Chapter 3

Investigating new reference and target sets in genotype imputation

3.1 Introduction

Genome-wide association studies (GWAS) are based on a tag SNP approach. Genotyping arrays use a set of SNPs chosen such that, between them, they are correlated with most of the common variants in the human genome. Any common causal variants will be then be well correlated with at least one SNP on the array, and (providing a large enough sample size is genotyped) such associations can be detected via signals at these tag SNPs.

While tag SNP sets are picked using a high-density reference set, the approach of testing these tag SNPs for association in a GWAS cohort makes no assumptions about what untyped SNPs are being tested. However, it

is possible to use the data in the reference set to improve the coverage of the study. The reference set tells us (at least some of) the common SNPs that exist, and allows us to place them together into multi-SNP haplotypes. We can therefore use the tag SNPs we have genotyped to match the haplotypes in our GWAS samples to haplotypes in our reference set, and use this matching to infer these samples' genotypes at other sites. This process is called "genotype imputation", and we refer to the dataset we are predicting genotypes for as the "target set".

Genotype imputation has a number of advantages over tag SNP testing. Firstly, it allows meta-analyses to be performed even when the component studies have been performed using different sets of tag SNPs, by allowing a common set of SNPs to be imputed. Secondly, imputed genotypes, while only probabilistically predicted, are imputed using information from many surrounding SNPs, and thus are often more strongly correlated with the true genotype than any single tag SNP. This gives improved power to detect associations, especially for variants that are not well tagged by the array, and can lead to significant associations being detected that would have been missed otherwise (Huang et al., 2012). Thirdly, it allows test statistics to be produced at all sites in the reference set, which (if the reference set is high enough density) is likely to contain the true causal variant, and thus can allow the function of associated variants to be inspected.

3.1.1 Overview of imputation software and methods

The vast majority of the human genome is diploid, meaning that it is made up two copies. Each copy contains its own set of alleles, which together make up the two multi-marker haplotypes that an individual carries. To perform the haplotype matching that genotype imputation relies on, we first need

to reconstruct these two haplotypes from the diploid genotypes produced by genotyping chips, by determining the phase of the alleles at each site (i.e. inferring which alleles are present on the same copy of the chromosome). This process is known as "phasing", and is the most statistically challenging aspect of imputation. The history of imputation is therefore, to a first approximation, a history of phasing techniques.

Experimental and family-based phasing techniques are as old as genetics itself, but statistical phasing techniques began being applied in the 1990s (Browning and Browning, 2011). The first statistical imputation method for unrelated individuals was the Clarke algorithm published in 1990 (Clark, 1990), which inferred the existence of haplotypes based on parsimony. Soon after methods based on Expectation-Maximisation (EM) were developed to estimate haplotype frequencies and phase small numbers of SNPs. Both of these methods are computationally expensive and relatively inaccurate, and thus did not generalise outside of small haplotype blocks (Browning and Browning, 2011). The EM method is still in use, however, for instance in the imputation function of the popular statistical genetics toolkit Plink (Purcell et al., 2007).

Most modern phasing and imputation methods are based on approximate coalescent techniques. Coalescent theory was developed in the 1980s as a way of linking population genetics to genealogy at a single gene or site (Kingman, 1982), and was extended in the 1990s to include recombination (Griffiths and Marjoram, 1996). Because coalescent theory models both polymorphism frequency and stretches of the genome shared by descent it is particularly well suited to modelling haplotype frequencies. While full coalescent theory is computationally difficult to apply in most circumstances, approximate methods have been developed that are computationally

tractable (McVean and Cardin, 2005). The most widely used approximation is the Li and Stephens model (Li and Stephens, 2003), which partitions the coalescence likelihood into a series of sequential conditional approximations, which are in turn calculated using a Hidden Markov Model that includes recombination and mutation.

The first piece of software to use the approximate coalescent was PHASE (Stephens et al., 2001). A faster technique, fastPhase (Scheet and Stephens, 2006) (also implemented in BIMBAM (Servin and Stephens, 2007)), was introduced in 2006; this was also the first software to perform genotype imputation *per se*. Other imputation programs using the same approach include IMPUTE (Marchini et al., 2007), IMPUTE2 (Howie et al., 2009), MaCH (Li et al., 2010) and SHAPEIT (Delaneau et al., 2012).

Not all imputation programs are based on an approximate coalescent model. The imputation program Beagle (Browning and Browning, 2007, 2009), while also based on a Hidden Markov Model approach, does not explicitly model mutation and selection, instead using a haplotype clustering model to perform phasing and imputation (Browning, 2006). In contrast, QCall (Le and Durbin, 2011) (an imputation program for sequencing data) performs imputation by directly fitting mutations to a sequence of sampled ancestral recombination graphs.

3.1.2 New reference and target sets in imputation

Imputation methods in GWAS originally used the HapMap phase 2 reference set (Frazer et al., 2007), which contained data on 400 haplotypes from three ethnic groups. This served as a successful reference set for common SNPs in Europeans for the first wave of GWAS, allowing around 75% of common SNPs to be imputed with accuracies of above 98% (Marchini et al., 2007),

and allowing meta-analysis of studies from different technologies (Zeggini et al., 2008).

However, in the last five years reference sets have developed substantially. The HapMap phase 3 expanded the dataset in the sample direction to include data from many populations, with five times the total number of samples as HapMap phase 2 (Altshuler et al., 2010). The 1000 Genomes pilot reference set expanded in the marker direction with 16M SNPs, indels and structural variants (Project, 2010), and the 1000 Genomes phase 1 reference set includes an unprecedented 40M SNPs on 1092 samples (Project, 2012), including data from genotyping chips, exome and whole genome sequencing. Many of the newly discovered variants are low frequency (MAF $< 5\%$). We have a far less detailed understanding of how well these new variants can be imputed, and how the changes in reference set will impact imputation.

Likewise, many of the original GWAS that used imputation were carried out on individuals of European descent. However, many important GWAS in recent years have been performed using sample collections from Africa (The MalariaGEN Consortium, 2009; Thye et al., 2012; Akinsheye et al., 2011). As we will discuss later in this chapter, these African populations tend to have a greater diversity (both within and between populations). They also have a lower correlation (linkage disequilibrium, or LD) between markers, and the patterns of LD tend to differ between populations. As a result, genotype imputation in these populations is more complicated and less well understood.

In this chapter I will investigate how changes in reference and target sets impact imputation. This will show how new reference sets allow us to use genotype imputation to fill gaps that old imputation reference sets left. This includes imputing variants at low frequency, and variants from specific

functional classes. It will also include imputation into populations where imputation has traditionally had more difficulty, such as African populations.

I will start by studying the impact of sample set and diversity on imputation of common and low frequency variation in Europeans, using HapMap imputation. I will then report two studies of imputation in Africa, including an investigation of HapMap imputation for GWAS meta-analyses, and the use of 1000 Genomes imputation in a single diverse population. Finally, I will discuss how these new imputation reference sets can be used to give us new biological insight into the relationship between variant function and disease association, by allowing us to impute loss-of-function variants into GWAS cohorts.

3.2 The impact of reference set diversity in Europeans

This section describes a study that I carried out and published (Jostins et al., 2011) in the first year of my PhD. The reference sets and software versions used are therefore largely out of date at the time of writing this thesis. However, the broader leasons learned about reference set diverse and genotype imputation are nonetheless still valid.

The HapMap phase 2 reference panel consists of genotype data from three homogeneous populations, with 120 haploid genomes each of European and African origin, and 180 of East Asian origin, genotyped at over 2 million sites. By contrast, the larger HapMap phase 3 (or HapMap3) reference set (Altshuler et al., 2010) is much larger, containing over 1000 samples genotyped at a restricted set of approximately 1.5 million variants. Unlike the HapMap2, this data is drawn from a set of 11 populations, providing a far more diverse dataset. Additionally, the HapMap3 benefits from a more mature genotyping technology, providing higher genotype quality. Taken together, these two HapMap datasets provide a significant and stable set of test data to investigate the impacts of the reference set on imputation quality.

I investigate the relationship between sample size and ancestry and imputation accuracy by comparing results obtained using HapMap2 and HapMap3 as the reference set. My comparative analysis focuses on three areas: (1) what effect does the higher quality of genotyping from HapMap3 compared to HapMap2 have on imputation? (2) what improvements can the large increase in sample size have on imputation accuracy and predicted quality scores, especially for low-frequency SNPs? and (3) what can we infer about the importance of closely matching ancestry of reference and target samples?

Population	Code	HapMap2	HapMap3
African Americans	ASW	0	63
North Europeans	CEU	60	117
Chinese Americans	CHD	0	85
Gujarati	GIH	0	88
Japanese and Chinese	JPT+CHB	90	170
Luhya	LWK	0	90
Mexicans	MEX	0	52
Maasai	MKK	0	143
Toscani	TSI	0	88
Yoruba	YRI	60	155

Table 3.1: A summary of the HapMap sample sets and their sizes in the HapMap2 and HapMap3 datasets. I used release 21 of the phased HapMap2 data, and release 2 of the phased HapMap3 data.

3.2.1 Performing and Scoring Imputation

For the target set, I used 1 374 individuals from the 1958 British Birth Cohort (Power and Elliott, 2006), genotyped on both the Illumina HumanHap550 BeadChip and Affymetrix GeneChip Human Mapping 500k chips as the target set. I used the Illumina data to perform imputation, and checked the answers using the Affymetrix data (Illumina chips having been previously shown to be more powerful for imputation (Anderson et al., 2008)). For the target reference sets, I used the approximately 2.5M polymorphic SNPs of the HapMap2 CEU samples, and various mixtures of HapMap3 samples, with approximately 1.4M polymorphic SNPs. Details on the HapMap reference sets are shown in Table 3.1, and the large-scale genetic relationships between these population (measured by principal component analysis) are shown in Figure 3.1.

To perform the imputation I used the imputation program Beagle (Browning and Browning, 2007) (version 3.0.2). I split the genome up into 500kb chunks, with 250kb buffer region on each side, and ran Beagle for 10 itera-

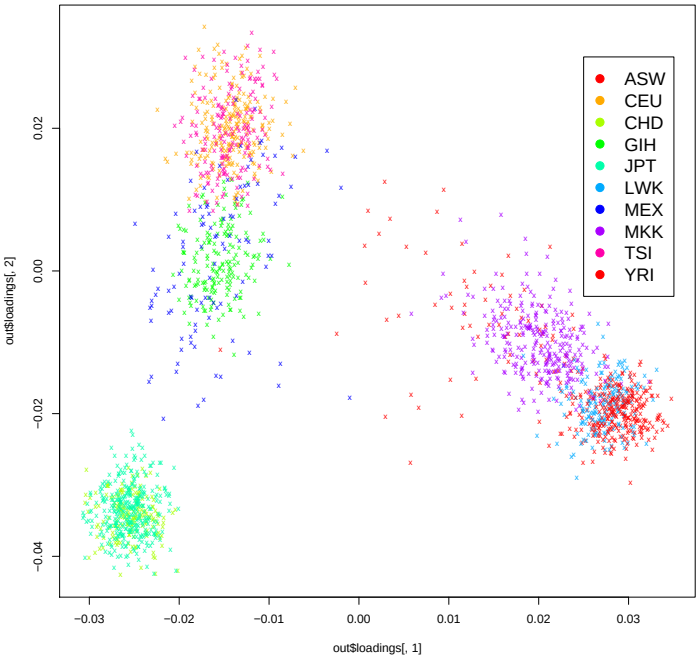


Figure 3.1: The first two principal components for each of the HapMap3 samples, coloured by population. Principal component analysis was performed on all genotypes on chromosome 17, using all founder samples.

tions. To remove poorly imputed SNPs, I applied a filter that removed SNPs with a predicted dosage r^2 of less than 0.9. For several analyses I compare common (MAF $> 5\%$) and low-frequency (MAF $\leq 5\%$) SNPs.

To score the imputation results, I measured both the accuracy of imputation and the usefulness of the predicted quality scores that the imputation method provides. Accuracy was measured using dosage r^2, defined as the square of the Pearson correlation coefficient between the imputed and the actual allele dosage across all imputed samples. The actual dosage is the count of minor alleles for each sample, and the imputed dosage is the expected minor allele count, defined as $2P(aa) + P(Aa)$, where a is the minor allele, and $P(G)$ is the posterior probability of a particular genotype. The

dosage r^2 is useful as it is not confounded by minor allele frequency, and thus can be used to compare low-frequency and common SNPs, as well as having a simple relationship to power in a GWAS (Anderson et al., 2008).

For predicted quality scores, most imputation programs (including Beagle) give a predicted dosage r^2 for each SNP, which was evaluated using four criteria: (1) the calibration, or mean difference between predicted and actual dosage r^2 (2) the quality r^2, or the correlation between predicted and actual dosage r^2, (3) the number of overconfident calls, i.e. the number of SNPs that are poorly imputed despite having high predicted dosage r^2, and, vice versa, (4) the number of under-confident calls. I am particularly interested in the number of overconfident SNPs, as these may lead to costly false positives.

3.2.2 Reference Set Quality

While the majority of SNPs in both HapMap2 and HapMap3 are of high quality, the genotyping for a number of previously poorly genotyped SNPs was improved in the development of HapMap3. To investigate whether this increase in reference set quality had a significant effect on imputation, I performed genome-wide imputation on the target set using two 'reduced' HapMap reference sets, and measured differences in dosage r^2. These reduced sets contained only the 56 CEU samples and 1M SNPs that HapMap2 and HapMap3 have in common. I found a small but significant difference due to genotyping quality (mean dosage r^2 0.841 vs 0.845, Figure 3.2), but not enough to explain a meaningful difference in imputation quality between HapMap2 and HapMap3.

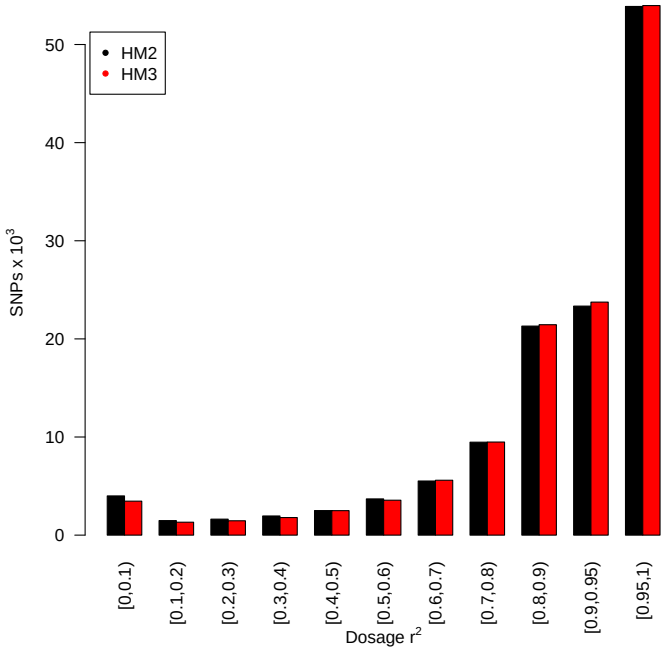


Figure 3.2: A histogram of dosage r^2 for a genome-wide imputation using the reduced HapMap2 and HapMap3 sets, which contain only the 1,069,264 SNPs and 56 CEU samples that both HapMap2 and HapMap3 have genotype information for. The means of the distributions are 0.841 and 0.845, and the difference is significant (t = 7.59, df = 256480, p $<10^{-13}$).

3.2.3 Reference Set Size

To assess the effect of larger HapMap sample sizes, I performed genome-wide imputation on the target set, using five reference sets of increasing size and diversity. I used the HapMap2 and HapMap3 CEU samples (HM2CEU and HM3CEU), which should be the best match to the UK target set, as well as a mixed reference set of HapMap3 European samples (CEU+TSI). To give a large, but still partially matched reference set, I used the HapMap3 European samples mixed with the Indian and Mexican samples (CEU+TSI+GIH+MEX), as these populations cluster together on the first two principal components (see Figure 3.1). Finally, I examined all

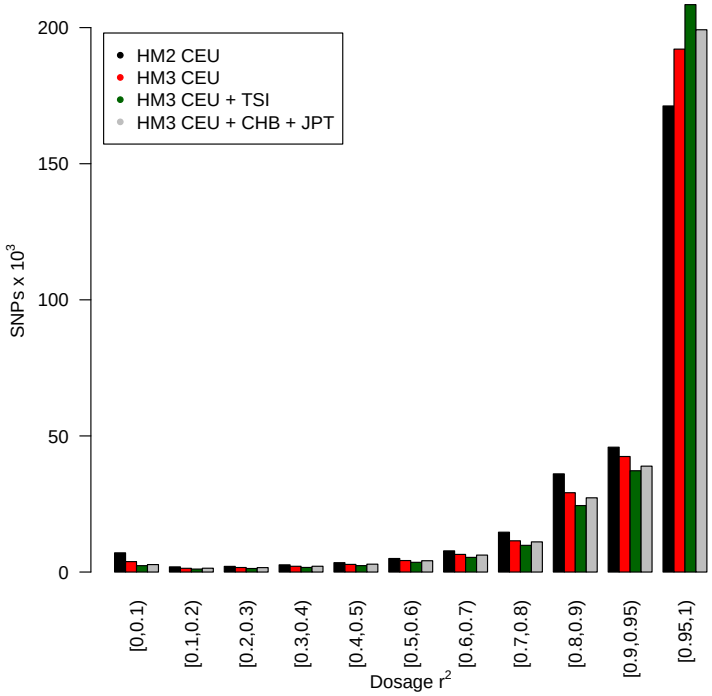


Figure 3.3: The effects of reference set on imputation accuracy. A histogram of dosage r^2 scores genome-wide for samples imputed with HapMap2 and HapMap3 CEU, as well as HapMap3 CEU+TSI, and a reference set consisting of HapMap3 CEU+JPT+CHB of the same size as the CEU+TSI set.

HapMap3 individuals (WORLD), in order to assess a very large and very diverse reference set. Sample sizes are shown in Table 3.2.

I found that HapMap3 yields a substantial increase in imputation accuracy compared to HapMap2, with the number of SNPs in the highest score category ($> 95\%$) increasing, and the number in all lower-scoring categories decreasing (Figure 3.3). A further increase in imputation accuracy is seen when adding the HapMap3 TSI samples. The number of SNPs that pass the filter (have a predicted r^2 greater than 0.9) rises as imputation accuracy increases, although this falls as samples from many populations are added due to a decrease in the imputation software's predicted confidence (see below).

Reference Set	Size	CPU	Passed Filter		Filtered Dosage r^2	
			Common	Low-frequency	Common	Low-frequency
HM2CEU	60	514h[a]	83.7%[b]	52.5%[b]	0.957	0.889
CEU	117	296h	85.1%	59.7%	0.968	0.921
CEU+TSI	205	350h	86.1%	63.1%	0.974	0.934
CEU+TSI +GIH+MEX	345	458h	85.3%	60.3%	0.978	0.957
WORLD	1010	1207h	83.8%	55.5%	0.979	0.968

Table 3.2: Information on Genome-Wide imputation using various reference sets. The CPU columns shows the number of CPU hours used in the imputation, which increases with the size and SNP density of the reference set. The proportion of SNPs that passed the filter (predicted dosage $r^2 \geq 0.9$), and the mean dosage r^2 of those that passed, are shown for common (MAF > 0.05) and low-frequency (MAF ≤ 0.05) SNPs. [a] HM2 has a large SNP set, hence the longer imputation time [b] HM2 has a larger number of SNPs in total

The dosage r^2 of filtered SNPs shows a trend of improved imputation with increasing sample sizes. This increase is statistically significant ($p < 10^{-16}$) for all increases in sample size, with the exception of the WORLD set (Table 3.2). A corresponding increase is seen in computational time, especially for the WORLD set; however, the CEU+TSI+GIH+MEX reference set only takes 55% longer to process than just CEU, despite being nearly 3 times larger.

The improvement for low-frequency SNPs is the most striking. The HM2CEU mean dosage r^2 score is low, especially compared to common SNPs (0.89 vs 0.96). If all samples from all HapMap3 populations are included, this gap nearly disappears (0.96 vs 0.98). In general, fewer low-frequency SNPs pass the imputation quality filter (63% at most), but the accuracy of these imputed low-frequency SNPs can become very high. The improvement in dosage r^2 is inversely proportional to the frequency of the SNP, with the

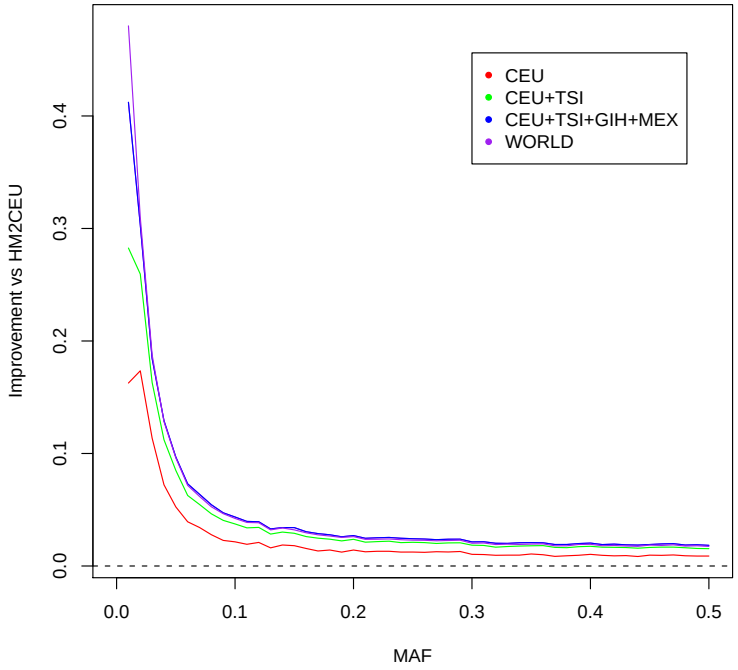


Figure 3.4: The genome-wide increase in dosage r^2 for imputed SNPs relative to HapMap2 CEU, plotted against minor allele frequency, for the four HapMap3 sample mixtures.

greatest improvement observed for the very rarest SNPs (Figure 3.4).

For small reference sets, the calibration of predicted quality scores tends towards overconfidence. As the reference set increases in size, the calibration improves, though very diverse reference sets lead the confidence scores towards under-confidence (Table 3.3). The correlation between predicted and actual dosage r^2 improves, though with a slight decrease for the most diverse sets. These trends are stronger in low-frequency variants than in common ones; low-frequency variants tend to have less well calibrated and correlated predicted quality scores. Larger reference sets decrease the number of overconfident mistakes and the number of under confident mistakes

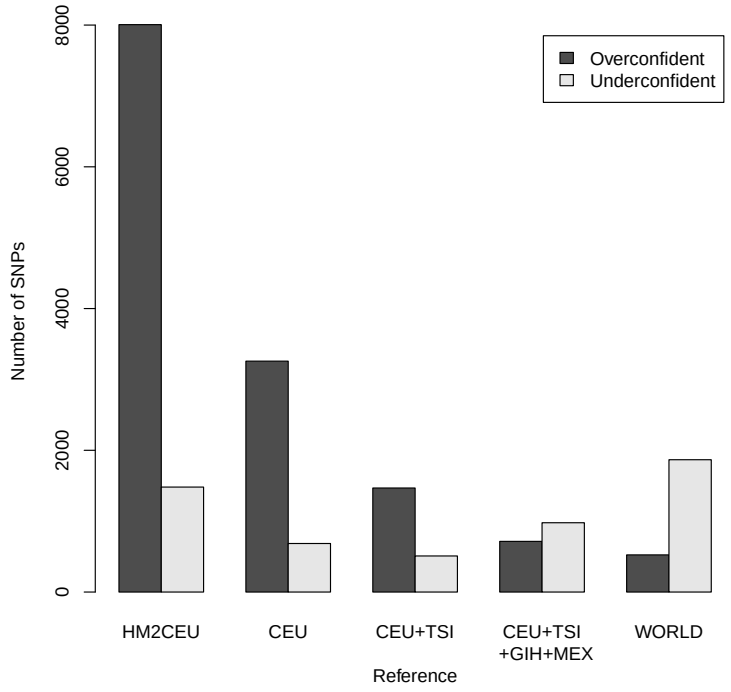


Figure 3.5: The rates of overconfident and under-confident mistakes in imputation, using various reference sets. An overconfident mistake is any SNP that is imputed with a predicted dosage $r^2 > 0.9$, but an actual dosage $r^2 \leq 0.8$, and an under-confident mistake has a predicted dosage $r^2 \leq 0.8$ and an actual dosage $r^2 > 0.9$.

(with the exception of the WORLD set, which causes a slight inflation in under-confident calls, Figure 3.5).

3.2.4 Reference Set Diversity

I investigated the importance of population matching, independent of sample size, in two ways. Firstly, I compared genome-wide imputation using the HapMap3 CEU+TSI reference set to a CEU+JPT+CHB reference set of the same size and non-CEU proportion. This allows us to investigate the effect of adding poorly matched samples on imputation. Second, I created a num-

Reference Set	Calibration		Quality r^2	
	Common	Low-frequency	Common	Low-frequency
HM2CEU	0.019	0.038	0.78	0.73
CEU	0.008	0.027	0.88	0.76
CEU+TSI	0.002	0.009	0.92	0.79
CEU+TSI +GIH+MEX	-0.006	-0.019	0.93	0.79
WORLD	-0.010	-0.043	0.91	0.76

Table 3.3: Calibration data for Genome-Wide imputation using the five reference sets. Quality calibration is defined as the mean difference between the actual and predicted dosage r^2; a negative value represents conservative quality scores, and a positive value represents liberal quality scores. The quality r^2 is the correlation between the predicted and actual r^2. The SNPs are split into common (MAF $>$ 0.05) and low-frequency (MAF $\leq$ 0.05).

ber of equally sized reference sets for chromosome 17 by combining a range of mixture proportions of either CEU and TSI , or CEU and CHB+JPT. I measured the accuracy of imputation using these reference sets for low-frequency variants. I denote these constant-sized mixed reference sets as CEU/TSI and CEU/CHB+JPT, in order to distinguish between reference sets in which sample size is not held constant (e.g. CEU+TSI).

I found that, while the mismatched CEU+JPT+CHB reference set gives a lower imputation accuracy than CEU+TSI, it still yielded a substantial improvement over the CEU reference set alone. Half of the improvement in imputation accuracy from CEU to CEU+TSI was also gained with the CEU+JPT+CHB reference. This implies that while matching the reference set to the target set is important, even the addition of unrelated samples yields increases in imputation accuracy.

Increased diversity initially correlates with increased imputation accuracy for both CEU/TSI and CEU/CHB+JPT (Figure 3.6), though the former is

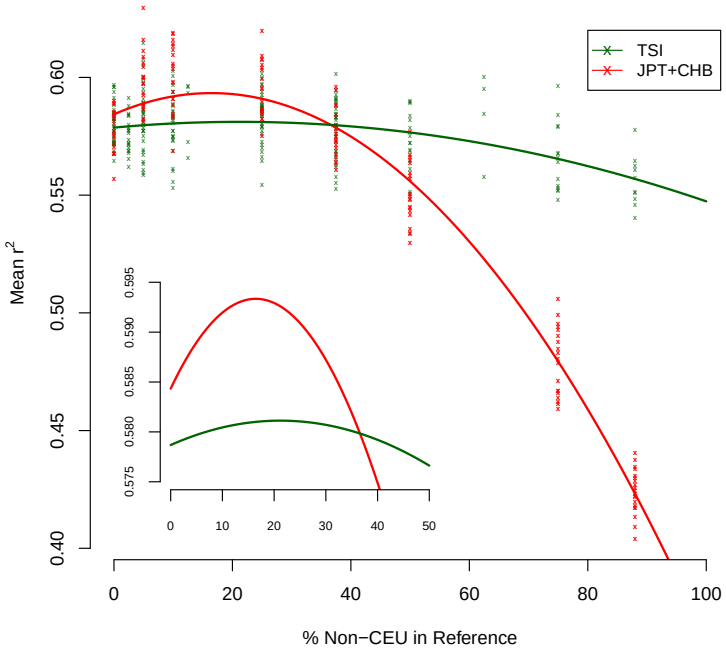


Figure 3.6: The relationship between the dosage r^2 and the proportion of non-CEU samples in a 100-sample reference set. The trend lines are quadratic least squared regression curves, and both explain the data significantly better than a linear relationship (N = 207, p $< 10^{-4}$ and N = 159, p $< 10^{-16}$ for TSI and CHB+JPT respectively). The insert shows an expansion of the trend lines between 0 and 50%.

far less marked than the latter. Beyond a certain proportion of non-CEU samples accuracy starts to fall off as the effect of diversity is outweighed by the effect of mismatching. The optimum population mix is 22% for CEU/TSI, and 17% for CEU/CHB+JPT. It is only above 43% TSI do we see a decrease in imputation accuracy for adding TSI over pure CEU; for CHB+JPT this figure is 33%. This relationship is specific to low-frequency variants.

3.2.5 Discussion

Higher quality reference data and larger sample sizes yield improved imputation accuracy. Using HapMap3 as a reference set compared to using HapMap2 demonstrates this improvement, especially at sites with a low minor allele frequency. While this result was expected I did not anticipate the substantial improvement achieved with large and genetically diverse reference sets. Including samples from such diverse populations as MEX and GIH can provide significant improvement in imputation into UK samples of alleles with a minor allele frequency of less than 5%. Larger reference sets also improve predicted quality scores, with a decrease in overconfident mistakes without inflating under-confident calls.

Overall, an imputation reference set consisting of CEU, TSI, MEX and GIH improves the quality of imputation in all frequency ranges, and greater improvement for very low-frequency SNPs was achieved with very large and highly mixed reference sets. The latter came at the cost of computational power, as well as overly conservative predicted quality scores. The quality scores are likely to be lowered due to the poor match of haplotype frequencies between the reference and target samples, which will in effect decrease the prior on correctly matched haplotypes. Imputation is robust to the precise mix of samples of closely related ancestry (such as CEU/TSI), and small amounts of divergent ancestry can actually improve accuracy (such as CEU/CHB+JPT). However, crude population matching is important, as demonstrated by the reduced accuracy of the CEU+JPT reference compared to CEU+TSI.

My results are consistent with those of Huang et al. (2009), who found that the imputation of Yoruba samples had higher accuracy with a YRI+CHB+JPT HapMap2 reference than with a pure YRI. However, Huang *et al* did not con-

trol for reference size, and showed a much smaller improvement compared to my results, probably due to the highly divergent nature of the HapMap2 populations.

These results imply a set of relatively simple rules for picking imputation reference sets: for the best trade-off between accuracy and computation time, the most diverse mixture of populations that still approximately cluster with the target samples of interest on a world-wide PCA plot should be used. However, if imputing genotypes for low-frequency variants with high accuracy is required, all samples available should be used, with the understanding that this will increase computational time, and cause quality scores to be somewhat conservative.

More recent developments in genotype imputation

Since I wrote the above section additional papers have been published by other researchers that shed further light out the relationship between reference set diversity and genotype imputation. Marchini and Howie (2010) performed imputation using HapMap2 data and demonstrated that combining reference haplotypes across continents gives greater imputation accuracy for low-frequency variation regardless of whether IMPUTE2, Beagle or fastPHASE was used, though IMPUTE2 being the most computationally efficient. Similar experiments using 1000 Genomes data carried out by Sung et al. (2012) showed a similar improvement in imputation low-frequency variation with larger and more diverse reference sets, this time while using the MaCH imputation program.

Over the last few years a concensus has emerged that imputation using world-wide datasets (including data from all available populations) is the simplest way of performing high-quality imputation. For instance, Howie

et al. (2011) demonstrated that such world-wide datasets give optimal or near optimal imputation results using both cross-validation experiments and imputation into real African GWAS data. The rise of pre-phasing techniques (Howie et al., 2012), which allow fast phasing that is independent of reference set size, has made the use of very large reference sets more computational tractable. The appeal of using world-wide reference sets is that they do not require careful selection of reference haplotypes to match the target panel, and thus can be used out-of-the-box on any set of samples.

3.3 Imputation in African populations

The previous section, and indeed most work on imputation to date, focused on imputing variants into European and East Asian datasets. However, many important GWAS datasets have been generated in African populations, notably studies of malaria (The MalariaGEN Consortium, 2009), tuberculosis (Thye et al., 2012) and sickle cell disease (Akinsheye et al., 2011). Just like European studies, these African studies require imputation, particularly where meta-analyses are performed.

Imputation in Africa provides us with its own unique set of difficulties. African populations show a higher degree of genetic diversity than European populations (both within and between populations (Altshuler et al., 2010)). They show less linkage disequilibrium (Altshuler et al., 2010), and substantial differences in patterns of LD between populations (Teo et al., 2009). Given this, it is unsurprising to note that imputation generally performs less well in African populations (Huang et al., 2009; Altshuler et al., 2010; Howie et al., 2011). However, while imputation is more difficult, the rewards are potentially greater. Good quality imputation can greatly improve power when the causal variant is not well tagged (The MalariaGEN Consortium, 2009), and can also allow well-powered meta-analyses in cases where LD differs between populations (Teo et al., 2010).

In this section I will discuss two studies of imputation in African populations. The first investigates HapMap3-based imputation in a GWAS meta-analyses to discover common associations, and the second looks at using a 1000 Genomes Project high-density reference set to impute into a single, diverse African population.

3.3.1 HapMap-based imputation in a GWAS meta-analyses

Description of the study and data

A large collection of blood samples from individuals diagnosed with severe malaria (including cerebral malaria and severe malarial anaemia), along with matched population controls, have been collected by MalariaGEN consortium partners in 9 African countries. 5425 cases and 6891 controls from three of these collections (Gambia, Malawi and Kenya) were genotyped on three different technologies (Illumina 650K, Illumina 1M and Illumina 2.5M respectively). The aim of the experiment was to identify and investigate genetic loci that correlate with severe malaria, and to investigate changes to standard methodology (including QC, imputation and association techniques) that are required to study these African collections.

Due to the difficulty of taking blood from severely ill children, only a small amount of DNA could be extracted and whole-genome amplification was performed, increasing noise in the genotype data. To produce a robust set of genotype calls, three different calling algorithms were used to process intensity data from the Illumina arrays, separately in each of the three cohorts. A set of consensus calls were obtained by treating as missing any genotype that was discordant among algorithms. SNPs with a missing data rate of $> 2.5\%$ were removed. Sample with outlying missingness of heterozygosity were also removed prior to imputation.

Performing and QCing imputation

Imputation was performed using Impute 2.12, using the phased release 2 of HapMap3 from the Impute website (http://mathgen.stats.ox.ac.uk/impute/). As we saw in section 3.2, a diverse reference set provides maximal imputation

accuracy, so I used all HapMap3 haplotypes from all populations (African and non-African) to perform imputation.

The genome was split up into chunks which are either 5Mb, or have 20 000 reference SNPs (whichever is smaller), with an additional 500kb buffer on either side of the segment. I used imputation parameter settings of k = 80 and Ne = 14000. Imputation was performed in parallel for each segment, and segments were reconstructed into chromosomes once all imputations had finished.

To ensure that imputation was performing correctly, I developed a manual imputation QC strategy for examining the output. For each sample cohort I manually examined the following quality-control diagnostic plots to ensure that imputation had performed properly:

(a) a histogram of certainty quality scores across SNPs

(b) a histogram of info quality scores across SNPs

(c) a histogram of per-individual type2 r2 scores, averaged across segments

(d) a histogram of per-segment heterozygous imputation accuracy (proportion of genotyped heterozygous calls that are also confidently imputed as heterozygous)

(e) a plot of per-segment mean type2 r2 scores against the segment's position along the genome

Examples of these plots (taken from the imputation of the Kenya dataset) are shown in Figure 3.7. This imputation run has completed without problems, as the quality scores peak near to 1 (Figures 3.7a and 3.7b), no chunks have abnormally low quality (Figure 3.7d), and the imputation performance shows no significant variation genome-wide (Figure 3.7e). One anomaly is

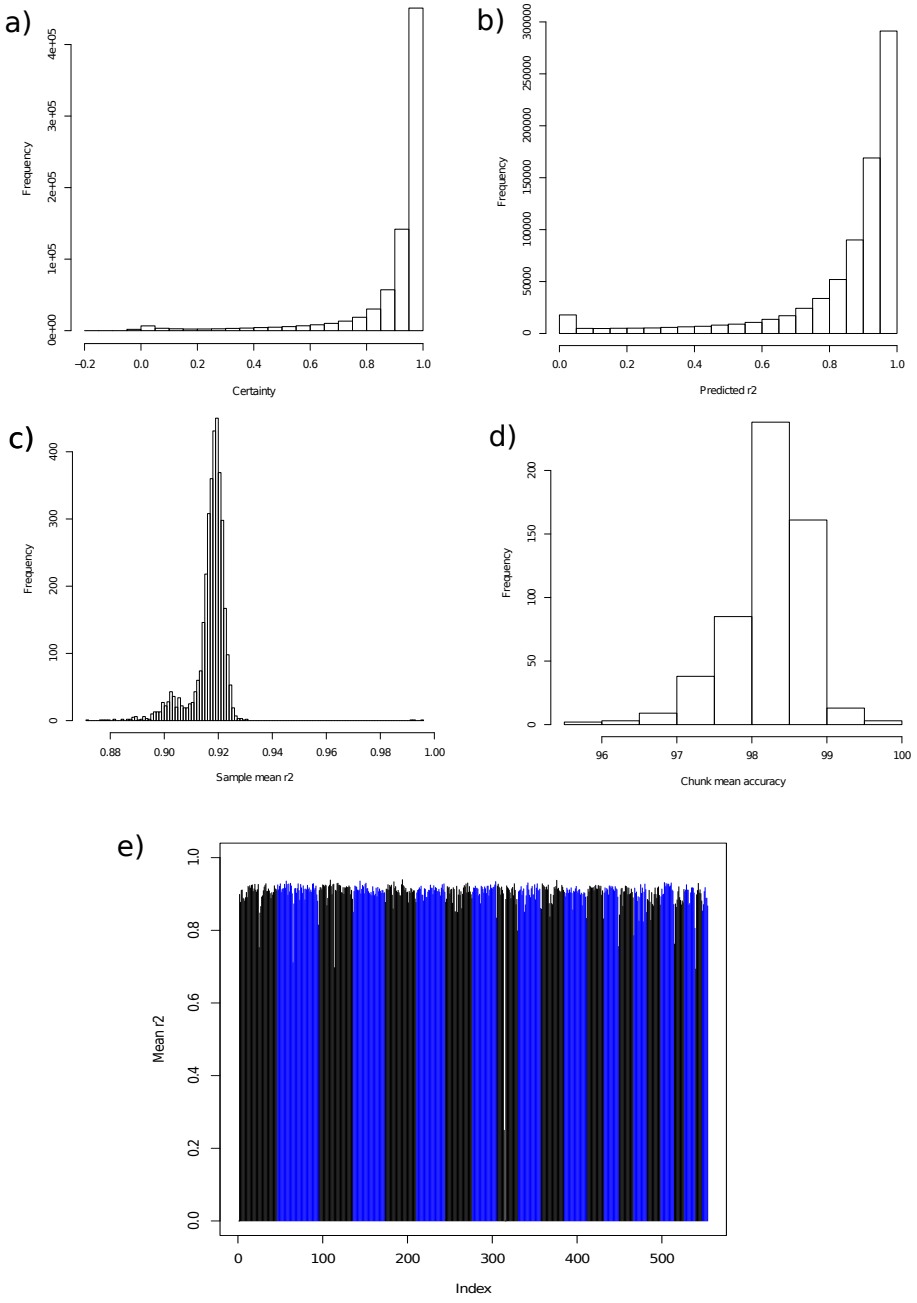


Figure 3.7: Example output from the imputation quality control pipeline for the Kenya imputation. Panels a) and b) show the distribution of two quality scores (certainty and predicted r2) across SNPs, figures c) and d) show the distribution of quality scores across samples and across chunks, and figure e shows the distribution of quality genome-wide (blocks of colour represent chromosomes).

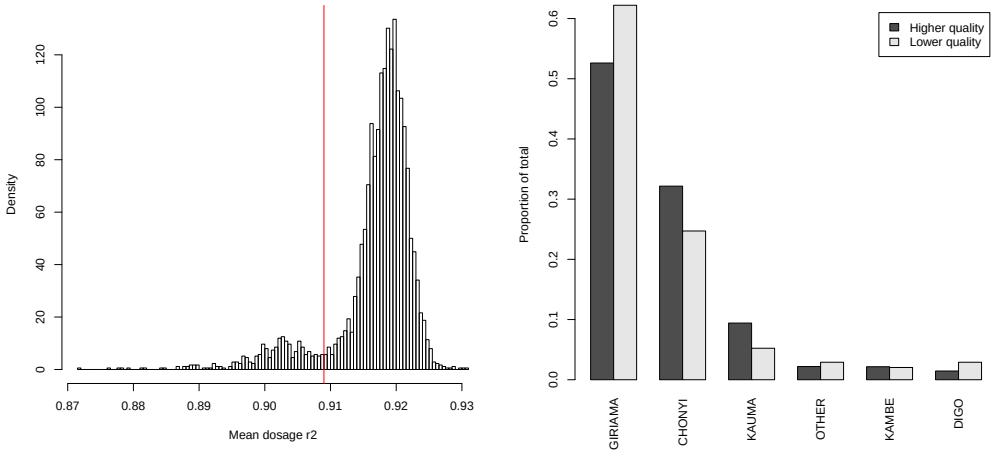


Figure 3.8: a) The distribution of imputation quality (measured by type2 r2) across imputed Kenyan samples. The red line is at r2 = 0.909, and is the minimum between the two peaks. b) The distribution of ethnic groups in the samples in the two peaks. The difference in the two distributions is highly significant (Fisher's exact test, p = 4 x 10^{-4}), suggesting that ethnic differences contribute to the bimodal distribution of imputation quality.

the unusual "bump" in the per-sample imputation plot (Figure 3.7c). Further investigation reveals that this "bump" arises at least in part from ethnic differences within Kenya (Figure 3.8).

Accuracy of imputation across populations

I assessed the accuracy of imputation using the dosage r^2 between imputed and true allele count at directly typed SNPs (This is generated internally by IMPUTE2, and called the type 2 r2). Figure 3.9 shows per-individual dosage r^2 broken down by country. While less accurate than typically achieved in European populations, imputation still captures the majority of common variation in these three populations (a mean dosage r^2 of 0.93 in Malawi, 0.92 in Kenya and 0.87 in Gambia). As in Europeans, common SNPs were better imputed than low-frequency SNPs.

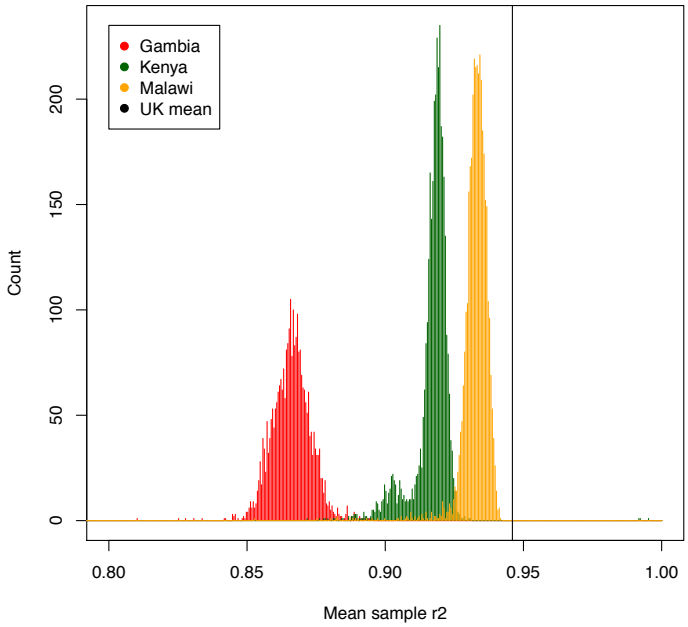


Figure 3.9: Per-sample imputation accuracy measured by dosage r^2, averaged over imputation chunks. Black vertical line shows typical imputation accuracy in a UK population, taken from Section 3.2. Gambian samples (red) perform worst due to the poor coverage of African variation by the Illumina 550K platform, followed by Kenyan samples (green) on the Illumina Omni2.5M, which while dense has limited overlap with our HapMap3 reference, with Malawian samples (yellow) performing best.

As I discussed above, as well as imputation accuracy we are also interested in the numbers of overconfidently and under-confidently imputed SNPs. I evaluated the calibration of the confidence of IMPUTE2 (measured by the info score) against its actual performance at genotyped SNPs. The calibration of confidence was high across our three samples (quality r^2s of 0.93 in Malawi, 0.92 in Kenya, 0.96 in Gambia) but, like overall accuracy, on average worse than in European samples (0.96). I included only SNPs with info score > 0.75 for downstream analyses, leaving a high quality set with mean r2 $>$ 0.9 in all samples, and less than 1% of either very overconfident (predicted

r2 $>$ 0.75, actual $<$ 0.6) or very under-confident (predicted $<$ 0.75, actual $>$ 0.9) SNPs. Taken together, these results suggest the underlying model of IMPUTE2, combined with our diverse reference panel, is generally applicable to samples from African populations.

Despite the high performance of imputation overall, I discovered a number of factors that influenced relative imputation performance, including (i) genotyping platform, (ii) ethnic matching of target GWAS samples to the imputation reference panel, and (iii) homogeneity of individual GWAS collections. The Gambian samples (typed on the Illumina 650Y array) show much poorer imputation quality (Figure 3.9) than our Kenyan and Malawian samples (typed on Illumina chips with $>$ 1 million SNPs). While genotyping array represents the single most important factor to imputation accuracy, two aspects of population genetics are also critical: good matching between reference and target samples and homogeneity within a GWAS sample (illustrated by the small number of samples of differential ancestry in Kenya with poorer imputation quality seen in Figure 3.8).

3.3.2 1000 Genomes-based imputation in a single, diverse population

Description of the data

The MalariaGEN Kenya dataset, included in the previously discussed meta-analysis, was genotyped on Illumina's Omni2.5 genotyping chip. This high-density SNP array is the first of a new generation of genotyping chips designed to assay a subset of the large numbers of SNPs discovered by resequencing studies, such as the 1000 Genomes Project. The Kenya malaria dataset is the first of many MalariaGEN datasets that will be genotyped on this chip,

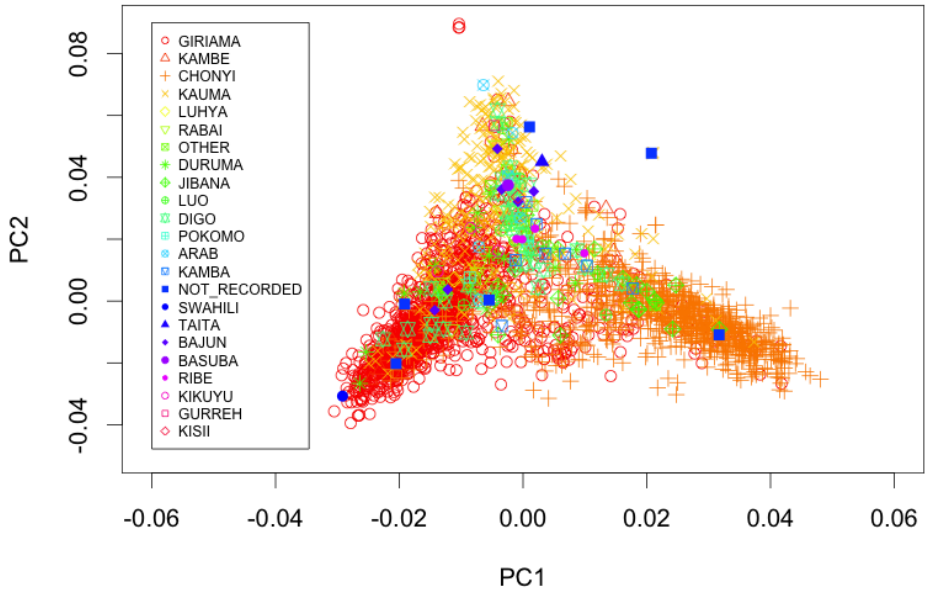


Figure 3.10: A PCA of the 2502 Kenyan samples, coloured by ethnicity.

as it is believed the higher density will allow us to overcome the LD issues that can confound cross-population meta-analysis.

However, this dataset also provides us with an opportunity to make a detailed assessment of the accuracy of high-density imputation into a diverse African population. Two factors make this a particularly good dataset for such assessment. Firstly, the 2502 Kenyan samples are ethnically diverse, as shown by their large number of stated ethnicities, and their significant structure on a principal component plot (both shown in Figure 3.10). We can use this to investigate the impact of target set diversity and structure on imputation accuracy. Secondly, the Omni2.5 is a particularly good system to assess GWAS imputation, as it is built on the backbone of an OmniExpress (a typical, middle cost GWAS chip), with a large number of 1000 Genomes

Reference Set	N. haplotypes	CPU use	Memory use
Pilot Yoruba	120	143hrs	20.5 Gb
Pilot (all samples)	360	163hrs	20.9 Gb
Phase I Yoruba+Luhya	400	165hrs	21.1 Gb
Phase I (all samples)	2420	220hrs	25.4 Gb

Table 3.4: Reference sets used for testing 1000 Genomes imputation, with resources required for imputation.

SNPs added. The OmniExpress backbone, as a model of a GWAS chip, can be imputed into from a high-density dataset, and the additional content can then be used as a validation set.

Performing imputation

Because the Omni2.5 can only be used to assess imputation results for SNPs on that chip, I decided to reduce imputation complexity by only using the Omni2.5 data generated as part of the 1000 Genomes Phase 1. I made a set of four test reference sets from this data, consisting of two 1000 Genomes pilot and two phase 1 datasets, with one containing only African samples, and one containing all samples (Table 3.4).

Imputation was performed only on Chromosome 1, using the Impute2 pipeline described in section 3.3.1. This took between 140 and 220 CPU hours and 20 to 26 CPU Gbs, and was only weakly dependent on reference set size (Table 3.4).

Imputation accuracy was measured using dosage r^2 between imputed and true genotyped at non-OmniExpress SNPs. For per-individual accuracy, I used heterozygous certainty (the mean heterozygous posterior probability at truly heterozygous sites).

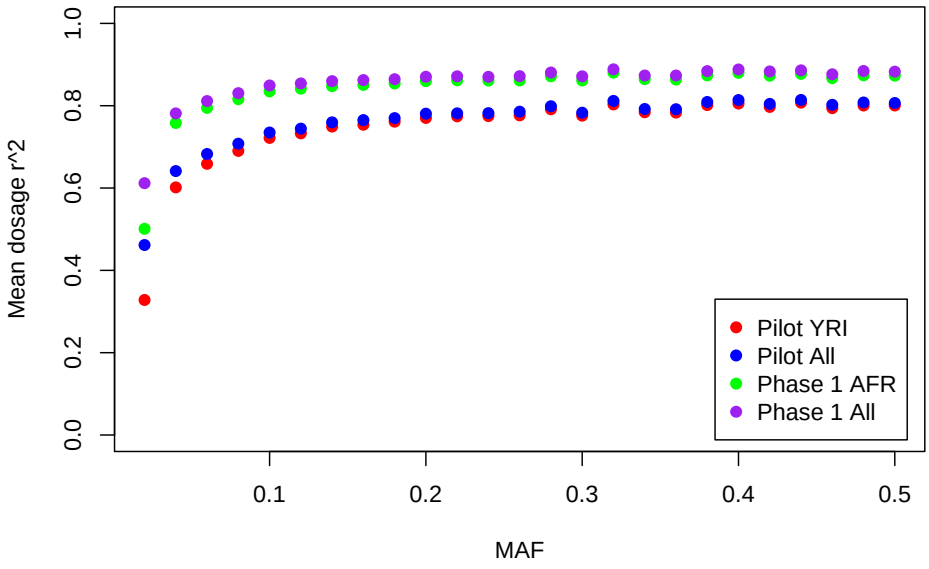


Figure 3.11: The relationship between imputation accuracy and call rate using the various reference sets. YRI=Yoruba, AFR=African. Note that these data has not been filtered by quality score.

Impact of reference set on imputation

Looking first at the pilot data, imputation of 1000 Genomes variants into Kenya performed very badly (Figure 3.11). Even common variants had a mean dosage r^2 of around 0.7. However, going to the Phase 1 data dramatically improved imputation performance, bringing the dosage r^2 up to over 0.8. Interestingly, the non-African haplotypes made almost no improvement to imputation for common SNPs in either the pilot or the phase 1 data. However, for the very low-frequency SNPs (MAF $< 2\%$), introduction of non-African haplotypes dramatically improved imputation, both for the pilot data (0.33 to 0.45) and for the Phase 1 data(0.51 to 0.61). This again reinforces the value of distantly related haplotypes to improve imputation

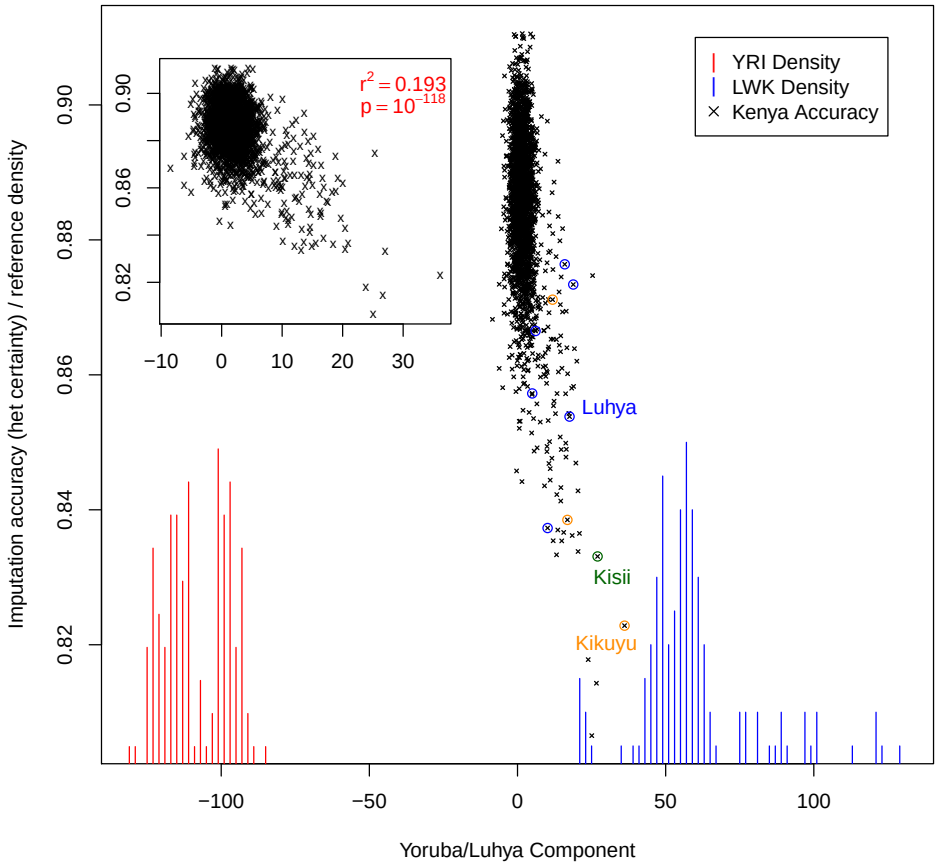


Figure 3.12: Individual variation in imputation accuracy with YRI/LWK principal component. Coloured bars represent the location of reference individuals. A few outlier ethnicities are circled. Inset expands the Kenyan region of the component.

for low-frequency variation.

Impact of target sample on imputation

To investigate the impact of population structure on imputation accuracy, I found the first principal component for the Luhya and Yoruba Phase 1 reference sets, and projected all Kenyan samples onto this axis (using the R

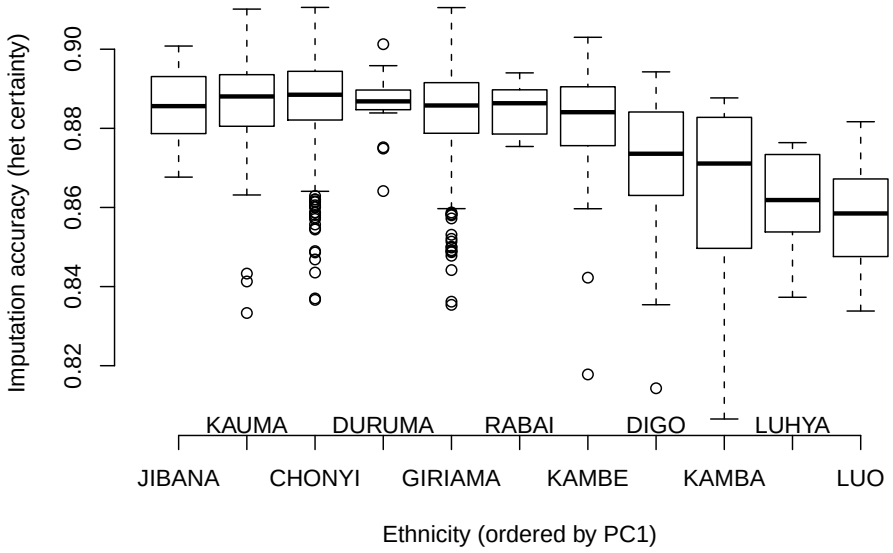


Figure 3.13: The variation in imputation accuracy between the major ethnic groups, ordered by distance from YRI

package snpMatrix). I then correlated this value with the imputation accuracy for the Kenyan samples imputed with the AFR Phase 1 dataset (Figure 3.12). Surprisingly, I found a significant inverse correlation, with samples that lay closer to the Luhya cluster having lower imputation accuracy.

The same relationship appeared to hold if median accuracy across ethnicity was considered, with ethnicities that were genetically more similar to the Luhya having lower median quality (Figure 3.13). However, it also appears that samples that are closest to the Yoruba also show a slight decrease in imputation quality. This suggests that the decrease in quality is in fact due to being ethnic outliers from the main Kenyan cluster, rather than due to similarity to reference populations. This may due to the effect of phasing: IMPUTE2 uses the entire target set to perform phasing, which will lead to

samples that are not closely related to the rest of the target set having worse phasing, and thus lower imputation accuracy.

Conclusions

I believe that the results above allow us to draw four conclusions about high-density imputation in diverse populations:

1. The Phase 1 1000 Genomes reference set grants significant improvements in imputation for African populations
2. Low-frequency imputation benefits from extreme diversity, illustrating the need for world-wide genotype reference sets
3. Imputation accuracy in Kenya varies significantly by ethnic group
4. The relationship between accuracy and target/reference match can be complex and counter-intuitive

3.4 Using imputation to explore the impact of loss-of-function variants on complex disease

3.4.1 Loss-of-function variants and the 1000 Genomes project

Loss of function (LoF) variants are SNPs, indels or CNVs where one allele entirely removes the function of one or more genes. These can include SNPs that disrupt a start codon, create a new stop codon or disrupt an essential splice site, indels that create a frame-shift and CNVs that partially or entirely delete a gene. Clearly these mutations are major candidates for having phenotypic effects, and many of the known Mendelian diseases are caused by LoF mutations, but it is also clear that many LoF variants are relatively benign and circulate at high frequency in the population. As part of the 1000 Genomes project, the LoF Group (now the Functional Integration Group) was founded to identify and investigate both common and rare LoF variants.

After extensive filtering, we discovered 1285 high quality LoF mutations in the 1000 Genomes pilot (MacArthur et al., 2012). This was a particularly challenging project, largely due to the high proportion of false positives in this dataset: 1666 putative loss-of-function variants were excluded due to possible mapping artefacts, errors in gene model and systematic sequencing errors. In total, we concluded that the average human genome contains around 100 loss-of-function mutations, with approximately 20 genes homozygously inactivated.

As well as identifying these mutations, an important aim of the project was to shed light on the biology of these mutations. This included identifying differences in the property of genes that harbour common LoF mutations and those where LoF mutations cause Mendelian disease, as well as using RNA-

Seq to study the impact of LoF mutations on gene expression. In this section, I will describe a study that I carried out, using genotype imputation to assess the impact of loss-of-function variants on human complex disease.

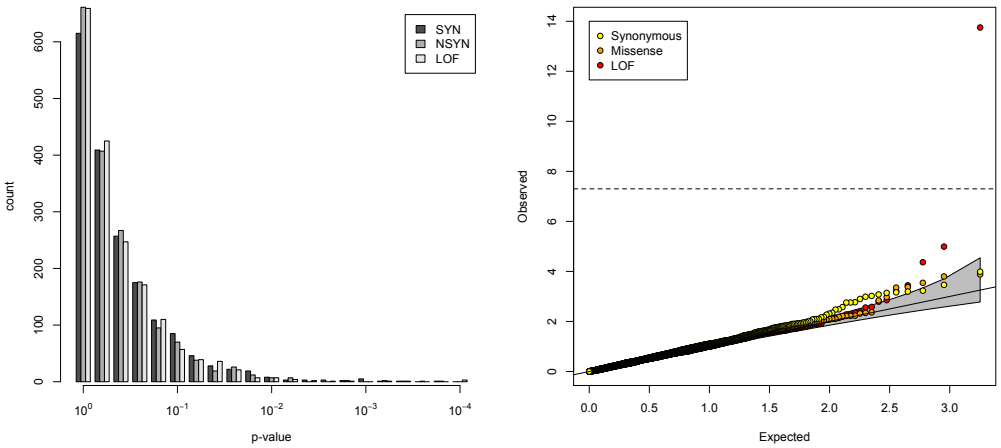
3.4.2 Performing imputation and association analysis

To assess whether LoF variants were enriched for effects on complex disease risk, I imputed all SNPs and indels genotyped in the CEU population in the 1000 Genomes low-coverage pilot (Project, 2010) into the complete Wellcome Trust Case Control Consortium 1 (WTCCC1) dataset (Wellcome Trust Case Control Consortium, 2007), comprising 2,938 controls and 13,241 cases that pass sample QC.

Genotypes for CEU SNPs and indels were obtained from the July 2010 release, and were merged with SNP genotypes from HapMap3 release 2. Imputation of these variants into the WTCCC1 dataset was performed using the IMPUTE2 pipeline described in section 3.3.1.

I investigated potential associations with complex disease risk for 625 high-confidence LoF variants identified as polymorphic in the CEU population. Of these variants, 417 imputed well enough in both controls and at least one cohort to go ahead with association (using an info score threshold of 0.2), resulting in a total of 2901 association tests in the seven disease cohorts. Only 3 variants were close enough to the threshold to be assessed in some cohorts but not others.

I performed a frequentist association analysis using the program SNPTest (Marchini et al., 2007), version 2.2.0. I used an additive model of risk, and a likelihood score test to account for uncertainty in imputed genotypes. Matched synonymous and missense sets were calculated using allele frequencies in controls, taking random draws without replacement of synonymous



(a) Histogram **(b)** QQ plot

Figure 3.14: Association of coding variants with complex disease risk. Observed -log10(P) values for disease association in 16,179 individuals from seven complex disease cohorts and a shared control group, following imputation of variants identified by the 1000 Genomes low-coverage pilot, are plotted against the expected null distribution for all LoF variants and frequency-matched missense and synonymous SNPs.

and missense variants from the same 1% frequency bin as each LoF variant. In both cases, five random draws were made.

3.4.3 Results

There were no significant detectable enrichments of associations for LoF variants compared to missense variants at P value thresholds of 10^{-5}, 10^{-4} or 10^{-3} (Fisher's exact P values 0.4994, 0.1245 and 0.8034, respectively), suggesting that common LoF variants are not substantially over-represented among complex disease risk variants compared to other functional coding polymorphisms.

The major caveat of this analysis is that the systematically low frequencies of LoF variants result in a decrease in imputation accuracy, and a subsequent

drop in power to detect association. However, note that the *NOD2* frameshift indel, with an allele frequency of <3% and an odds ratio of approximately 3, achieved a P value of 1.78 x 10^{-14} for association with Crohn's disease despite having a low info score for imputation (0.25). This suggests that my analysis would have successfully identified other LoF variants with large effects, even where allele frequency and imputation accuracy was relatively low. Additionally, imputation quality was high for common LoF variants, allowing us to positively rule out a major role of common LoF variants in complex disease.

In addition to the *NOD2* variant that achieved genome-wide significance, two LoF variants achieved Bonferroni-corrected significance: rs16380, a frameshift indel in *ZNF3* (associated in type 1 diabetes), and a novel frameshift indel at chr1:152018423 in the gene *SLC27A3* (associated in hypertension). I pursued the evidence for association for the *ZNF3* variant using data from a meta-analysis of genome- wide association studies of type 1 diabetes incorporating 7,514 cases and 9,045 controls (Barrett et al., 2009a). 3 SNPs were in strong linkage disequilibrium with rs16380 based on 1000 Genomes pilot data that were also examined in the meta-analysis; these showed only nominal significance in the meta-analysis (P = 0.03-0.04), and this association was driven entirely by the samples overlapping with the WTCCC1 analysis: looking only at samples that were not overlapping with WTCCC1, the P value was 0.4012. This suggests that the marginally significant association in the WTCCC1 samples is a chance finding rather than a genuine association.

3.5 Concluding remarks

Throughout this chapter we have seen how new reference sets can add significant value to genome-wide association studies via genotype imputation. This has included allowing assessment of low-frequency variations from both HapMap and 1000 Genomes reference sets, as well as facilitating meta-analysis of diverse African populations and inferring the impact of newly discovered loss-of-function variants in human disease.

However, we have also seen that imputation is most useful when we have access to large, diverse and high-density reference sets. The well-matched but small HapMap2 reference set is not sufficient to allow accurate imputation of low-frequency variation in Europeans (section 3.2). Likewise, despite its high marker density, the 1000 Genomes pilot data is not able to produce accurate imputation in a diverse African population (section 3.3.2). These experiments have shown that to accurately impute all markers down to low frequency, we require sample sizes on the scale of the HapMap3, but with the high-density granted by sequencing.

In essence, this is what has now been achieved by the 1000 Genomes Project Phase 1 release (Project, 2012), which we have seen is capable of imputing low-frequency variation even in a diverse African population (section 3.3.2). This reference set, and subsequent imputation sets from the 1000 Genomes Project and other sequencing projects, presented a new opportunity to extend the reach of genome-wide association studies into new frequency ranges and classes of variation. As such, they represent a valuable, and continually growing, resource for adding value to GWAS.