Chapter 6

Genotype Imputation

6.1 Motivation

There are two types of missing data in a case-control association study: some individuals have missing genotype values at loci which are otherwise successfully typed; and there are loci which are not typed at all. Most analysis methods use only the data that has been collected, but an alternative approach that has received a lot of attention in the statistical literature is to predict, or impute, missing values based on the observed data, and use the complete data for analyses (Rubin, 1987).

In what follows, *imputing missing genotypes* refers to imputing genotype values at loci which are typed in the sample of interest, but which are missing for a small fraction of the individuals in that sample. *Imputing untyped loci* refers to imputing genotypes at loci which are entirely untyped in the sample of interest, but which are typed in some other sample.

The ARG mapping approach described in Chapter 3 tests all branches on the marginal tree at a typed marker, that is, all possible SNPs (real or otherwise) that are compatible with the genealogy. A more direct approach is to test only those branches that correspond to known SNPs, typed or untyped. One way to do this would be to merge a sample of more densely genotyped individuals, such as from the HapMap Project, with the case-control sample. The ARG inference algorithm can then be used to impute genotypes at loci that are missing in the cases and controls but present in the denser sample. The imputed loci could

then be tested directly for association.

There are at least three additional reasons why we might want to estimate, or impute, genotypes that are not present in observed data.

First, any large-scale genotyping effort is likely to produce incomplete data. Depending on the subsequent analyses, and the processes generating the missing genotypes, incomplete data can lead to biased results. For example, in case-control studies, a systematic bias in the genotyping quality of an allele may result in false, or missed associations (Clayton et al., 2005), and incomplete data can hinder multi-SNP analyses of haplotypic or interaction effects. Furthermore some analysis methods are unable to operate on data that has any missing genotypes.

Second, imputation of untyped loci may help fine map disease causing variants. By imputing untyped loci at an association peak, additional informative association structure in that region may be found. If an imputed locus shows stronger association than the observed loci around it, it may be causative, or more strongly linked to the causative polymorphism(s).

Third, imputing untyped loci has the potential to give greater power to detect disease associations than relying on pairwise LD. As discussed in the Introduction, when the single marker chi-square test is applied to case-control association study data, the association signal at a SNP in LD with a causative polymorphism is dependent on the r^2 LD between those two loci (Pritchard and Przeworski, 2001). If no typed SNP is in sufficient r^2 LD with the untyped causative polymorphism, the association will be missed. Furthermore, it is conceivable that allelic heterogeneity and interaction can cause single marker tests at SNPs in strong r^2 LD with an untyped causative variant to show no significant signal (Terwilliger and Hiekkalinna, 2006). However, accurate imputation of the untyped causative polymorphism will recover the association signal. The scenario (albeit, a very unlikely one) constructed by Terwilliger and Hiekkalinna (2006) is as follows:

Consider a three SNP haplotype, with loci A, B and C, with alleles a, A; b, B; and c, C. Assuming that there is no recombination between these alleles, and no recurrent or back mutations, there can be at most four possible haplotype configurations. Suppose we observe the following haplotypes: A-B-C, A-B-c, a-B-C and A-b-c, each with equal frequencies, 0.25, in

the population. Now suppose that SNPs A and B are untyped and modify disease risk, while SNP C is typed but is not causative. If a and b affect the disease phenotype in a dominant fashion, with equivalent effect, i.e., $P(Disease|aa) = P(Disease|Aa) = P(Disease|bb) = P(Disease|Bb)$ and $P(Disease|AA, BB) = 0$, then the following contingency tables are possible:

Allele at A	Freq in Cases	Freq in Controls
A	0.667	0.757
a	0.333	0.243
		Odds Ratio 1.55

Allele at B	Freq in Cases	Freq in Controls
B	0.667	0.757
b	0.333	0.243
		Odds Ratio 1.55

Allele at C	Freq in Cases	Freq in Controls
C	0.5	0.5
c	0.5	0.5
		Odds Ratio 1.00

In this scenario, although loci A and C have $r^2 = 0.33$, the association at A will never be detected by genotyping C alone, regardless of sample size, because the odds ratio at C is 1.00. The same holds for detecting the association at B via C. In such a situation, A and B must be tested directly, or a multimarker method used. Accurate imputation of A and B and then "direct" testing would successfully identify the signal.

In this chapter, I evaluate the imputation performance of MARGARITA and compare it with fastPHASE (Scheet and Stephens, 2006), and apply the MARGARITA imputation approach to the 8q24/Prostate Cancer data of Yeager et al. (2007) to test loci not typed in the original

study.

6.2 Existing Methods for Imputing Genotype Data

There are a number of methods for imputing missing genotypes. The widely used PHASE (Stephens et al., 2001; Stephens and Donnelly, 2003; Stephens and Scheet, 2005) and fastPHASE (Scheet and Stephens, 2006) methods impute missing genotypes while inferring haplotype phase.

fastPHASE, the model for which is described in Chapter 2, can be used to impute missing genotypes in the following way: Once the model is fitted, the probability that a missing genotype takes a particular value is dependent on the probabilities of its haplotypes belonging to particular clusters, and the frequencies of the observed genotypes within those clusters. A point estimate of the missing genotype is taken by choosing the genotype with the greatest probability.

Within the context of association studies, there has been some discussion in the tagSNP literature (Goldstein et al., 2003) on explicitly estimating the genotypes of untyped SNPs from genotyped tagSNPs. Methods have been developed to do this (Evans et al., 2004; Souverein et al., 2006). It is also possible to select tagSNPs that optimise subsequent prediction accuracy of untyped SNPs (Nicolae, 2006; Paschou et al., 2007; Eyheramendy et al., 2007). However, only Souverein et al. (2006) and Paschou et al. (2007) apply their method to case-control association study data. I now briefly describe each of these approaches.

In Souverein et al. (2006), a linear or logistic regression model was fitted to a training data set containing genotype data for all SNPs, and then used to impute loci missing in the less densely typed data. The authors selected manually which predictor SNPs to use in the regression, and their method requires that the predictor SNPs are complete.

In Evans et al. (2004) windows of SNPs were taken, and population haplotype frequencies were determined from a training set. The probability that an individual has a particular value at a missing genotype was calculated by taking all the haplotypes matching the observed genotypes for the individual, and using these to fill in the missing value. The contribution of

each matching haplotype to the missing genotype was weighted by haplotype frequency. A similar approach was taken in Nicolae (2006).

Goldstein et al. (2003) discusses the idea of selecting tagSNPs on the basis of how well models involving those predict the untyped SNPs. This approach was adopted in Nicolae (2006); Eyheramendy et al. (2007) and Paschou et al. (2007).

Eyheramendy et al. (2007) describe a method for predicting non-tags using the Li and Stephens model (Li and Stephens, 2003). The model is fitted to training data, such as the HapMap, and then individuals sampled from the case-control study are introduced, and the missing genotypes imputed from the model. However, the method was only applied to the dense ENCODE HapMap data, not to an association study. They selected tagSNPs for the ENCODE data, fitted the model using a training set of haplotypes, and then measured how well the non-tags were reconstructed in the rest of the data, the same population.

Paschou et al. (2007) describe a method for choosing tagSNPs from a data set. During selection of tagSNPs, a singular value decomposition is performed on the genotype data matrix. The resultant eigenvectors give linear combinations of SNPs that capture the structure of the data, and can be used to select the SNPs that contribute the most information, which become the tagSNPs. When a population is typed with those tagSNPs, the information from the decomposition can be applied to reconstruct the missing genotypes. A drawback of this approach is that it requires exactly those tagSNPs, as defined by the singular value decomposition procedure, to be typed. Hence, when they applied it to an association study, they split the data into two: selecting tagSNPs from one half, then "assaying" those tagSNPs in the other half, and then imputing the untyped SNPs.

The advantage of methods such as fastPHASE and Margarita is that they can be applied to data with any pattern of missingness. Therefore, in this chapter, I compare the imputation performances of these most flexible methods.

Just prior to submission of this thesis, Servin and Stephens (2007) published a method that tackles the problem of imputing and testing untyped loci. They used fastPHASE to impute untyped loci in quantitative trait association studies, and then tested the imputed loci directly using a Bayesian regression approach. The advantage of using Bayesian regression

is that it naturally handles the uncertainty in imputed values. They found that compared to single SNP ANOVA tests and a linear regression test on tagSNPs only, the approach of imputing and testing with Bayesian regression appears to have greater power to detect rare variants.

The Wellcome Trust Case Control Consortium (2007) describes a genome wide association study for seven diseases, typed for 469,557 SNPs in 2,000 cases each, and using 3,000 shared controls. An multilocus method (Marchini et al., 2007) based on Li and Stephens (2003) was used to impute data at over 2 million HapMap SNPs not typed in the study, and these were tested for association. In one of the peaks associated with Type 2 Diabetes, an untyped SNP (imputed from the Phased II HapMap) was identified with stronger significance that the surrounding typed SNPs.

6.3 Imputing Missing Genotypes and Untyped Loci, and Testing for Association

As described in Chapter 2, missing data can be imputed using inferred ARGs. Missing alleles are imputed when two compatible sequences coalesce, where one sequence has an observed allele at the position which is missing in the other sequence. The missing position takes the allele of the other sequence, and this assignment is propagated down the ARG to the leaves.

I used two ARG strategies for imputing data:

- Margarita-full constructs ARGs for all the individuals in the data together. This approach can always be used to impute genotypes which are observed in some of the individuals.

- Margarita-one is only used when imputing loci that are untyped in a sample, by merging in a more densely typed sample. Rather than constructing ARGs for all the individuals together, ARGs are inferred for all the densely typed individuals and only one of the sparsely typed individuals at a time. The motivation for this is that additional individuals from the sparse data do not contribute to imputation at the untyped loci.

In order to obtain a single "best" estimate for missing data, I infer multiple ARGs and take the most frequently imputed genotype at each position as the consensus imputation. Where I compare the imputation accuracy of MARGARITA's consensus imputation with that from fastPHASE (Servin and Stephens, 2007), I use fastPHASE's default parameters.

However, for association testing, it is important to handle the uncertainty in genotype imputation, and I therefore analyse each of the ARG imputations rather than only the consensus imputation, and incorporate the uncertainty in imputation by using the methodology of Multiple Imputation (MI) (Rubin, 1987; Little and Rubin, 1987; Cordell, 2006; Souverein et al., 2006; Dai et al., 2006; Mensah et al., 2007).

MI is a simulation-based approach in which missing data are imputed multiple times, to give a number of complete data sets. Each complete data set is then analysed using some standard method. The statistic from the analysis is averaged over the imputations to give a single estimate, and the within- and between- imputation variances of the estimate are calculated. It is then possible to calculate significance for the estimated statistic (Rubin, 1987).

Specifically, I inferred $k = 30$ ARGs for each data set, giving k imputations. The choice of 30 was selected by considering the variance in the estimated odds ratios for imputed loci, which is given below. To test an imputed locus for association I calculated the log odds ratio $\hat{L}_i$ using the genotypes in imputation i. I then took the mean log odds ratio over the k imputations,

$$\bar{L} = \frac{1}{k} \sum_{i=1}^{k} \hat{L}_i,$$

to obtain an estimated log odds ratio for that locus. For imputing untyped loci in the Yeager et al. (2007) data, discussed later, $k = 30$ imputations gives a standard error of the mean, $\bar{L}$, of, on average, 9% of $\bar{L}$ (range 0.1% to 20%).

Note that the denser samples from which the missing loci are imputed are not included in calculating the log odds ratios.

In order to calculate confidence intervals and P-values for these estimated log odds ratios, the additional variance due to imputation uncertainty must be taken into account. I used

Rubin's rules to do this (Rubin, 1987), which combine the within-imputation variances (the variance for each $\hat{L}_i$), with the between-imputation variance (the variance of the sample mean $\bar{L}$). The estimated log odds ratio is distributed as a Student's t-distribution, with degrees of freedom and variance determined by the variance components just described. Testing for departure from the null hypothesis of no association is then achieved in the usual way, by comparison to this distribution. Rubin's rules are as follows:

The average within-imputation variance is

$$\bar{W} = \frac{1}{k}\sum_{i=1}^{k} W_i,$$

where the variance, W_i, for a log odds ratio, $\hat{L}_i$, is the sum of the reciprocals of the counts in each cell in the contingency table.

The between-imputation variance is

$$B = \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{L}_i - \bar{L}\right)^2,$$

and then the total variability associated with the estimate $\bar{L}$ is

$$T = \bar{W} + \frac{k+1}{k}B.$$

Then, for significance testing against $\bar{L} = 0$ and confidence interval estimation,

$$\frac{\bar{L}}{\sqrt{T}} \sim t_v$$

where the degrees of freedom for the t-distribution is

$$v = (k-1)\left(1 + \frac{1}{k+1}\frac{\bar{W}}{B}\right)^2.$$

MI approaches have already been used in two ways in association studies: First, to handle the uncertainty in haplotype phase when testing for haplotype specific effects (Cordell, 2006;

Mensah et al., 2007); and second, to handle the uncertainty in imputed genotypes (but not entirely missing loci) when testing for single SNP effects (Souverein et al., 2006; Dai et al., 2006). In Cordell (2006) and Mensah et al. (2007) haplotypes are estimated multiple times from the data, and disease model parameters are estimated by taking the mean of the estimates derived from each of the haplotype estimations. In Dai et al. (2006) three missing genotype imputation methods are compared for SNP data; missing genotypes were imputed multiple times and odds ratios calculated by MI.

In the experiments described below I compare my imputation approach to one where only the loci typed in the case-control sample are tested. The association signal at a typed locus is determined by calculating the log odds ratio from the observed data, and P-values calculated by comparison to a Normal distribution.

I performed experiments to test (1) the accuracy of missing genotype imputation; (2) the accuracy of untyped locus imputation; and (3) whether additional insights can be gained for fine-mapping.

6.4 Results for Imputing Missing Genotypes

To test the imputation accuracy for missing genotypes, I used two data sets:

- ASH. 163 Ashkenazi controls and 293 cases from a case-control study of association between a 10Mb region of chromosome 20 and Type 2 Diabetes (Barroso et al., 2007), typed with an average density of 1 SNP/2.5kb.
- NBS. 400 UK controls from the UK National Blood Service Control Cohort (The Wellcome Trust Case Control Consortium, 2007). I used chromosome 20, where there is an average density of 1 SNP/5kb.

The experiments were parameterised according to the population (ASH controls, ASH cases and controls, or NBS), and proportion of genotypes removed at random from the data (1%, 5%, 10% or 20%). For each parameterisation, 50 experiments were simulated, each one involving approximately 1Mb (400 SNPs for ASH, 200 SNPs for NBS) from a randomly chosen region, with the specified fraction of genotypes removed. Each data set had some

Population	% missing	MARGARITA-FULL	FASTPHASE
ASH controls	1	**0.019 (0.018,0.020)**	0.020 (0.019,0.021)
ASH controls	5	0.021 (0.020,0.022)	0.021 (0.020,0.022)
ASH controls	10	0.022 (0.021,0.023)	0.022 (0.021,0.023)
ASH controls	20	0.026 (0.025,0.027)	**0.025 (0.025,0.026)**
ASH cases and controls	1	**0.011 (0.011,0.012)**	0.017 (0.016,0.018)
ASH cases and controls	5	**0.013 (0.012,0.013)**	0.020 (0.019,0.021)
ASH cases and controls	10	**0.013 (0.012,0.013)**	0.019 (0.019,0.020)
ASH cases and controls	20	**0.017 (0.017,0.018)**	0.024 (0.023,0.025)
NBS	1	0.048 (0.044,0.051)	**0.034 (0.031,0.036)**
NBS	5	0.049 (0.047,0.052)	**0.036 (0.034,0.038)**
NBS	10	0.051 (0.048,0.054)	**0.036 (0.034,0.038)**
NBS	20	0.057 (0.054,0.060)	**0.038 (0.036,0.040)**

Table 6.1: Mean imputation error rates for missing genotypes, with standard error intervals in brackets.

missing data itself (0.8% for NBS chromosome 20 and 0.5% for ASH) which was imputed but not assessed.

MARGARITA-FULL and FASTPHASE were applied to these data sets, and their imputations compared to the held out observed genotypes. Table 6.1 reports the mean imputation error rate (the number of incorrect genotype imputations divided by the number of removed genotypes) for each parameterisation.

As can be seen from Table 6.1 and Figure 6.1 both methods perform better on the ASH data (with error rates of 1-3%) than on the NBS data (error rates 3-6%). There may be a number of reasons for this. First, the ASH data is typed more than twice as densely. Second, the markers in the ASH data were chosen to be non-redundant ($r^2 < 1$), whereas the NBS markers were not chosen with such a strong tagging requirement. Third, the Ashkenazi population is smaller and more homogeneous than the UK population; this can potentially give LD that extends over longer regions.

On the ASH data, the imputations are more accurate for the cases and controls combined than for the cases alone. This improvement is likely due to the increased sample size and increased homogeneity within the cases.

I also tested whether differences in the allele frequency spectrum between the ASH and NBS populations could affect imputation error rate. Figure 6.2 shows the minor allele fre-

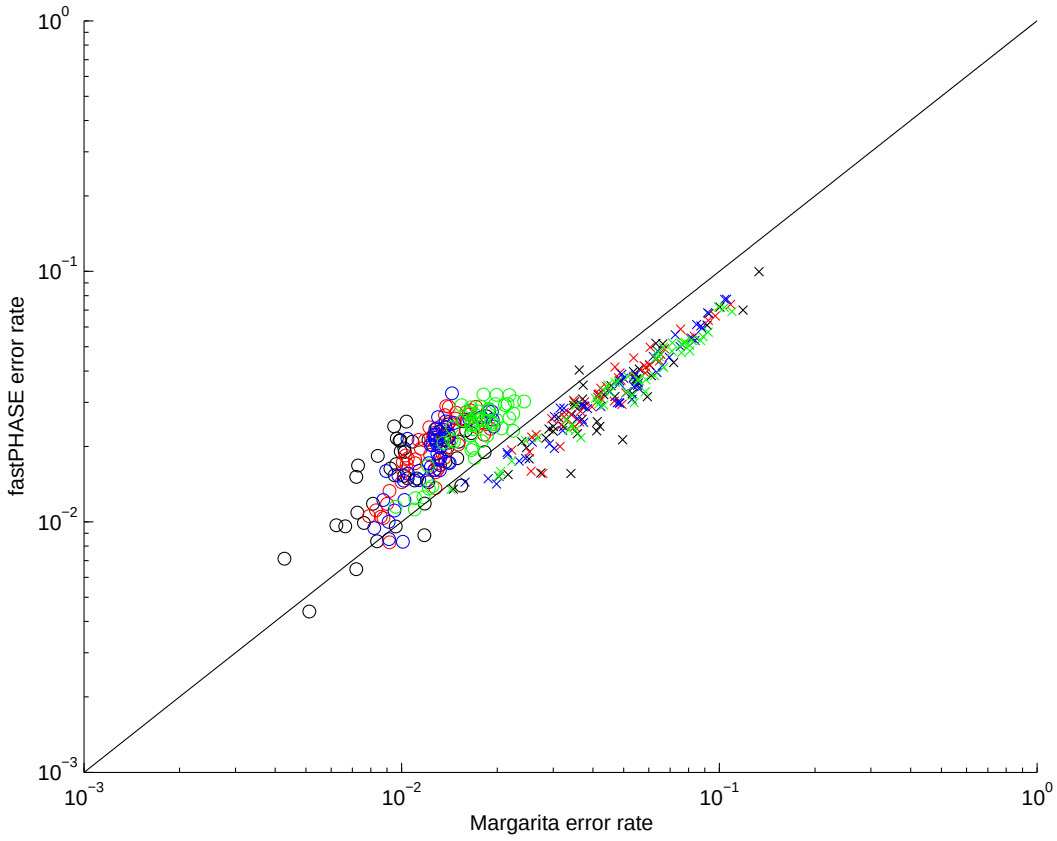

Figure 6.1: Error rates for imputation of genotypes removed at random from the data. Crosses are NBS, circles are ASH. Black is 1% missing data, red is 5%, blue is 10%, green is 20%.

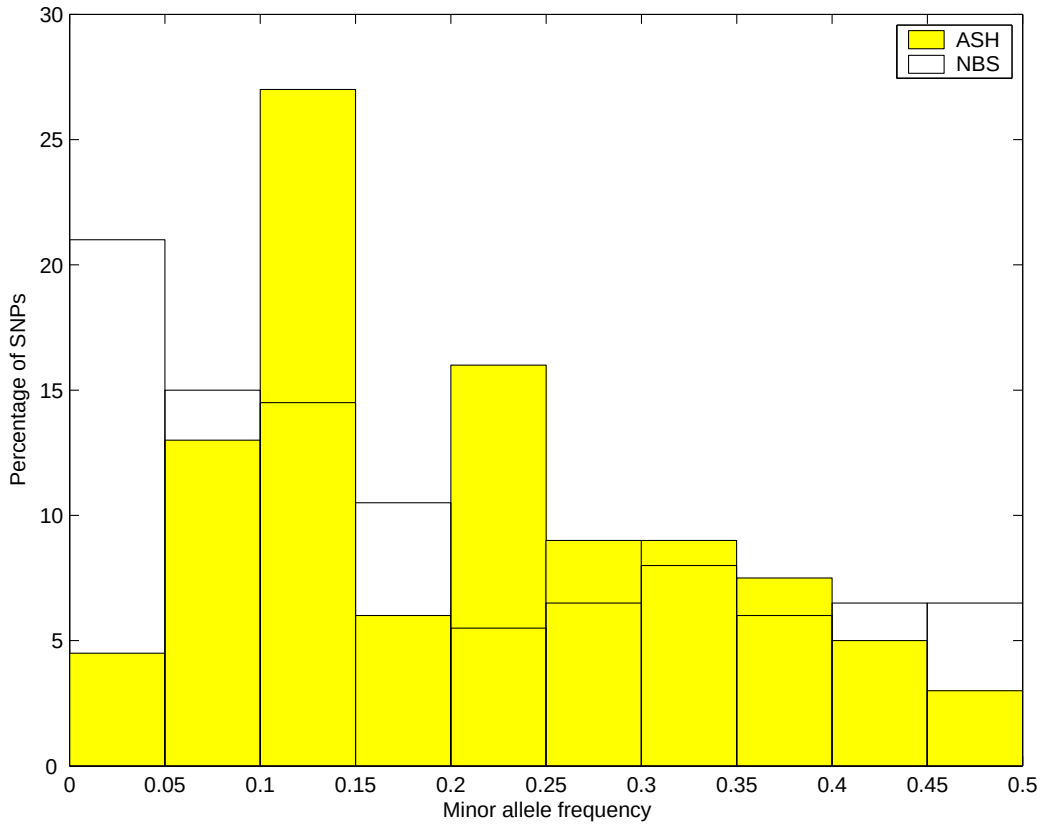

Figure 6.2: Minor allele frequency spectrum for a randomly chosen 1Mb of the ASH controls and NBS data.

quency spectrum for the two populations, and shows that the NBS has more rare alleles than the ASH controls. MARGARITA-FULL typically outperforms FASTPHASE on the ASH data, while FASTPHASE outperforms MARGARITA-FULL on the NBS data. One explanation may be that the NBS data has more rare alleles than the ASH data. We will observe in the next section that the ratio of MARGARITA-ONE error to FASTPHASE error has a slight tendency to be greater for less frequent alleles.

6.5 Results for Imputing Untyped Loci

In order to test how well loci which are untyped in one sample can be imputed from a more densely typed sample, I combined the NBS and HapMap data sets. For each simulation, a SNP typed in both samples was removed entirely from the NBS data, but kept in the HapMap data. I then took 200 NBS and/or HapMap SNPs either side of the removed locus,

Population	MARGARITA-FULL	MARGARITA-ONE	FASTPHASE
CEU+NBS	0.060 (0.049,0.071)	**0.058 (0.047,0.070)**	0.097 (0.081,0.114)
CEU+YRI+NBS	0.087 (0.072,0.102)	0.087 (0.071,0.103)	0.087 (0.068,0.106)
YRI+NBS	0.193 (0.159,0.227)	**0.173 (0.142,0.204)**	0.250 (0.211,0.289)

Table 6.2: Mean imputation error rates for untyped loci with standard error intervals in brackets.

corresponding to around 300-600kb, for 400 NBS individuals and the unrelated and unphased CEU and/or YRI HapMap. 50 of these experiments were performed using MARGARITA-FULL, MARGARITA-ONE and FASTPHASE. An additional 450 CEU+NBS experiments were performed using MARGARITA-ONE only, in order to gain further insight into how imputation performance varies with minor allele frequency at the untyped locus.

Table 6.2 gives the error rates for untyped locus imputation.

Using the CEU HapMap to impute missing loci gives a much lower error rate than using the YRI HapMap. This is not surprising; the NBS is a European population, and the CEU is a European-derived population, whereas the YRI is an African population. This result suggests that ensuring reasonable matching between the dense and case-control samples is important for accurate imputation.

When imputing using the CEU and YRI samples together, FASTPHASE achieves its best performance. The larger sample size may help. However, the performance of MARGARITA-ONE is intermediate between the NBS+CEU and NBS+YRI configurations.

Both MARGARITA-FULL and MARGARITA-ONE tend to perform better than FASTPHASE, with MARGARITA-ONE marginally better than MARGARITA-FULL.

Figure 6.3 shows how the relative error rate of MARGARITA-ONE and FASTPHASE varies with minor allele frequency at the missing locus. The relative error on the y-axis is the $\log_{10}$ of the MARGARITA-ONE error rate divided by the FASTPHASE error rate, for each of the 50 CEU+NBS experiments. There is a slight (non-significant) tendency for the ratio of MARGARITA-ONE error rate to FASTPHASE error rate to be lower for loci with greater minor allele frequency (for 50 CEU+NBS experiments, Spearman's rank correlation coefficient -0.07, $P = 0.64$).

Figure 6.4 gives the MARGARITA-ONE error rate versus minor allele frequency of missing

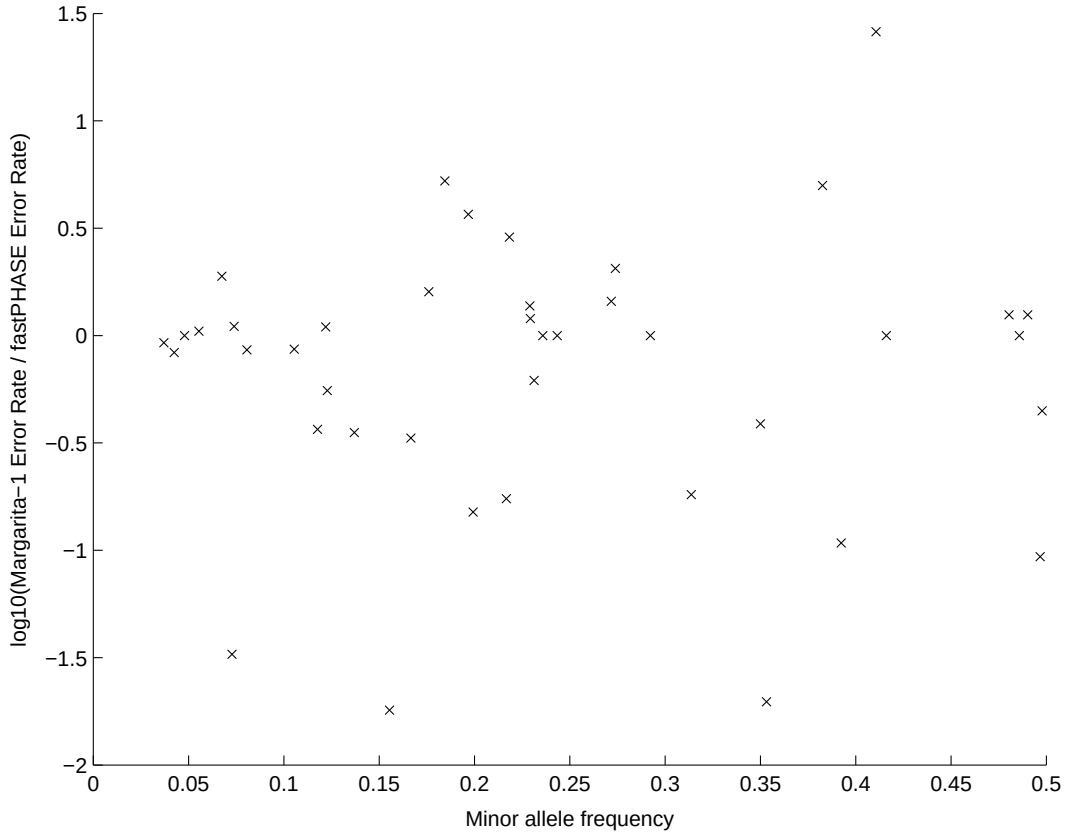

Figure 6.3: MARGARITA-ONE versus FASTPHASE error rates for imputation on 50 NBS+CEU data sets. Relative performance is plotted against minor allele frequency.

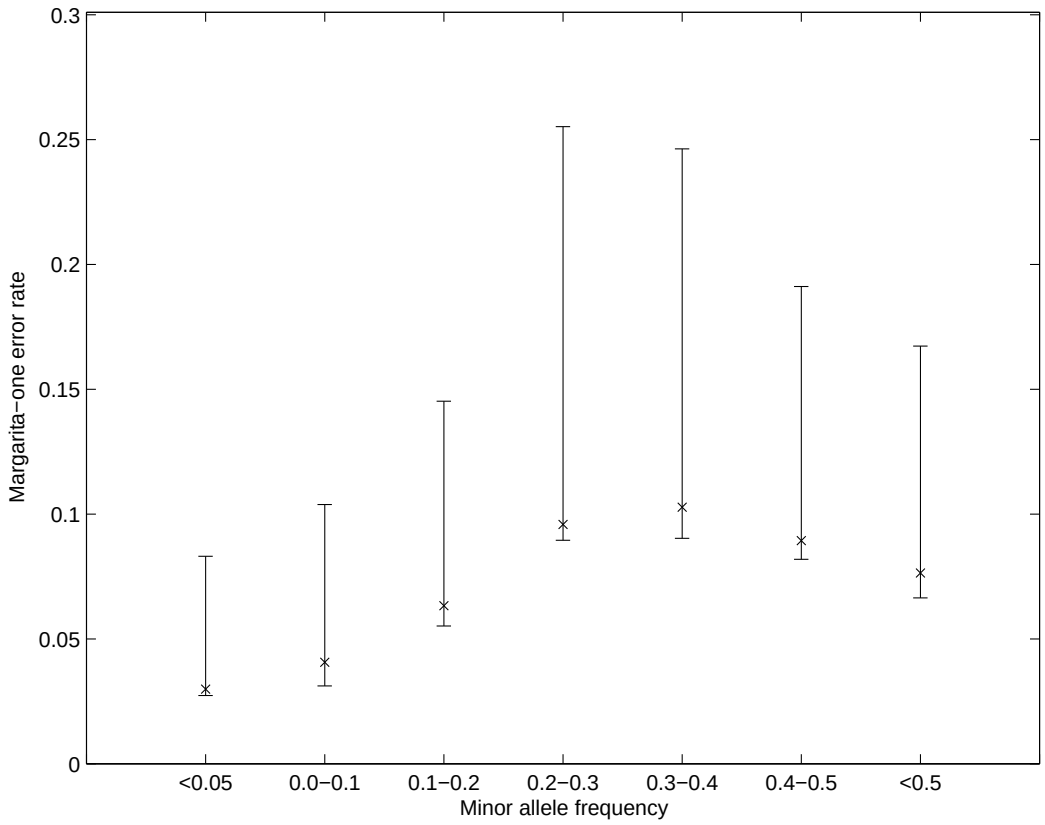

Figure 6.4: MARGARITA-ONE error rate (mean and inter-quartile range) versus minor allele frequency of imputed loci in 500 NBS+CEU experiments.

loci for 500 CEU+NBS experiments. The imputation error rate is lower for loci with lower minor allele frequencies (Spearman's rank correlation coefficient 0.11, P = 0.02). This may be expected, because there is inherently more uncertainty in the value of a genotype with greater minor allele frequency.

6.6 Results for Imputation of Untyped Loci in 8q24

A data set containing the European American samples analysed in Chapter 5 (Yeager et al., 2007) and the CEU HapMap was provided by G. Thomas of the NIH National Cancer Institute. This data set contains the 1169 prostate cancer cases and 1094 controls, and the 60 CEU parents, covering 1Mb of 8q24, with 250 SNPs typed in both the Yeager et al. (2007) sample and the HapMap, and 898 SNPs typed in the HapMap alone.

In order to test how well the association signal for an imputed locus matches the true

association signal, I constructed 250 experiments: one for each SNP typed in both the Yeager et al. (2007) sample and the HapMap. This was done by removing that SNP from the Yeager et al. (2007) data, but keeping it in the HapMap sample. 200 SNPs either side of the removed locus were then taken, corresponding to 300-600kb, and presented to MARGARITA-ONE.

The mean imputation error rate is 0.1213 and the median is 0.0898.

The P-values for the imputed log odds ratios and the observed log odds ratios are shown in Figure 6.5. These results show that loci which are significant when observed tend to be significant when imputed, and that imputation does not create false positives.

I also took the full 1Mb of data and imputed all loci typed in the HapMap, but untyped in the Yeager et al. (2007) sample. The results for all imputed loci are shown in Figure 6.6.

Table 6.3 gives the imputed odds ratios for SNPs untyped in the Yeager et al. (2007) sample but typed and found significantly associated across multiple populations in either Haiman et al. (2007) or Gudmundsson et al. (2007). Along with the imputed odds ratios, the table also gives the odds ratios reported in Haiman et al. (2007) and Gudmundsson et al. (2007). It should be noted that different samples are used between studies, which may have different allele frequencies and vary in size and population. Therefore, I only report the odds ratios for European-derived populations.

The SNPs rs13254738, rs6983561 and rs16901979 correspond to the most centromeric association signal found in Haiman et al. (2007) and Gudmundsson et al. (2007), for which there appears to be no significant evidence of association in our data set, imputed or observed.

However, rs13254738 and rs6983561 are not significant when only the European American population in Haiman et al. (2007) is considered, thus the most likely explanations for lack of imputed significance is that these SNPs are not causative, or have lower allele frequencies, or weaker effects in European-derived populations. Indeed, the odds ratios for these SNPs in the observed Haiman et al. (2007) European American data and the imputed odds ratios show rough agreement (Table 6.3).

Gudmundsson et al. (2007) report rs16901979 as being significant across multiple European populations; however, I do not find an imputed significant signal. This is most likely due to lack of power of the imputation approach, caused by there being only two copies of

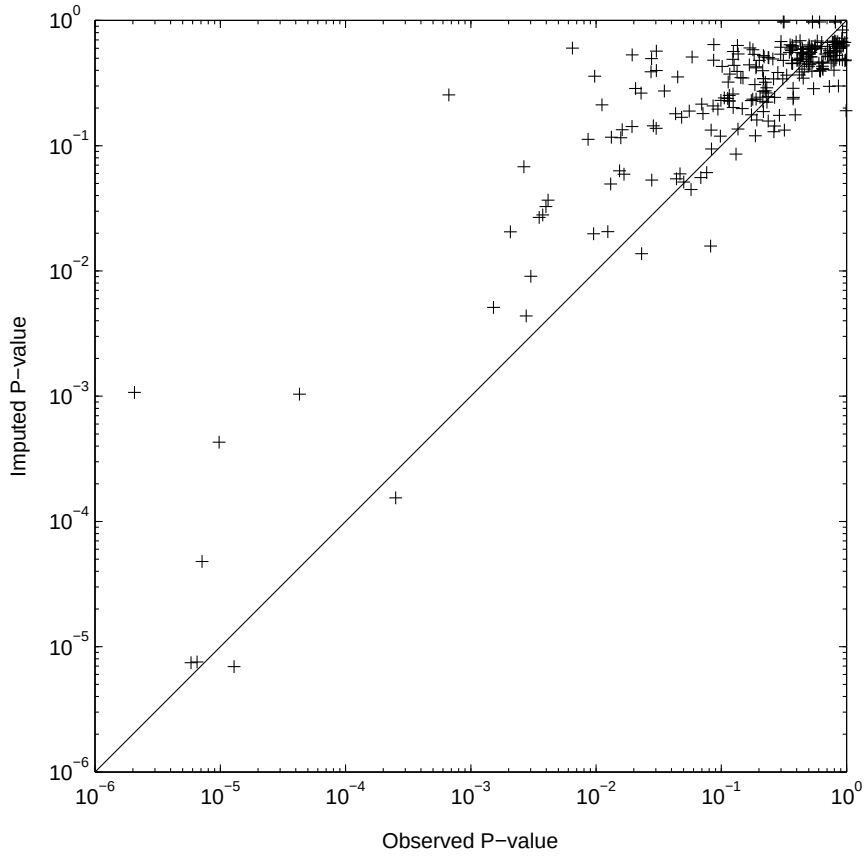

Figure 6.5: Observed association test P-values compared with those obtained by imputation, for SNPs typed in the Yeager et al. (2007) sample.

Study	SNP	Study OR	Imputed OR
Haiman et al.	rs13254738	1.11 (0.97-1.26)	1.06 (0.92-1.21)
Haiman et al.	rs6983561	1.16 (0.86-1.58)	1.22 (0.85-1.76)
Gudmundsson et al.	rs16901979	1.79 (1.53-2.11)	1.17 (0.85-1.63)
Haiman et al.	rs7000448	1.14 (0.98-1.40)	1.20 (1.06-1.36)
Haiman et al.	rs10090154	1.44 (1.17-1.76)	1.50 (1.24-1.81)

Table 6.3: Imputed odds ratios (ORs) and 95% confidence intervals for SNPs untyped in the Yeager et al. (2007) sample but which were found to be significant across multiple populations in other studies. Study ORs are from the European populations only (European Americans in Haiman et al. (2007) and all European populations combined in Gudmundsson et al. (2007)).

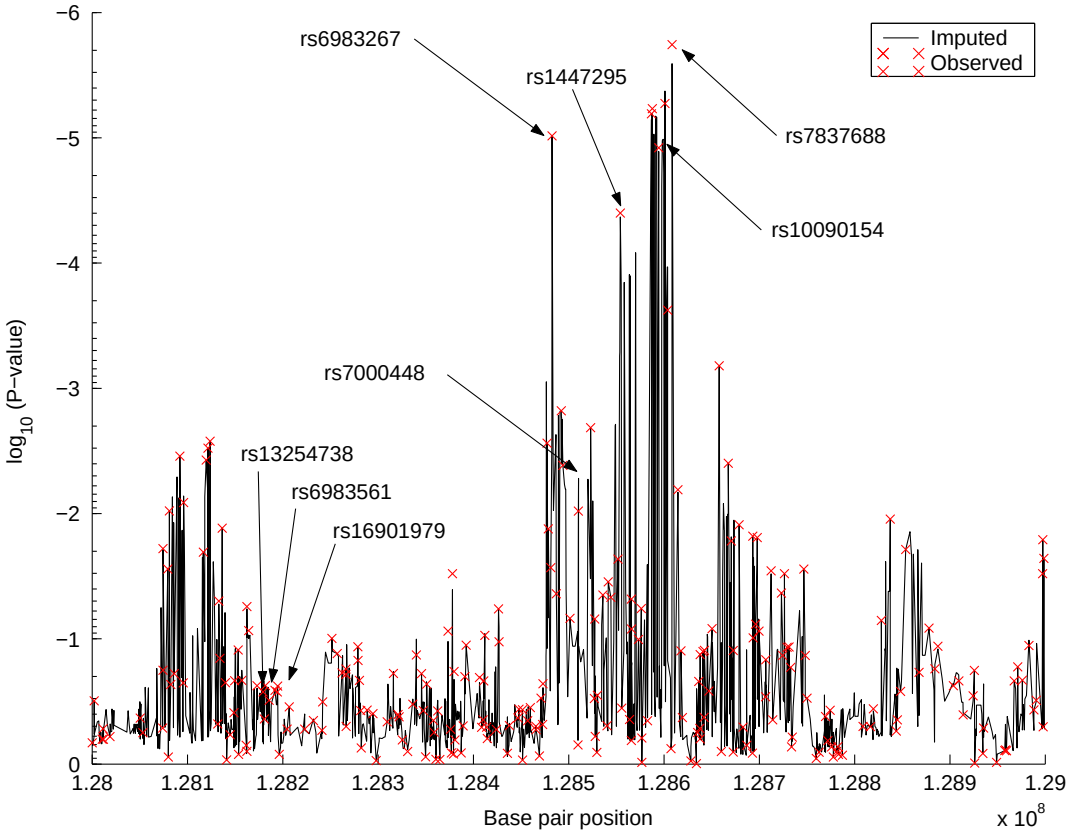

Figure 6.6: Imputed P-values for the Yeager et al. (2007) data. rs13254738, rs6983561, rs16901979, rs7000448 and rs10090154 are untyped in the Yeager et al. (2007) sample, but were found to be associated with prostate cancer in either Haiman et al. (2007) or Gudmundsson et al. (2007). rs6983267, rs1447295 and rs7837688 are typed in the Yeager et al. (2007) sample.

the rs16901979 risk allele in the CEU sample from which I impute.

The imputed association at rs7000448 has a greater odds ratio than that given in Haiman et al. (2007), where this association was originally detected. The significance at this imputed locus is SNP-wise P=0.005, which is suggestive of association. This SNP is in a region of high recombination and so could correspond to a causative polymorphism distinct from the association peaks either side of it (rs6983267 and rs1447295 in Figure 6.6).

The imputed SNP rs10090154 has strong significance (SNP-wise $P = 3.3 \times 10^{-5}$). This was previously detected in the Haiman et al. (2007) study, and is located within the telomeric peak of the observed Yeager et al. (2007) data. Many of the imputed and typed SNPs in the proximity of rs10090154 show strong significance, and are likely to correspond to the same causative variant. In the CEU HapMap, rs10090154 has $r^2 = 1$ with rs1447295 and rs7837688, which are typed in the Yeager et al. (2007) sample and are highly significant.

Although no new association peaks are found, the imputed data does show additional structure around rs1447295. The imputed odds ratios for untyped SNPs around rs1447295 are stronger than for rs1447295 itself, although after accounting for the increased variance due to imputation, the P-values for these SNPs are less significant. The fact that many of the imputed loci have greater odds ratios suggests that rs1447295 may be in LD with the causative variant(s) and not itself causal.

6.7 Discussion

Rather than testing each branch of inferred ARGs for disease association, a more direct approach is to test only those branches that correspond to segregating sites. I have done this by inferring ARGs for case-control data combined with more densely genotyped controls.

The performance of MARGARITA for imputing missing genotypes and untyped loci is in general comparable to that of FASTPHASE, with each appearing better in some conditions. In Scheet and Stephens (2006), FASTPHASE achieves an error rate of 3-4% on CEU HapMap data where genotypes are removed at random. In my experiments on more individuals less densely typed, similar error rates are achieved. Locus imputation, however, is harder, and

the error rates tend to be above 5%, which is likely to have a significant impact on the power of the imputation approach to detect true associations.

Because I am using HapMap samples to impute loci in cases and controls we might expect the association signals to be deflated, since the cases are being imputed from a relatively small sample which may not be ideally matched. I observed that the error rate for imputing loci in cases and controls from the Yeager et al. (2007) sample is greater than that for imputing loci in the NBS controls. Therefore, one way to increase the imputation accuracy at untyped loci may be to densely genotype (or in the future, resequence) case and control individuals from the case-control study, rather than using an external generic sample. This also suggests that a larger densely typed sample, e.g. a larger HapMap, would help imputation based approaches.

Another scenario in which this method may be valuable is where different case-control studies are combined, involving individuals from the same or related populations but genotyped on different platforms, and thus with mostly different typed SNPs. However, I do not consider this scenario here.

Analysis of the prostate cancer data in 8q24 of Yeager et al. (2007) suggests that untyped locus imputation may be useful for identifying additional association structure. For most SNPs that have been reported as associated in other publications, the imputation approach makes fairly accurate estimates of the odds ratios. One SNP which was found to be significant in Gudmundsson et al. (2007) was not imputed as significant in my experiment; however this SNP has minor allele frequency 1.7% in the CEU HapMap sample. This again suggests that a much larger set of densely typed individuals would be valuable.