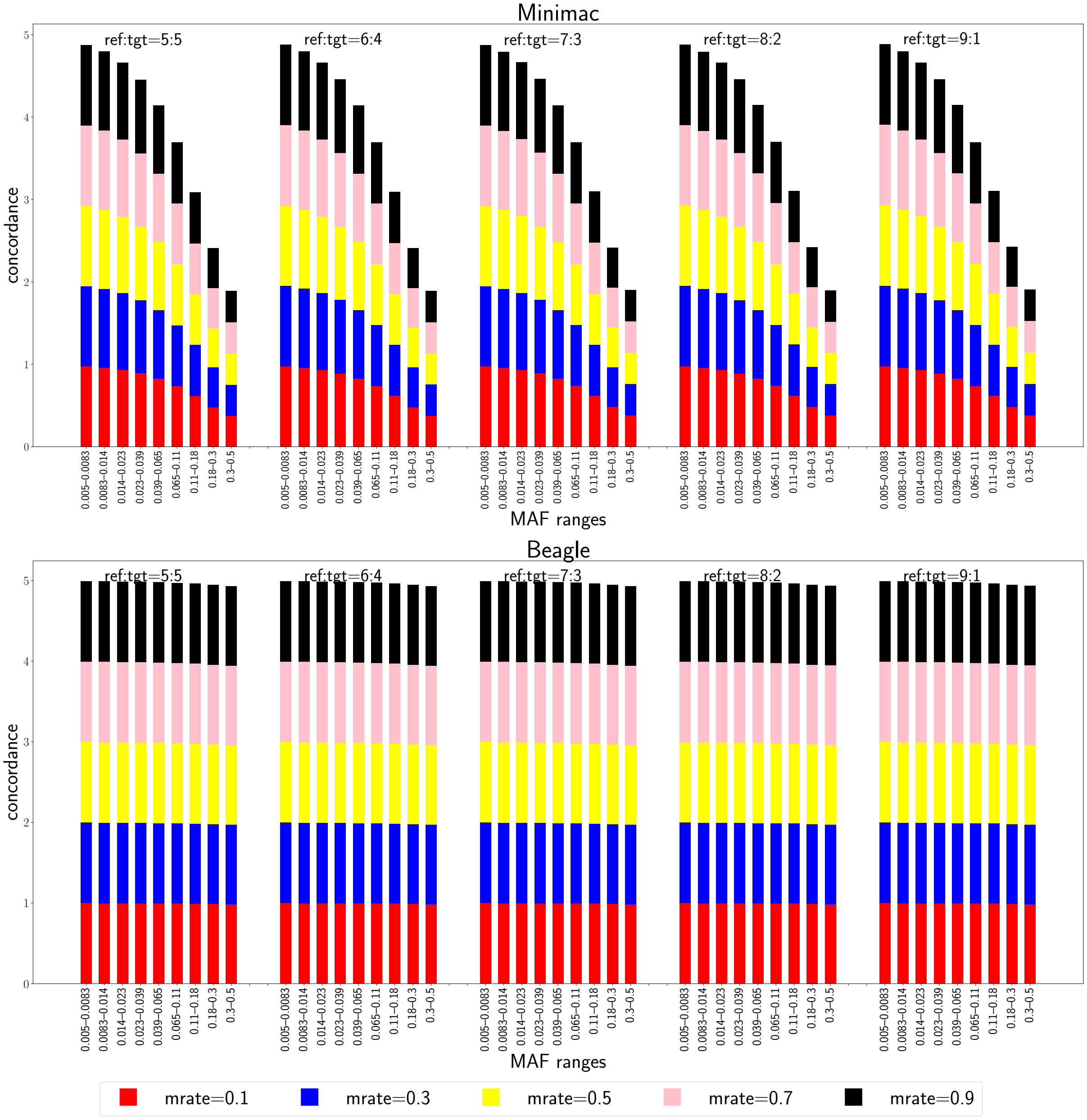

Figure 4.9: Concordance results for the human chromosome 13 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for both programs: [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

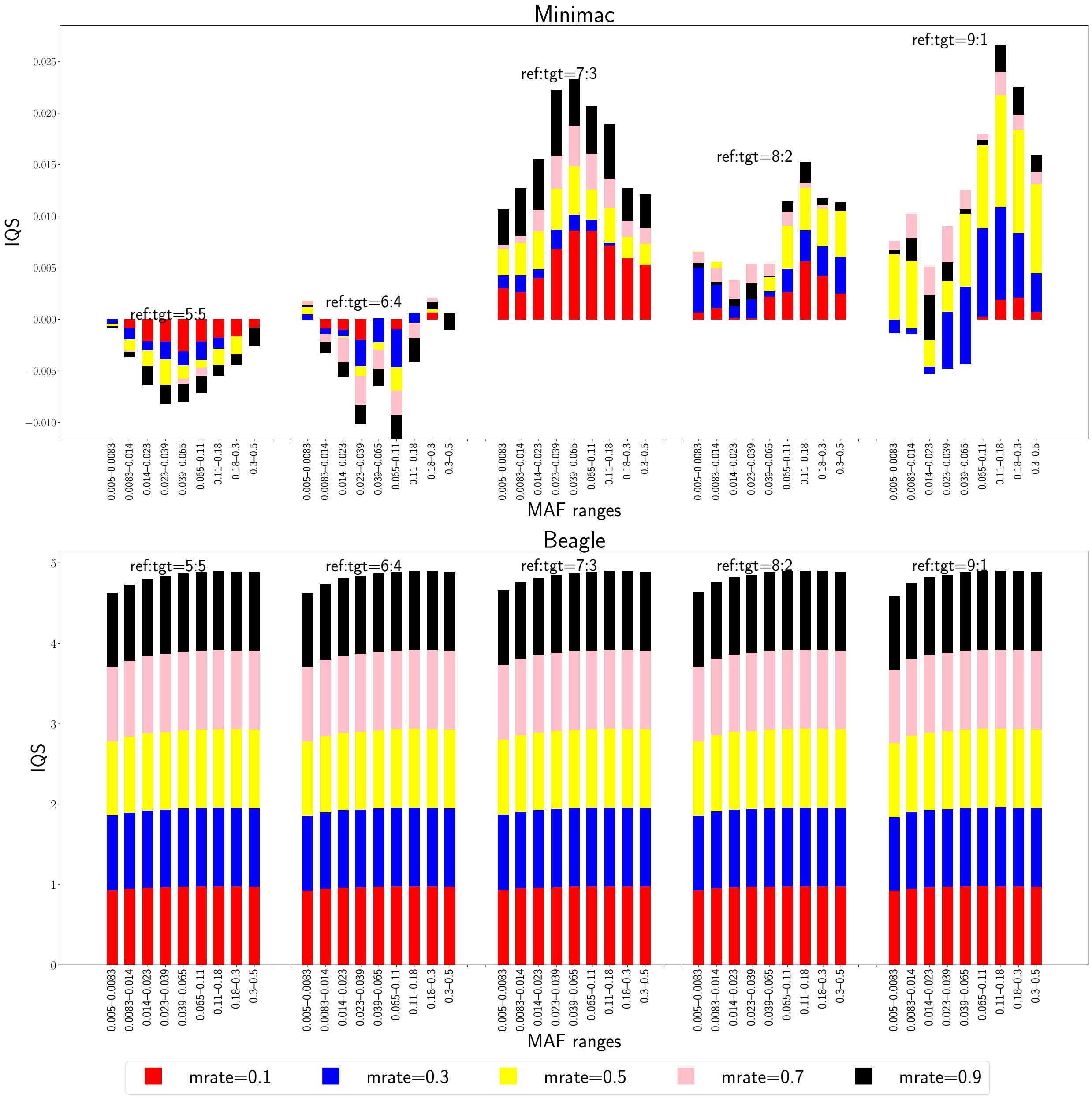

Figure 4.10: IQS results for the human chromosome 13 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [-0.010, 0.025]; y-axis range for Beagle (lower row): [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

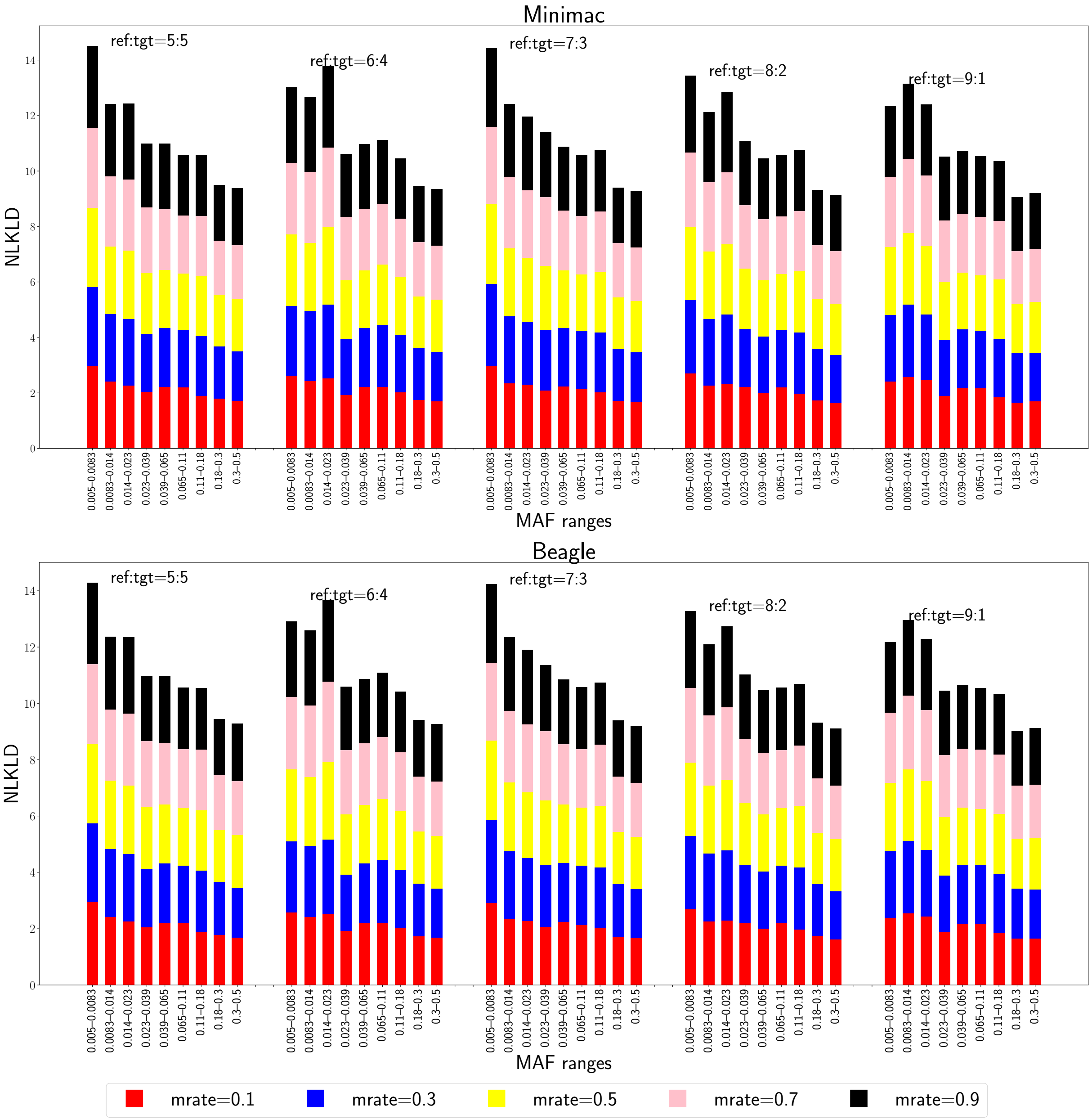

Figure 4.11: NLKLD results for the human chromosome 13 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 14]; y-axis range for Beagle (lower row): [0, 14]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

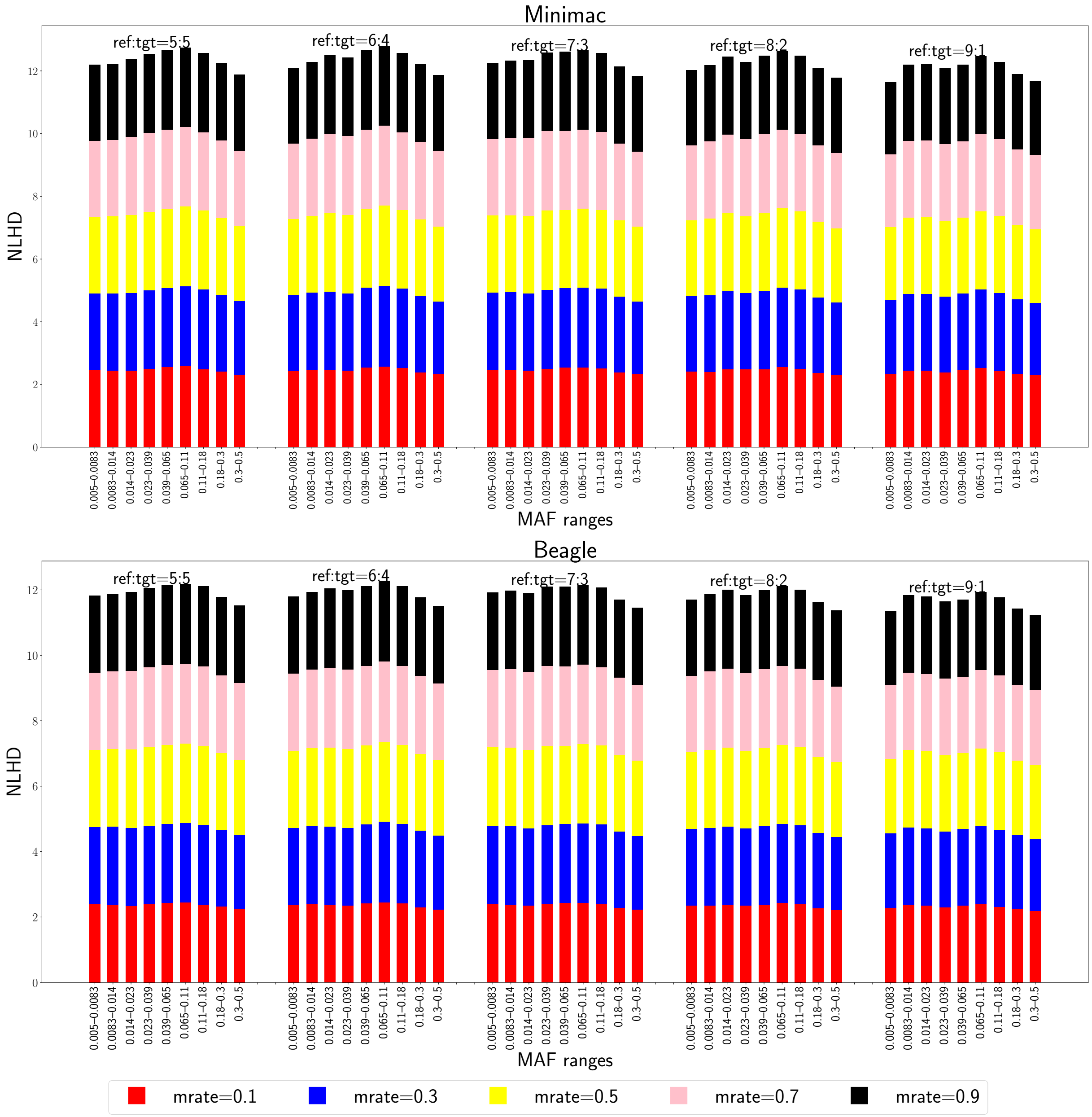

Figure 4.12: NLHD results for the human chromosome 13 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 12]; y-axis range for Beagle (lower row): [0, 12]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

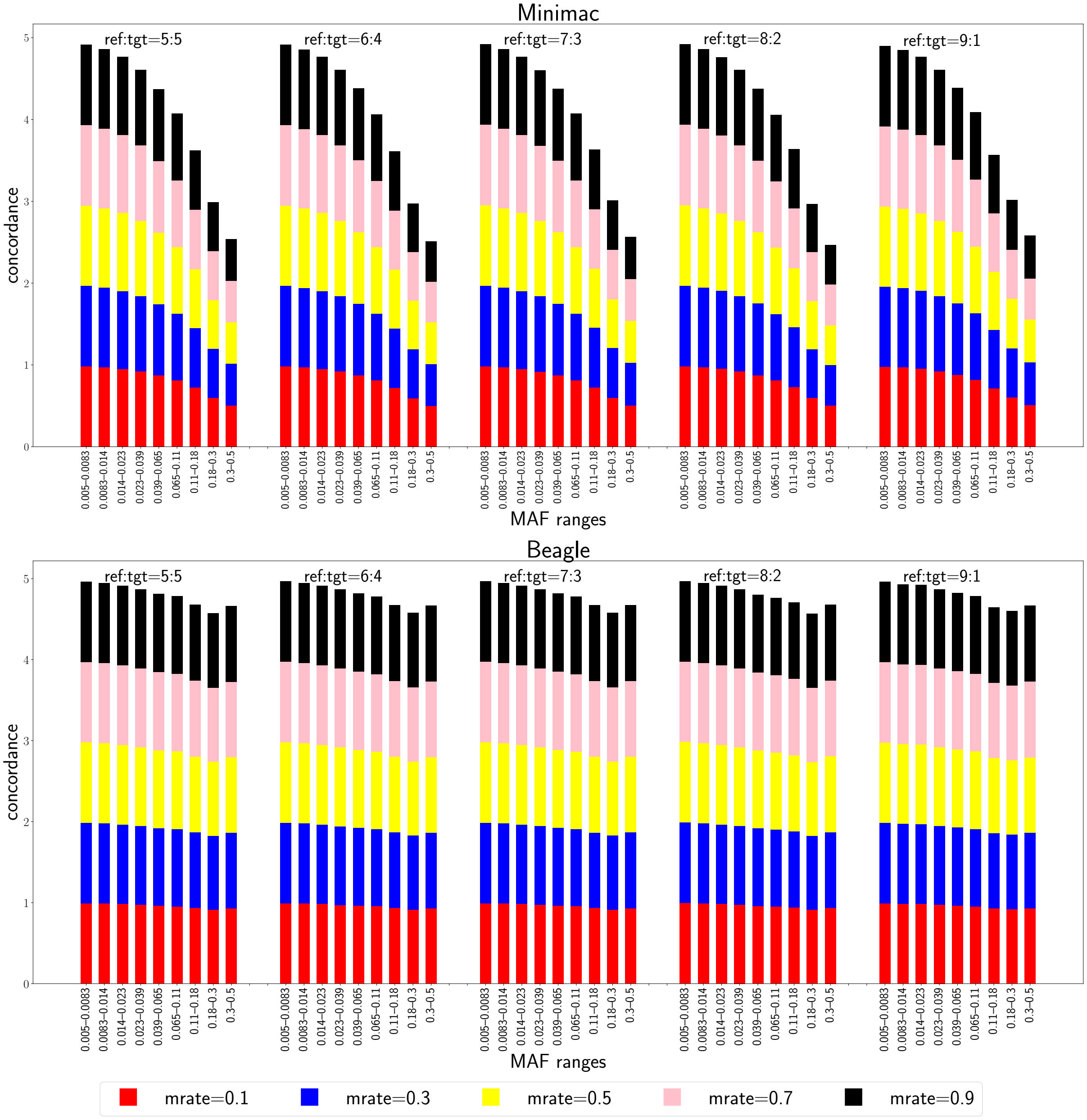

Figure 4.13: Concordance results for the rice chromosome 12 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

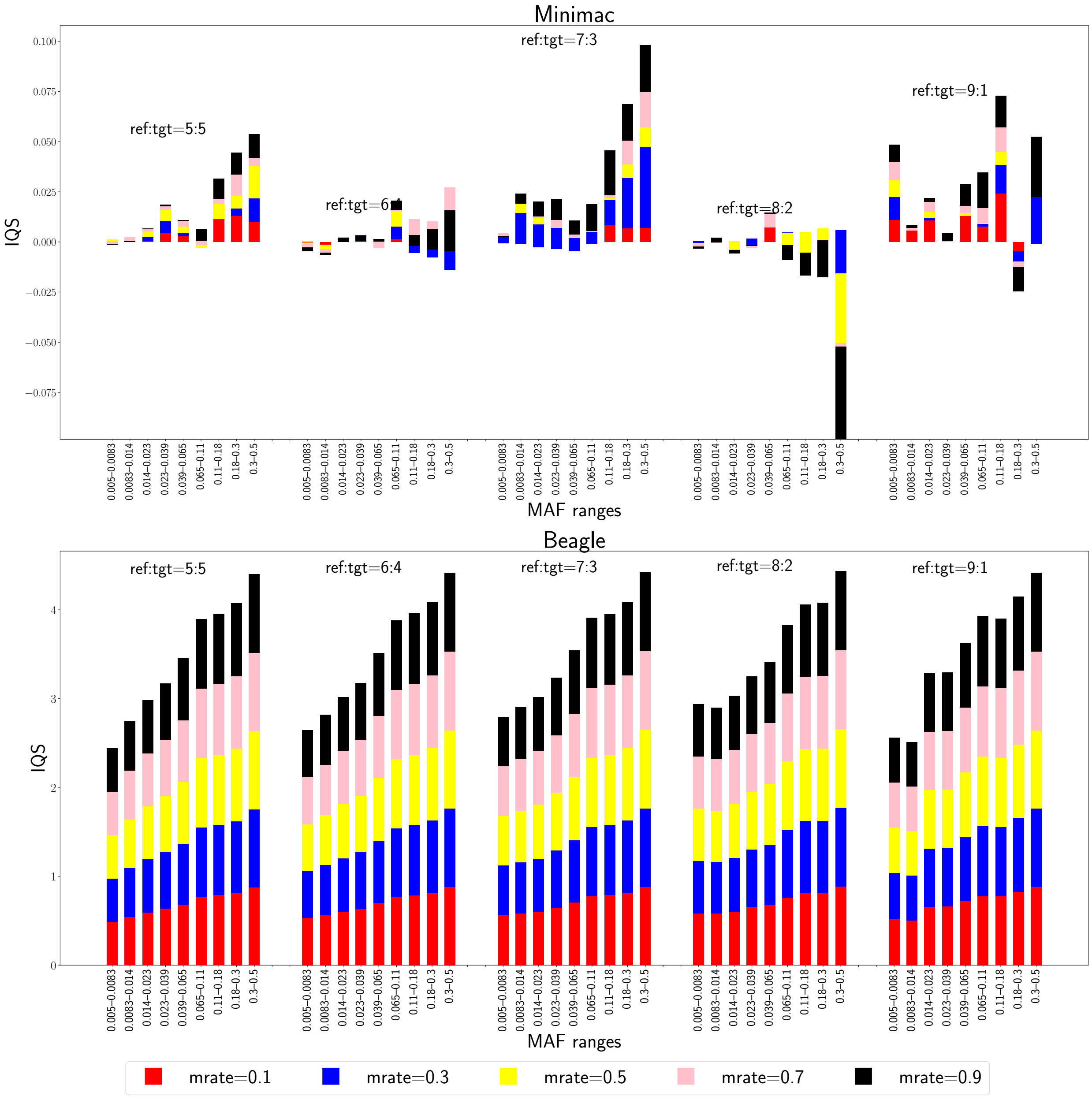

Figure 4.14: IQS results for the rice chromosome 12 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [-0.075, 0.100]; y-axis range for Beagle (lower row): [0, 5]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

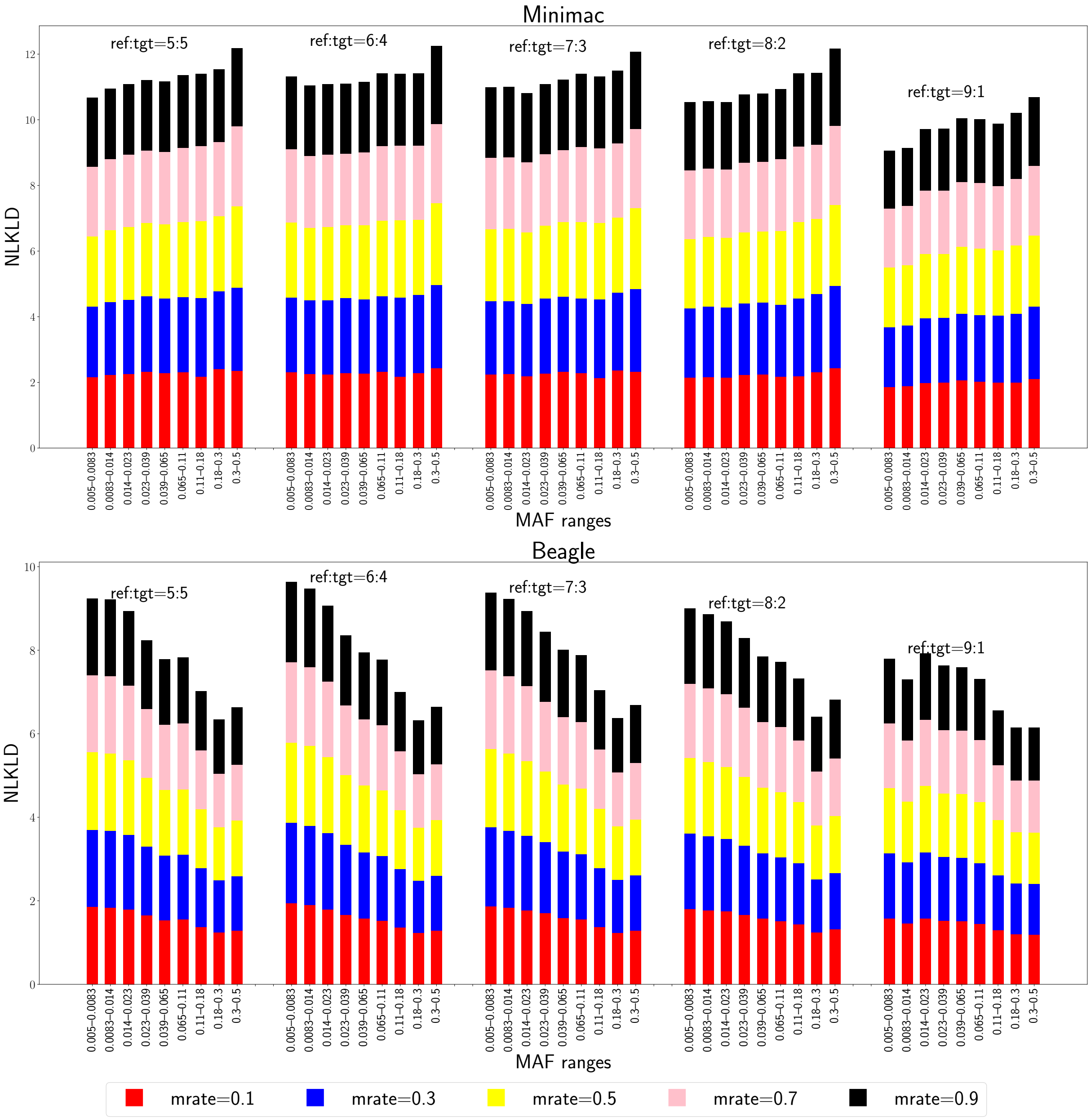

Figure 4.15: NLKLD results for the rice chromosome 12 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 12]; y-axis range for Beagle (lower row): [0, 10]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

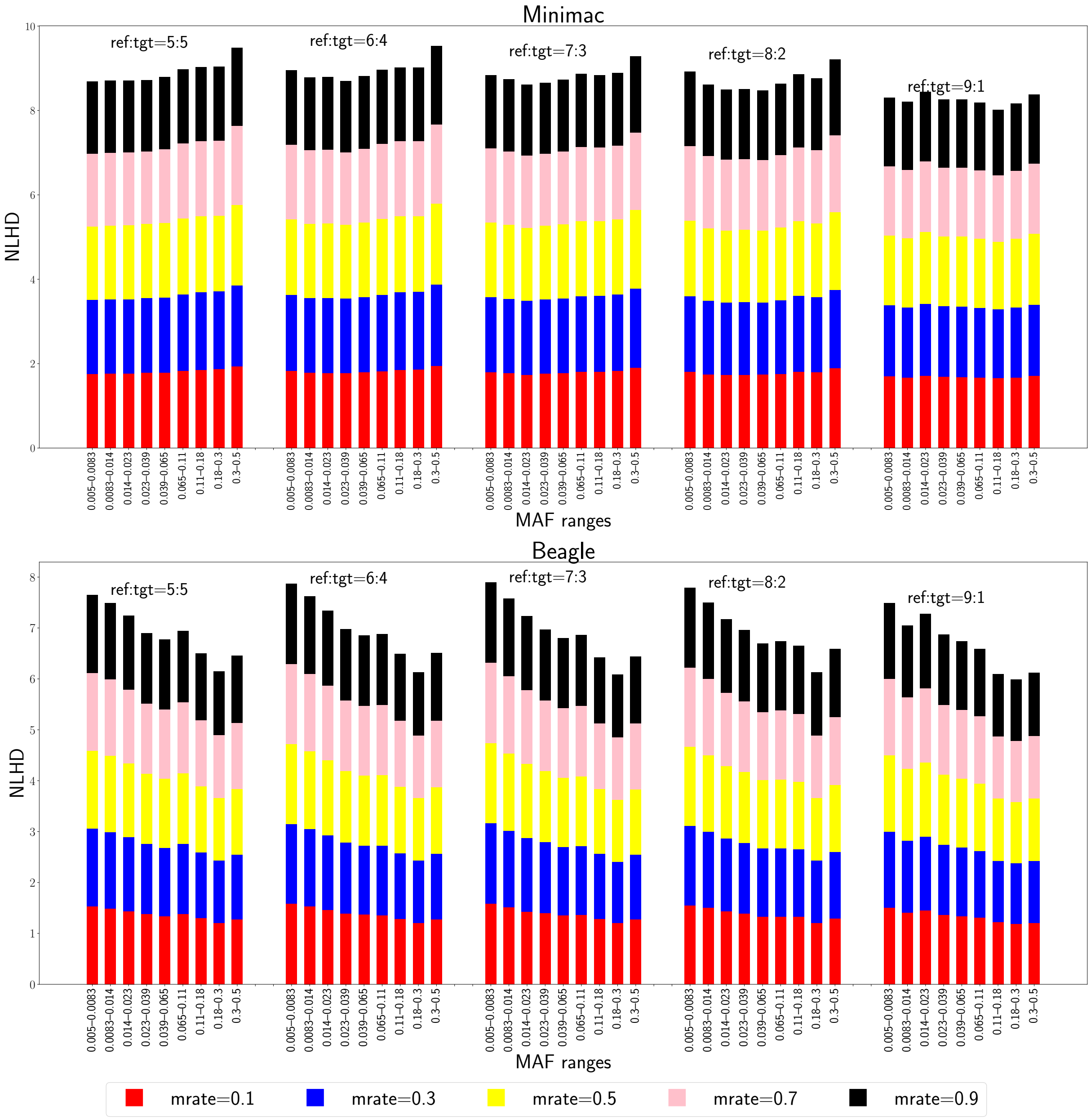

Figure 4.16: NLHD results for the rice chromosome 12 experimental data. Results from Minimac (upper row) and Beagle (lower row) are shown for different MAFs with different missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). y-axis range for Minimac (upper row): [0, 10]; y-axis range for Beagle (lower row): [0, 8]; MAF ranges (x-axis) for both programs: 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, 0.3–0.5.

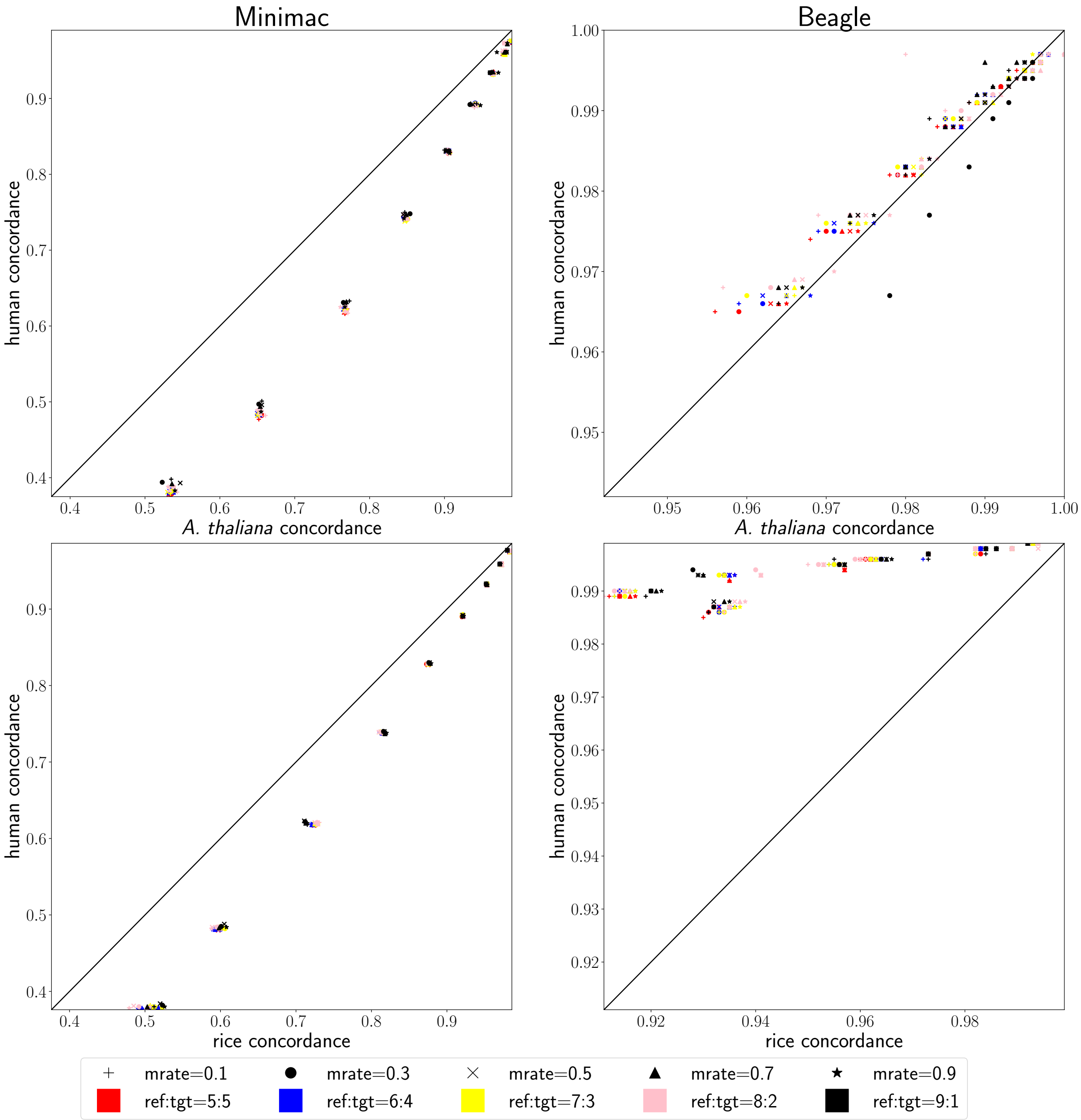

Figure 4.17: Comparisons of concordance between plant and the human with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each column of scatterplots shows the comparisons for one program. Each row of scatterplots shows the comparisons between a plant (x-axis) and human (y-axis). The diagonal line in each plot represents the case of the program performing equally well on both datasets, so the organism with superior results has more data points on its side.

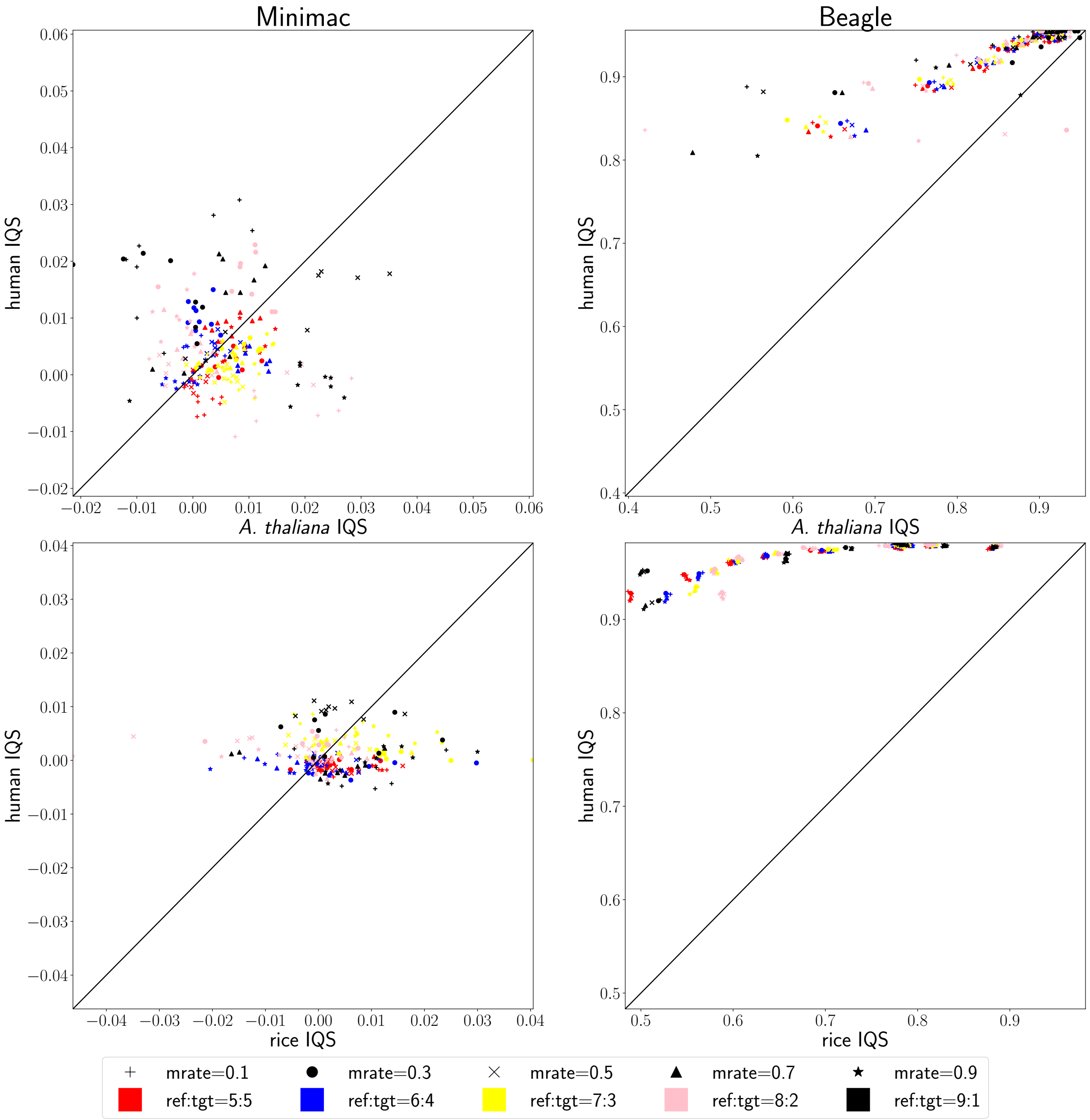

Figure 4.18: Comparisons of IQS between plant and the human with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each column of scatterplots shows the comparisons for one program. Each row of scatterplots shows the comparisons between a plant (x-axis) and human (y-axis). The diagonal line in each plot represents the case of the program performing equally well on both datasets, so the organism with superior results has more data points on its side.

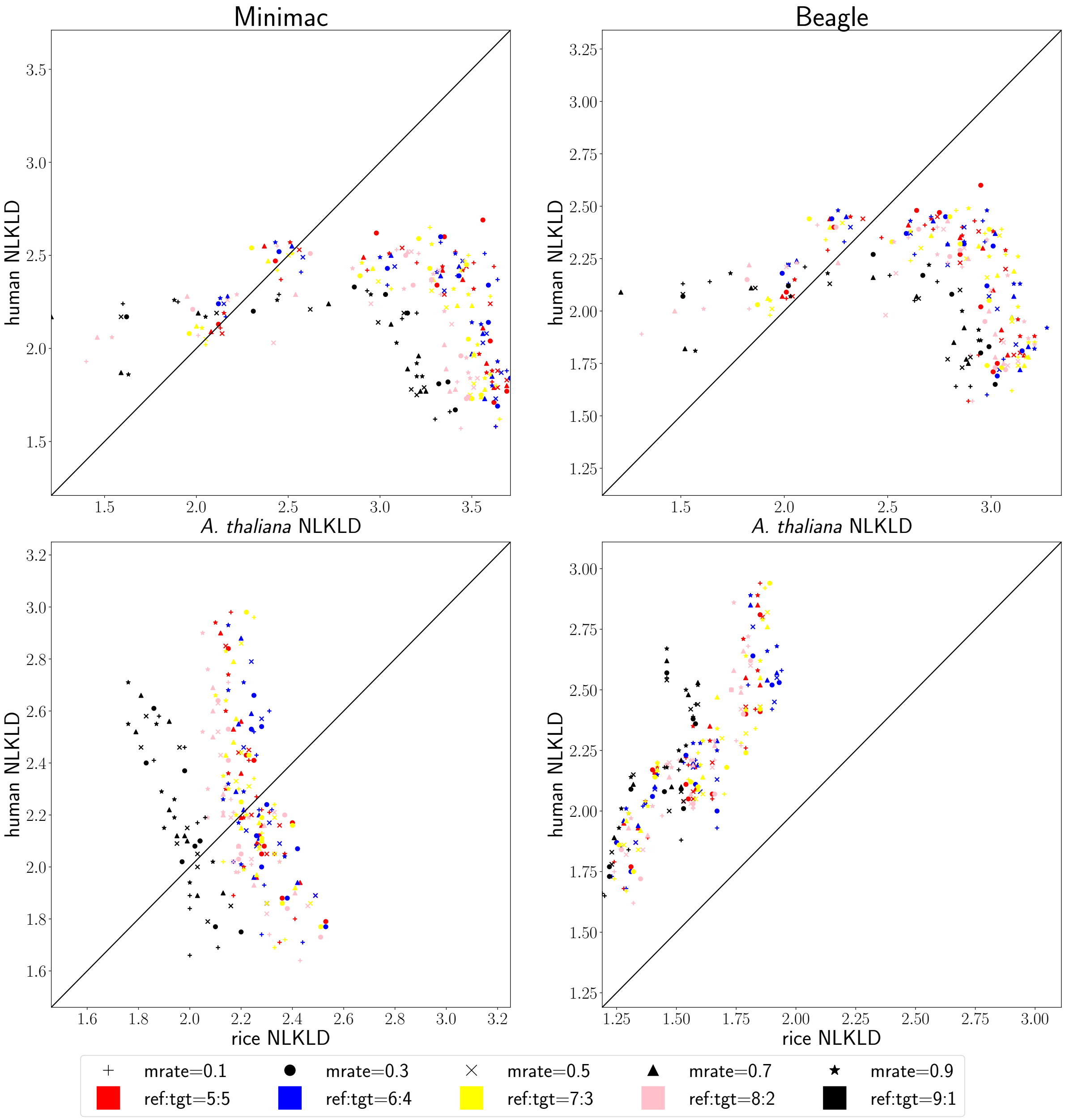

Figure 4.19: Comparisons of NLKLD between plant and the human with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each column of scatterplots shows the comparisons for one program. Each row of scatterplots shows the comparisons between a plant (x-axis) and human (y-axis). The diagonal line in each plot represents the case of the program performing equally well on both datasets, so the organism with superior results has more data points on its side.

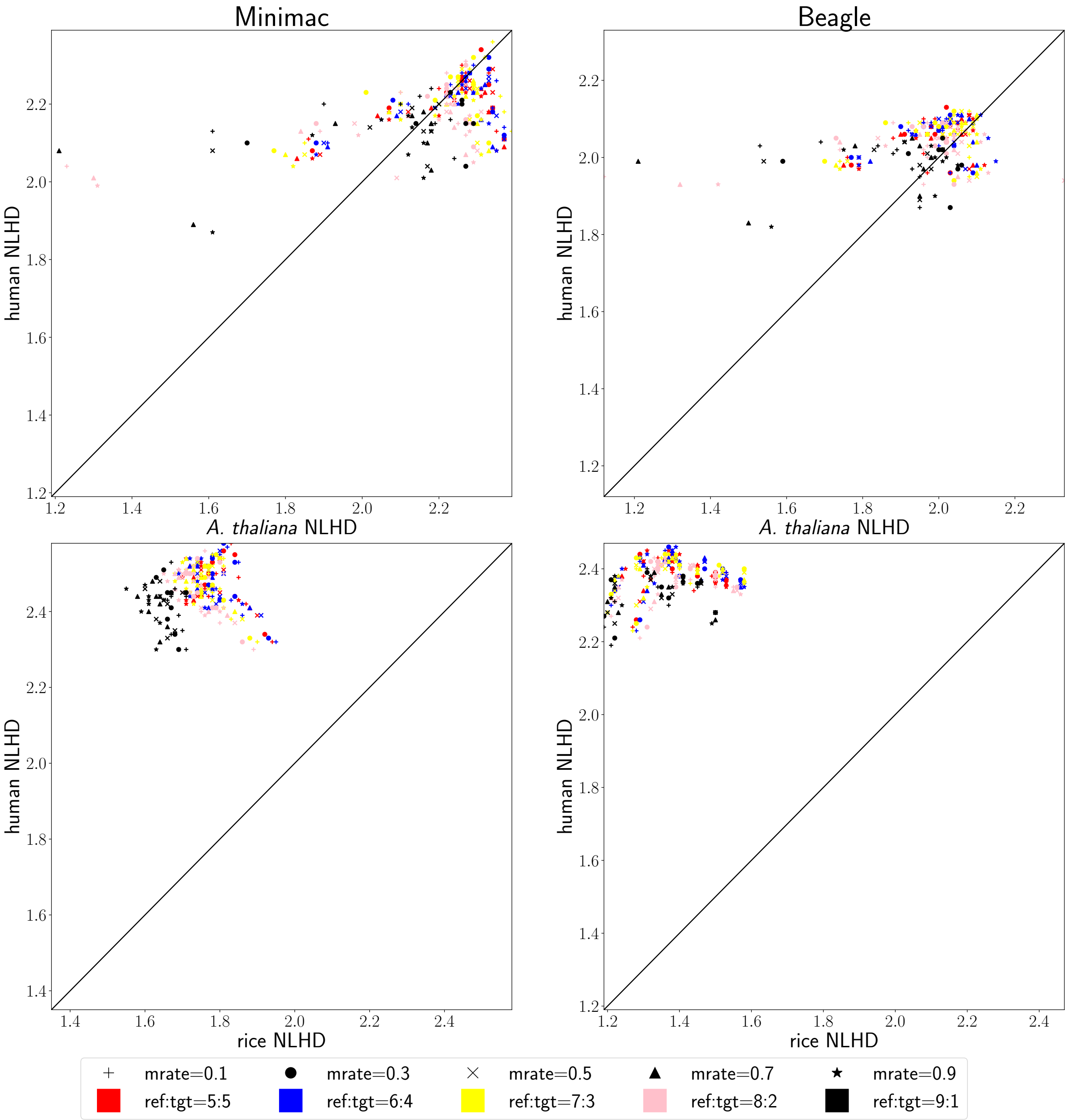

Figure 4.20: Comparisons of NLHD between plant and the human with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each column of scatterplots shows the comparisons for one program. Each row of scatterplots shows the comparisons between a plant (x-axis) and human (y-axis). The diagonal line in each plot represents the case of the program performing equally well on both datasets, so the organism with superior results has more data points on its side.

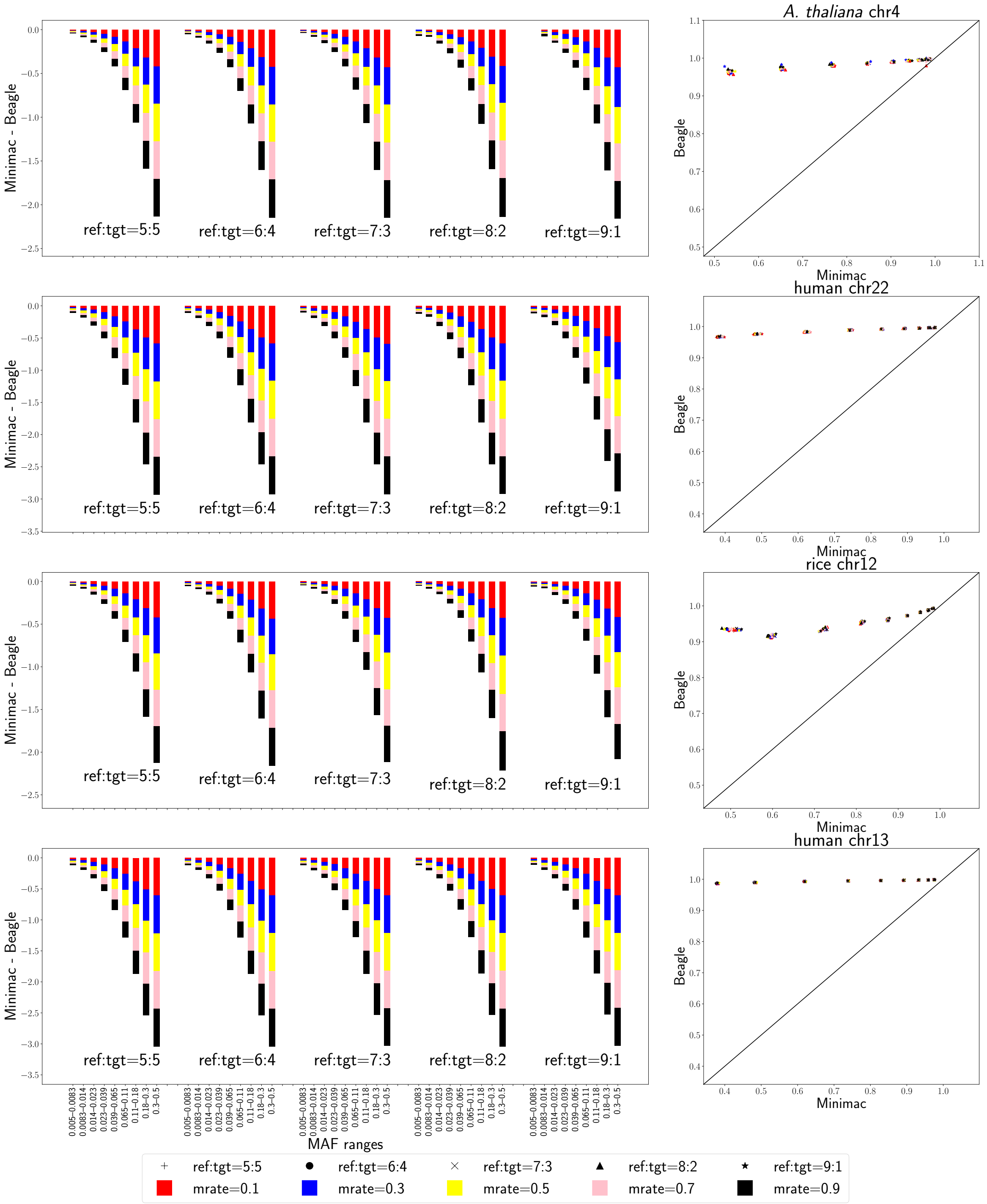

Figure 4.21: Comparisons of concordance between Minimac and Beagle with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each row shows a comparison between the two programs for one set of experimental data. In each bar-plot, the y-axis represents the results of Minimac minus that of Beagle so bars being below the y=0 line mean that Beagle performs superior to Minimac and bars being above the y=0 line mean that Minimac performs superior to Beagle. In each scatterplot, the x-axis represents the Minimac results whereas the y-axis represents the Beagle results. The same pair of datasets is considered in each row. The diagonal line in each scatterplot represents the case where both programs perform equally well on the dataset and hence the program with superior performance has more data points on its side.

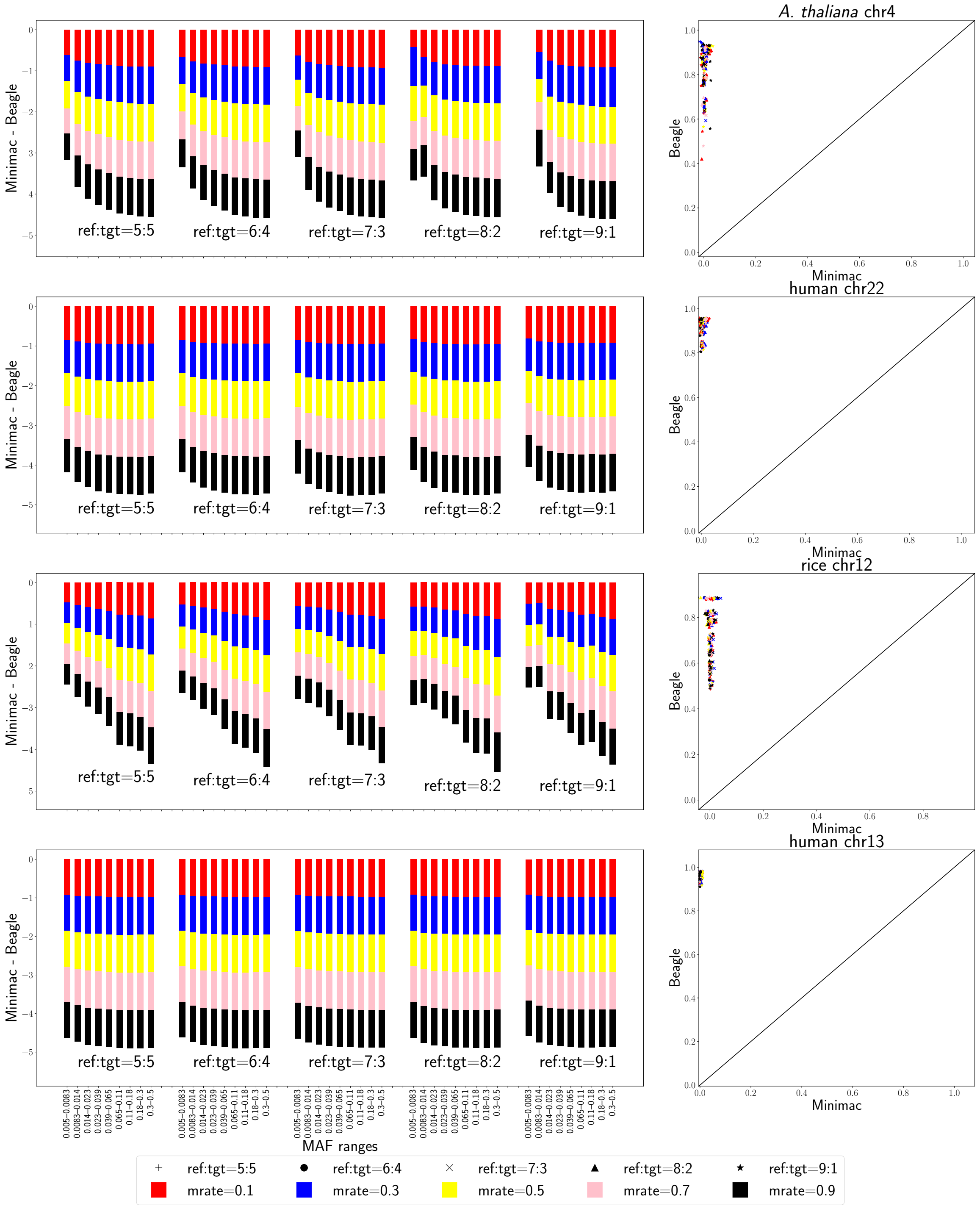

Figure 4.22: Comparisons of IQS between Minimac and Beagle with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each row shows a comparison between the two programs for one set of experimental data. In each bar-plot, the y-axis represents the results of Minimac minus that of Beagle so bars being below the y=0 line mean that Beagle performs superior to Minimac and bars being above the y=0 line mean that Minimac performs superior to Beagle. In each scatterplot, the x-axis represents the Minimac results whereas the y-axis represents the Beagle results. The same pair of datasets is considered in each row. The diagonal line in each scatterplot represents the case where both programs perform equally well on the dataset and hence the program with superior performance has more data points on its side.

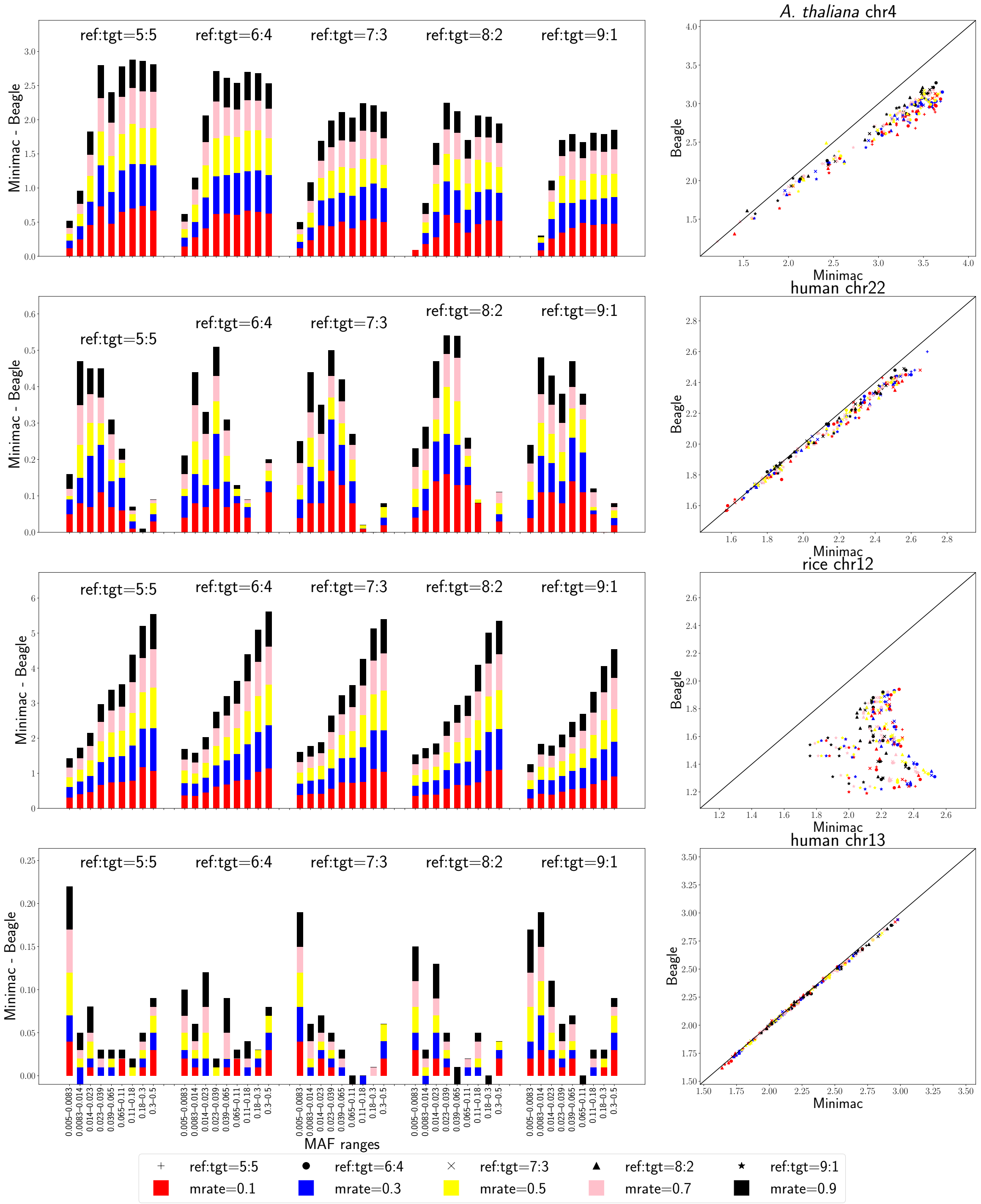

Figure 4.23: Comparisons of NLKLD between Minimac and Beagle with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each row shows a comparison between the two programs for one set of experimental data. In each bar-plot, the y-axis represents the results of Minimac minus that of Beagle so bars being below the y=0 line mean that Beagle performs superior to Minimac and bars being above the y=0 line mean that Minimac performs superior to Beagle. In each scatterplot, the x-axis represents the Minimac results whereas the y-axis represents the Beagle results. The same pair of datasets is considered in each row. The diagonal line in each scatterplot represents the case where both programs perform equally well on the dataset and hence the program with superior performance has more data points on its side.

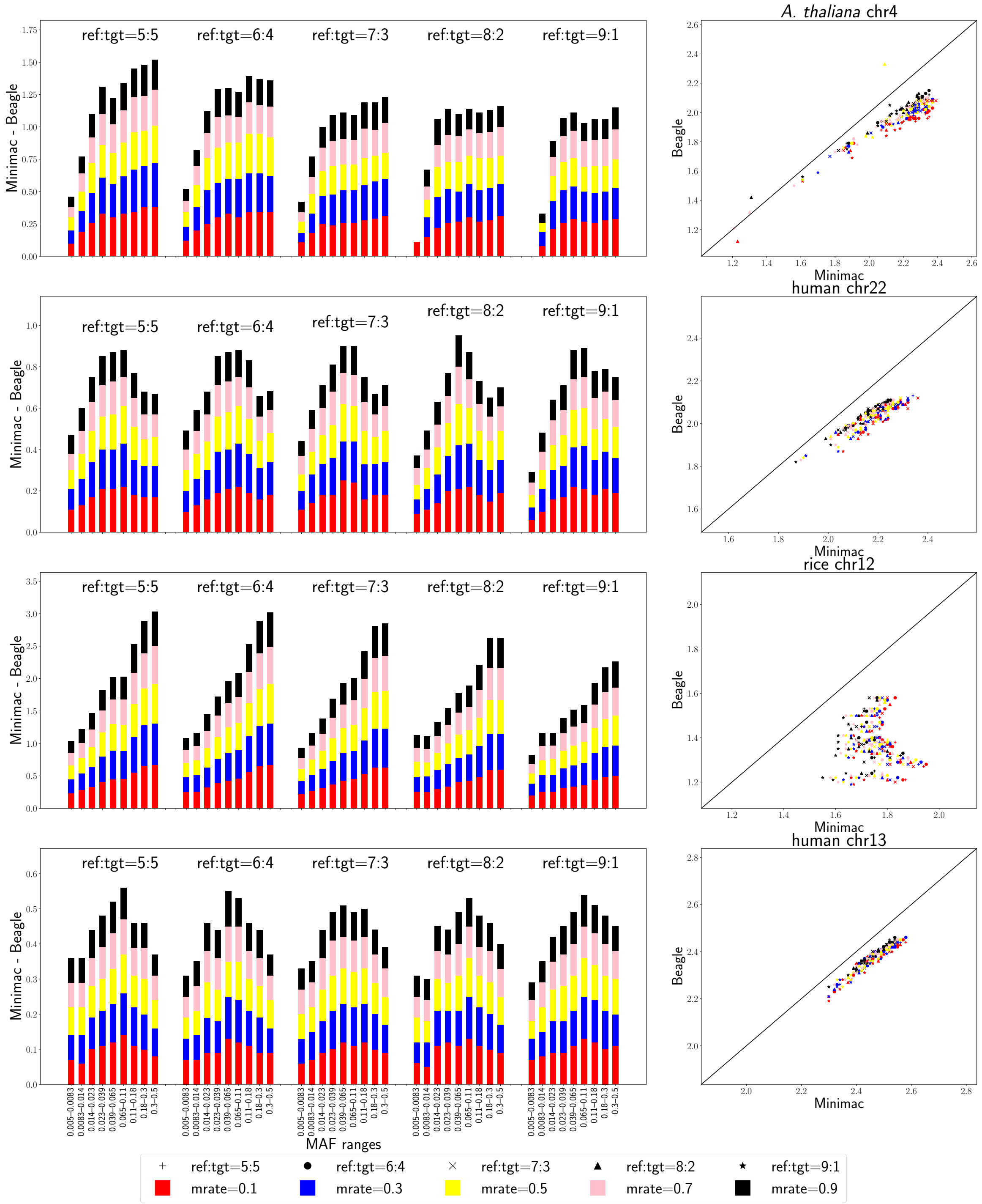

Figure 4.24: Comparisons of NLHD between Minimac and Beagle with varying missing rates (mrate) and sample size ratios (ref:tgt) between reference (ref) and target (tgt). Each row shows a comparison between the two programs for one set of experimental data. In each bar-plot, the y-axis represents the results of Minimac minus that of Beagle so bars being below the y=0 line mean that Beagle performs superior to Minimac and bars being above the y=0 line mean that Minimac performs superior to Beagle. In each scatterplot, the x-axis represents the Minimac results whereas the y-axis represents the Beagle results. The same pair of datasets is considered in each row. The diagonal line in each scatterplot represents the case where both programs perform equally well on the dataset and hence the program with superior performance has more data points on its side.

Chapter 5

Discussion

In this thesis, we present novel evaluations of current genotype imputation programs. In this chapter, we discuss the limitation of the more standard measure IQS in section 5.1. In section 5.2, we present the results of testing LinkImpute, an imputation program that does not require a reference genome. In section 5.3, we address the limitations of our study. Finally, in section 5.4, we discuss other statistical measures that are potentially useful in evaluating genotype imputation programs.

5.1 Limitation of IQS

Although we utilized IQS in our study, it may have critical issues that make it uninformative and hence inappropriate for imputation studies. As discussed in section 2.5, IQS is based on a kappa statistic. In their paper, Pontius and Millones [35] proposed *quantity disagreement* and *allocation disagreement* as a substitute for a kappa statistic. Quantity disagreement is the amount of disagreement between the known and imputed alleles for a given genotype, regardless of which samples/participants are assigned to specific alleles. For instance, if a SNP is detected in a population with 100 samples, and genotype AA appears 12 times in imputed results whereas 10 times in the known ones, the quantity disagreement of genotype AA is 2/100. Allocation disagreement, on the other hand, is the amount of disagreement between the known and imputed genotypes with an optimal match after excluding the quantity disagreement. Again, we use the same example as for quantity disagreement to illustrate this term. For the same SNP site as in the previous example, suppose samples 1 through 5 all have the same genotype AA and their imputed genotypes are all AA (i.e., the imputed and known genotypes agree on the same SNP site for the same five samples). Also, suppose samples 6 through 10 have the known genotype AA whereas their imputed genotypes are all Aa. Samples 11 through 17 all have the same genotype Aa whereas their imputed genotypes are all AA. Finally, the remaining 83 samples all have agreed known and imputed genotypes Aa. In this case, we can tell that genotype AA is incorrectly assigned to 7 samples (samples 11 through 17) and 5 samples are incorrectly assigned to genotypes other than AA (samples 6 through 10). Hence, the allocation disagreement of AA is $(5 + 7)/100 - 2/100 = 10/100$, where 2/100 is the quantity disagreement that should be excluded. In order to compare the two disagreements with IQS, we draw Table 5.1 and calculate the IQS for this example. Here, the concordance p_0 is $(5 + 83 + 0)/100 = 0.88$, and the hypothetical chance agreement probability p_e is $(12 \times 10 + 88 \times 90 + 0 \times 0)/100^2 = 0.804$. Hence, the IQS value for this example is $(0.88 - 0.804)/(1 - 0.804) \approx 0.388$. For a given SNP site, while the IQS suggests a poor imputation result, the two disagreements do not. According to Pontius and Millones [35], 1) the same kappa statistic might not necessarily reveal the same level of allocation disagreement or quantity disagreement, leading to useless or confusing information; and 2) a larger kappa statistic might not necessarily correspond to a larger total of allocation and quantity disagreements, making the kappa statistics even harder to interpret. Therefore, allocation disagreement and quantity disagreement can be more useful than IQS in evaluating genotype imputation results and the these two disagreements should be considered in future genotype imputation studies.

Table 5.1: Counts of agreement and disagreement between the known and imputed genotypes.

		Known genotypes			
		AA	Aa	aa	Total
Imputed genotypes	AA	5	7	0	12
	Aa	5	83	0	88
	aa	0	0	0	0
	Total	10	90	0	100

5.2 Performance of non-reference-based imputation program

Since the majority of existing genotype imputation programs are designed to infer missing genotypes from reference genotypic data (possibly with the aid of genetic maps), imputation for organisms without reference genotypic data, e.g., lentil, is impractical. However, in 2015, Money et al. [28] developed LinkImpute that does not require reference genotypic data. Since little work has been done to show how effective such a non-reference-based imputation programs is, comparing LinkImpute with reference-based programs under same experiment settings (e.g., input datasets and default program parameters) can be beneficial. During our study, we attempted to compare the performance of LinkImpute with reference-based genotype imputation programs (i.e., Minimac and Beagle). Specifically, we designed similar experiments and used the same data as in the experiments to test Minimac and Beagle as described in section 3.4. To make the data "reference-free", we combined the reference and target data and used the combined data as input.

However, LinkImpute did not work properly on any of the experimental data as used to evaluate Minimac and Beagle; the program exited quickly without producing useful output. We speculate that LinkImpute is not yet capable of imputing large datasets with thousands of samples and that the program is not as versatile as its reference-based peers. Future studies can further evaluate non-reference-based genotype imputation programs and compare them with reference-based genotype imputation programs to determine which type of program is superior to the other.

5.3 Limitations of study

In subsection 3.4.1, we utilized a 5% missing rate threshold to filter the original SNP data for various organisms. However, this missing rate threshold did not make close the number of filtered SNP sites between the rice chromosome 12 and human chromosome 13 data. The difference of SNP sites would potentially affect the performance evaluation of genotype imputation programs. Hence, for filtering SNP data, future study may consider a missing rate threshold that makes close the number of SNP sites between organisms.

In subsections 3.5.1 and 3.5.2, we used negative-logarithm-base-ten of K-L divergence and Hellinger distance to magnify the signals. However, negative-logarithm-base-two of K-L divergence and Hellinger distance could further magnify the signals. This alternative could be explored in future work. In addition, we used default settings to test Minimac and Beagle. However, Minimac automatically adjusted parameters based on the input data whereas Beagle did not. This could be important since the default population size for input data assumed by Beagle is one million, and our datasets had thousands of samples rather than millions. The effect of this discrepancy is unknown and could be explored as future work. This would involve manually setting the "ne" parameter to "1e3" (thousands of samples/participants) prior to running Beagle. Further, the failure of the LinkImpute tests (section 5.2) suggests that effort is warranted on non-reference-based imputation programs that provide not only enhanced data format compatibility but also improved capability of imputing large-scale datasets. Finally, other aspects could have also caused the disagreement among performance measurements: (1) The *A. thaliana* and rice data were phased computationally. Unfortunately, the accuracy of this phasing process is not well understood. (2) Quality of SNP data varies among organisms and is heavily dependent on the read depth and log-likelihood of SNP calling. The impact of data quality is not well understood and could potentially result in misleading results. Therefore, future studies may consider investigating the effect of different phasing programs and the impact of data quality on imputation results.

5.4 Statistical measures for evaluating genotype imputation results

We demonstrated that Pearson correlation is not appropriate to evaluate the performance of genotype imputation programs. Also, we discussed that IQS was not ideal to compare imputation performances. Moreover, negative logarithmic K-L divergence and negative logarithmic Hellinger distance are appropriate measures of ranking genotype imputation results along with imputation quality score and concordance. To better interpret correspondence between the known and imputed genotypes, one can consider Kendall's *tau* since this non-parametric test is independent of distributions [31]. One can also find other useful ranking tools in Chapter 9.7 of Murphy's book [32]. Finally, one can consider quantity and allocation disagreements (described in section 5.1) to replace IQS and compare results with NLKLD and NLHD.

Chapter 6

Conclusions

In this study, we demonstrated that Pearson correlation is inappropriate to evaluate the correspondence between the known and imputed genotypes. Additionally, we explained in theoretical terms that both NLKLD and NLHD are more appropriate than IQS and concordance to rank imputation results. However, the four measures do not appear to agree with each other when used to compare imputation results either between plant and human, or between Minimac and Beagle.

Table 6.1 shows the results of the comparisons between plant and human for Beagle and Minimac. As can be seen, Beagle overall performed better on human than plant. In contrast, the Minimac results are less conclusive than Beagle.

Table 6.1: Results of the comparisons between human and plant for Beagle and Minimac. For a given comparison pair, each cell gives the organism with the superior results.

Comparison pair	Concordance		Comparison pair	IQS	
	Program			Program	
	Beagle	Minimac		Beagle	Minimac
human vs. *A. thaliana*	human	*A. thaliana*	human vs. *A. thaliana*	human	**inconclusive**
human vs. rice	human	rice	human vs. rice	human	**inconclusive**

Comparison pair	NLKLD		Comparison pair	NLHD	
	Program			Program	
	Beagle	Minimac		Beagle	Minimac
human vs. *A. thaliana*	*A. thaliana*	*A. thaliana*	human vs. *A. thaliana*	human	**inconclusive**
human vs. rice	human	**inconclusive**	human vs. rice	human	human

Table 6.2 shows the results of comparison between Beagle and Minimac for all experimental data. In Table 6.2, Beagle outperforms Minimac according to concordance and IQS. These results mean that Beagle has a superior percent accuracy over Minimac. However, the probability distribution of the imputed SNPs from Minimac better reflects the probability distribution of the known SNPs than in the case Beagle according to NLKLD and NLHD.

Table 6.2: Results of the comparisons between Minimac and Beagle for all experimental data. For a given measure and specific set of experimental data, each cell gives the program with superior performance.

Measure	Concordance			
Data	*A. thaliana* chr4	human chr22	human chr13	rice chr12
Minimac *vs.* Beagle	Beagle	Beagle	Beagle	Beagle

Measure	IQS			
Data	*A. thaliana* chr4	human chr22	human chr13	rice chr12
Minimac *vs.* Beagle	Beagle	Beagle	Beagle	Beagle

Measure	NLKLD			
Data	*A. thaliana* chr4	human chr22	human chr13	rice chr12
Minimac *vs.* Beagle	Minimac	Minimac	Minimac	Minimac

Measure	NLHD			
Data	*A. thaliana* chr4	human chr22	human chr13	rice chr12
Minimac *vs.* Beagle	Minimac	Minimac	Minimac	Minimac

6.1 Guidance on the use of existing imputation programs

With the experience from our trials of two existing genotype imputation programs using different datasets in various situations, we have the following suggestions for future study of genotype imputation programs: (1) Pearson correlation coefficient is not appropriate to evaluate performance of genotype imputation programs, and hence should not be considered; (2) in addition to concordance and IQS, NLKLD and NLHD are useful measures to evaluate performance of genotype imputation programs; (3) to impute human missing genotypes, Beagle should be considered because Beagle favours human over plant data.

6.2 Future work

In our results, Beagle had overall consistently superior results on human over plant data whereas Minimac did not have consistently superior results on either plant or human data. In addition, both NLKLD and NLHD suggest that Minimac has a superior imputation method over Beagle's. However, further studies with more data of the same and different kinds are in order to confirm these trends. In addition, although we theoretically explained that NLKLD and NLHD were more appropriate to measure imputation performance than concordance and IQS, future studies may provide more evidence to support this point of view. Further, we compared imputation performance between plant and human for Minimac and Beagle using data from different chromosomes. This study was based on the assumption that each imputation program had the same performance on different chromosomes of the same organism. However, such an assumption might not be valid. Therefore, future research may consider using data from different chromosomes from one organism to verify such an assumption. Finally, we drew diagonal lines in scatterplots to not only compare performance between imputation programs, but also compare imputation results between plants and humans. Such a visual inspection would suggest that 1) the performance was similar for both cases and 2) there was some random variation (as one would expect). To make the results quantitative, future study may consider measures such as sums of squared estimates of errors (SSE) for data points above and below the diagonal line in each scatterplot and then compare the two SSE values.

References

[1] 1000 Genomes human SNP data. ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, November 2017.

[2] 1001 Genomes *Arabidopsis thaliana* SNP data. https://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snpeff_v3.1/, Feburary 2018.

[3] Cohen's kappa. https://en.wikipedia.org/wiki/Cohen's_kappa, October 2018.

[4] The European Bioinformatics Institute rice SNP data. ftp.sra.ebi.ac.uk/vol1/, April 2018.

[5] Gene. https://www.genome.gov/genetics-glossary/Gene, November 2019.

[6] Genetic variation. https://www.genome.gov/genetics-glossary/Genetic-Variation, November 2019.

[7] Genome-wide association studies. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies, November 2019.

[8] Genotype. https://www.genome.gov/genetics-glossary/Genotype, December 2019.

[9] Phenotype. https://www.genome.gov/genetics-glossary/Phenotype, November 2019.

[10] Single nucleotide polymorphism. https://ghr.nlm.nih.gov/primer/genomicresearch/snp, November 2019.

[11] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[12] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[13] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2 edition, 2003.

[14] Thomas M. Cover and Joy A. Thomas. Elements of information theory. In *Wiley Series in Telecommunications and Signal Processing*, chapter 2.3 Relative entropy and mutual information, pages 19–20. Wiley-Interscience, New York, NY, USA, 2 edition, 2006.

[15] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284, 2016.

[16] Dana B Hancock, Joshua L Levy, Nathan C Gaddis, Laura J Bierut, Nancy L Saccone, Grier P Page, and Eric O Johnson. Assessment of genotype imputation performance using 1000 genomes in african american studies. *PLoS One*, 7(11):e50610, 2012.

[17] John M. Hickey, Jose Crossa, Raman Babu, and Gustavo de los Campos. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science*, 52(2):654–663, 2012.

[18] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955, 2012.

[19] Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011.

[20] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

[21] Sekar Kathiresan, Olle Melander, Candace Guiducci, Aarti Surti, Noël P Burtt, Mark J Rieder, Gregory M Cooper, Charlotta Roos, Benjamin F Voight, Aki S Havulinna, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, 40(2):189, 2008.

[22] Mingyao Li, Pelin Atmaca-Sonmez, Mohammad Othman, Kari EH Branham, Ritu Khanna, Michael S Wade, Yun Li, Liming Liang, Sepideh Zareparsi, Anand Swaroop, et al. Cfh haplotypes without the y402h coding variant show strong association with susceptibility to age-related macular degeneration. *Nature Genetics*, 38(9):1049, 2006.

[23] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10:387–406, 2009.

[24] Peng Lin, Sarah M Hartz, Zhehao Zhang, Scott F Saccone, Jia Wang, Jay A Tischfield, Howard J Edenberg, John R Kramer, Alison M Goate, Laura J Bierut, and John P Rice. A new statistic to evaluate imputation reliability. *PLoS ONE*, 5(3), 2010.

[25] Qian Liu, Elizabeth T Cirulli, Yujun Han, Song Yao, Song Liu, and Qianqian Zhu. Systematic assessment of imputation performance using the 1000 genomes reference panels. *Briefings in Bioinformatics*, 16(4):549–562, 2014.

[26] Po Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

[27] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906, 2007.

[28] Daniel Money, Kyle Gardner, Zoë Migicovsky, Heidi Schwaninger, Gan-Yuan Zhong, and Sean Myles. Linkimpute: fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genetics*, 5(11):2383–2390, 2015.

[29] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 2.8.2 KL divergence, pages 57–58. The MIT Press, 2012.

[30] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 21.2.2 Forward or reverse KL?, page 735. The MIT Press, 2012.

[31] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 9.7.4 Loss functions for ranking, page 304. The MIT Press, 2012.

[32] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 9.7 Learning to rank, pages 300–305. The MIT Press, 2012.

[33] Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Review Genetics*, 12(6):443–451, 2011.

[34] Marju Orho-Melander, Olle Melander, Candace Guiducci, Pablo Perez-Martinez, Dolores Corella, Charlotta Roos, Ryan Tewhey, Mark J Rieder, Jennifer Hall, Goncalo Abecasis, et al. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and c-reactive protein but lower fasting glucose concentrations. *Diabetes*, 57(11):3112–3121, 2008.

[35] Robert Gilmore Pontius Jr and Marco Millones. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011.

[36] Shelina Ramnarine, Juan Zhang, Li Shiun Chen, Robert Culverhouse, Weimin Duan, Dana B Hancock, Sarah M Hartz, Eric O Johnson, Emily Olfson, Tae Hwi Schwantes-An, and Nancy L Saccone. When does choice of accuracy measure alter imputation accuracy assessments? *PLoS ONE*, 10(10), 2015.

[37] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[38] Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[39] Cristen J Willer, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, Simon C Heath, Nicholas J Timpson, Samer S Najjar, Heather M Stringham, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, 40(2):161, 2008.

[40] Zhenming Zhao, Nadia Timofeev, Stephen W Hartley, David HK Chui, Supan Fucharoen, Thomas T Perls, Martin H Steinberg, Clinton T Baldwin, and Paola Sebastiani. Imputation of missing genotypes: an empirical evaluation of impute. *BMC Genetics*, 9(1):85, 2008.

Appendix

Results data for drawing figures

The data for drawing graphs of results is available in this thesis document via the following link: `https://drive.google.com/open?id=1SGAeRACLq-cv31dSsaCOBvQDFYFvuD1r`.

The root directory includes 3 sub-directories of results organized by organism names. For the human results, there are two subsubdirectories for results of two separate experiments. `report_s1135_chr22` contains files of results from the human chromosome 22 data with 1135 participants, and `report_s2504_chr13` contains files of results from the human chromosome 13 data with 2504 participants. Each result filename indicates the corresponding measurement. In each file, the header `#ratio` separates results by separation ratios. Under each separation ratio there are 18 rows of results corresponding to results of 9 missing rates for Minimac and Beagle. The odd rows are Minimac results whereas the even rows are Beagle results. In other words, the order of results rows is Minimac results with 10% missing rate, Beagle results with 10% missing rate, Minimac results with 20% missing rate, Beagle results with 20% missing rate, etc. Each row contains 9 values that are separated by minor allele frequency range, i.e., 0.005–0.0083, 0.0083–0.0014, 0.0014–0.023, 0.023–0.039, 0.039–0.065, 0.065-0.11, 0.11–0.18, 0.18–0.3, and 0.3–0.5.